論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	An Investigation of Context-Driven Caption Generation
著者(和文)	YANG Zhishen
Author(English)	Zhishen Yang
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第396号, 授与年月日:2025年3月26日, 学位の種別:課程博士, 審査員:岡崎 直観,德永 健伸,井上 中順,篠田 浩一,荒瀬 由紀
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第396号, Conferred date:2025/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Doctoral Dissertation

An Investigation of Context-Driven Caption Generation

Zhishen Yang

February 26, 2025

Department of Computer Science
School of Computing
Institute of Science Tokyo

A Doctoral Dissertation
submitted to School of Computing,
Institute of Science Tokyo
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Zhishen Yang

Thesis Committee:

Professor Yuki Arase
Professor Nakamasa Inoue
Professor Naoaki Okazaki (Supervisor)
Professor Koichi Shinoda
Professor Takenobu Tokunaga

Committee members listed alphabetically by family name

An Investigation of Context-Driven Caption Generation*

Zhishen Yang

Abstract

This thesis investigates context-driven caption generation through systematic studies of two specialized domains: news image and scientific figure captioning. Unlike traditional image captioning, which focuses solely on describing visual content, these two domains require sophisticated integration of contextual information to generate meaningful and accurate captions. The research addresses two fundamental questions: How can models effectively integrate information from visual and textual modalities to generate informative captions, and what is the relative importance of textual versus visual context in context-driven caption generation?

The first study on news image captioning demonstrates that generating appropriate captions requires understanding the visual content and its relationship to the broader news narrative. Traditional image captioning approaches are insufficient for this task as they cannot capture the journalistic significance of images within their news context. The study introduces a novel Transformer-based architecture from news articles that effectively integrates visual features with textual context. Through extensive experiments using both automatic metrics and human evaluation, the research reveals that while textual context from news articles provides the primary information for generating contextually appropriate captions, incorporating visual features through the proposed model leads to more context-relevant captions. The model outperforms previous state-of-the-art approaches across multiple evaluation metrics, demonstrating the effectiveness of the transformer-based architecture in handling multimodal information.

*Doctoral Dissertation, School of Computing
Institute of Science Tokyo, February 26, 2025.

The second study examines scientific figure captioning, which presents unique challenges distinct from natural image captioning. Scientific figures typically contain complex data visualizations, graphs, and technical diagrams that require domain-specific knowledge for proper interpretation. Unlike natural images, where visual content might be self-explanatory, scientific figures often need substantial context from the accompanying research papers to be understood and described adequately. The research introduces SciCap+, an enhanced dataset that augments scientific figures with two crucial contextual elements: mention-paragraphs (text segments referencing the figures) and OCR-extracted text from within the figures. The study reframes scientific figure captioning as a knowledge-augmented image captioning task, demonstrating that effective caption generation requires the integration of multiple context sources.

Using the M4C-captioner model as a baseline, the research shows that incorporating information from mention-paragraphs and OCR tokens improves captioning performance significantly compared to approaches using visual features alone. To validate the knowledge-augmented approach, we conducted human evaluation, which revealed that even expert annotators struggled to write accurate figure captions without access to contextual information. This finding highlights the need for context in scientific figure captioning.

The research demonstrates that textual context is crucial for news image and scientific figure captioning tasks. For news images, article text provides essential context for understanding news values, while mention-paragraphs supply scientific background and technical details for scientific figures. Visual features serve a complementary role. In news captioning, they enhance caption quality when combined with textual context, and for scientific figures, OCR text within figures provides important technical information. The studies also validate the effectiveness of transformer-based architectures for multimodal integration, with the proposed news captioning model successfully combining visual and textual features and the M4C-captioner architecture effectively integrating multiple knowledge sources for scientific figures.

The methodological contributions include the development of transformer-based architectures for multimodal context integration, attention mechanisms for cross-modal feature fusion, and establishing practices for handling domain-specific challenges. The research also makes resource contributions through SciCap+, an enhanced scientific figure dataset incorporating mention-paragraphs

and OCR-extracted text, a refined version of the news-image captioning dataset with complete article text, and the implementation of evaluation frameworks for both tasks.

Our research reveals two key insights into context-driven image captioning: the necessity of context and the dominance of textual context. Both experiments and human evaluation demonstrate that context plays a crucial role in context-driven image captioning. The human evaluation shows that even annotators struggled to write captions without proper context. Regarding textual dominance, textual context serves as the primary source of information for generating informative captions, while visual context plays a secondary role.

This study finds that while textual context plays a primary role in context-driven image captioning tasks, optimal performance requires effective integrating multiple modalities. These findings have meaningful implications for advancing multimodal learning and enhancing automated caption generation in specialized domains. Future research directions identified include developing visual encoders trained specifically for news images and scientific figures, improving methods for extracting and utilizing contextual information from research papers, and exploring more sophisticated attention mechanisms for handling complex relationships between text and images.

The primary applications of our research are in journalism and scientific publishing. In journalism, the system can assist editors in generating contextually aware captions instantly when journalists upload images with articles, streamlining workflows in the fast-paced news industry. The system can also help authors generate comprehensive and informative figure captions that accurately describe their methods and results for scientific publishing. This study advances our understanding of how contextual information enhances caption generation while providing practical frameworks for implementing context-aware captioning systems across specialized domains.

Keywords:

Multimodal Machine Learning, Image Captioning, Context-Driven Image Captioning, Multimodal Integration

Acknowledgements

Writing these acknowledgements for my PhD dissertation reminds me that my PhD journey has reached its conclusion. My appreciation first goes to Professor Naoaki Okazaki. As my PhD advisor, insightful advice and generous support from Prof. Okazaki helped me navigate research challenges. I would also like to thank my PhD committee members, Prof. Yuki Arase, Prof. Nakamasa Inoue, Prof. Koichi Shinoda, and Prof. Takenobu Tokunaga, for their valuable insights and suggestions, which greatly benefited me and helped refine my research and PhD dissertation.

I appreciate Ms. Yukiko Konishi and Ms. Naoko Furuya, the administrative staff of the Okazaki Lab, for their exceptional support, which allowed me to focus on my research without additional concerns. I am especially grateful to Ms. Konishi for her kind words and timely assistance.

My appreciation extends to my research collaborators: Dr. Hideki Tanaka, Dr. Dabre Raj, Dr. Tosho Hirasawa, and Dr. Edison Marrese-Taylor. Working with all of you was both enjoyable and inspiring.

I extend my gratitude to the Academy of Leadership (ToTAL) faculty members for their guidance. In particular, I appreciate Prof. Keisuke Yamada and Prof. Yuri Matsuzaki for their invaluable feedback and guidance during lectures, group work, and my final assessment presentation. I also thank the administrative staff, Ms. Yu Iwai and Ms. Keiko Tsuchiya, for their kindness and support.

Many thanks to all members of the Okazaki Lab for your kind help and support. Your presence made my time at the lab truly meaningful, and the conversations and moments we shared will remain cherished memories for years to come. In particular, I would like to extend my sincere gratitude to Vijay Daultani, Sangwhan Moon, Sakae Mizuki, Ayana Niwa, Ao Liu, Youmi Ma, An Wang, Anantaprayoon Panatchakorn, Tatsuya Hiraoka, Mengsay Loem, Hsuan-Yu Kuo, Wiem Ben Rim, Hongxiang Wan, Marco Cognetta, Ryuto Koike, Kosuke Endo, Yuki Maruyama, David Pohl, Hinari Shimada, and all colleagues I had the privilege of meeting in

the Okazaki Lab.

Last but not least, I am deeply grateful to all my family and friends for giving me the courage to overcome obstacles. My deepest gratitude belongs to my mother, whose unconditional support and silent dedication gave me the courage to follow my dreams.

PhD study aims to cultivate researchers who can identify problems, develop plans, overcome difficulties, and complete research projects. Through my PhD, I have realized that research is not just an individual effort but a result of the support from countless people. Words cannot fully express my gratitude, and though I may not name everyone who has helped me, I will never forget every act of kindness and support.

I would like to conclude my acknowledgements with wisdom from Chinese philosopher Xunzi (荀子) to remind and encourage me:

昨日之深淵，今日之淺談。路雖遠，行則將至。事雖難，做則可成。

昨日の深淵は、今日では浅き語り。道は遠くとも、歩めば必ず至る。事は難しくとも、為せば成る。

What was an abyss yesterday becomes shallow talk today. Though the path is distant, you will reach it if you walk. Though tasks are difficult, they can be accomplished if you do them.

This PhD journey has taught me that perseverance will lead to success, no matter how distant or difficult the path ahead may seem.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Humans use sophisticated biological mechanisms to perceive and interact with the environment. Human bodies seamlessly convert environmental stimuli through sensation into meaningful information that our brains can interpret. Humans experience their surroundings through five main sensory channels: vision, hearing, smell, taste, and touch, each supported by specialized organs and receptors [43]. These primary senses detect and process specific input types, with each sensory pathway precisely tuned to capture particular environmental inputs and convert them into interpretable neural signals.

The human brain has cognition functions to integrate diverse sensory inputs into unified perceptions, enabling rich environmental understanding [65, 43]. This multimodal integration facilitates sophisticated cognitive functions like reasoning and decision-making [41, 23]. A representative example of multimodal integration is to describe visual stimuli through natural language: vision-language understanding. Researching artificial intelligence (AI) systems that can achieve similar multimodal understanding represents a crucial step toward more human-like AI capabilities.

1.2 Motivation

Image captioning has emerged as a trending research in artificial intelligence, aiming to automatically generate natural language descriptions of visual content [32, 63, 66]. Humans have sophisticated capabilities in visual processing: receiving visual stimuli and integrating them with existing knowledge to produce informed scene interpretations. This cognitive process extends beyond traditional image captioning into context-aware image understanding, which requires inte-

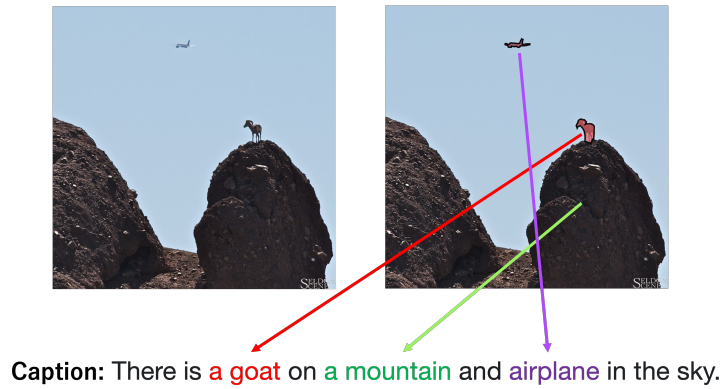


Figure 1.1: A typical example of traditional image captioning tasks where models identify key objects and basic spatial relationships without providing contextual descriptions. The caption demonstrates basic object detection (identifying goat, mountain, and airplane) and spatial relationship understanding (on, in the sky).

grating broader contextual information into caption generation.

Traditional image captioning methods focus on identifying and describing visible elements within images and their spatial relationships, such as objects, actions, and scenes, and generate descriptions based solely on the visual content present in images without referring to broader contextual information or background knowledge. As illustrated in Figure 1.1, such captions typically enumerate visible objects (goat, plane, mountain) and describe their basic spatial relationships.

Figure 1.2 demonstrates the role of context in image caption generation. A conventional captioning model, with only visual information access, produces a generic description: "A lit metal tower at night." While factually correct, this surface-level caption fails to capture the identity of the landmark and its cultural significance. In contrast, when granted appropriate contextual information, a context-driven model identifies the landmark as Tokyo Tower, generating a more precise and meaningful caption. This transformation from generic to specific description demonstrates how contextual knowledge allows caption generation systems to move beyond simple visual enumeration toward human-like scene interpretation that captures deeper semantic meaning.

Traditional image captioning approaches face significant limitations when applied to specialized domains that require deep contextual understanding. This

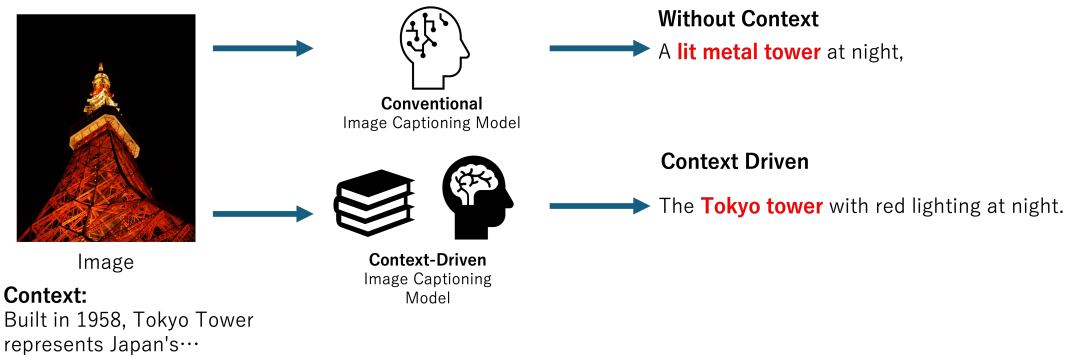


Figure 1.2: Context enhances caption generation quality. Without context, the model generates a generic description: "A lit metal tower at night." With proper contextual information, the model produces a more informative caption identifying the landmark as "Tokyo Tower," demonstrating how context enables more precise and meaningful caption generation.

research investigates context-driven caption generation through news media and scientific literature. These domains present distinct real-world challenges in multimodal integration: captioning news images must be interpreted within broader journalistic narratives while captioning scientific figures requires technical understanding within a research context. By examining these complementary domains, we advance practical applications and theoretical understanding of how AI systems can effectively combine visual and textual information.

In news media and scientific literature, visual content alone is insufficient for generating meaningful captions for images. News images derive significance from their relationship to broader narratives described in associated articles, requiring sophisticated integration of textual context for proper interpretation. Our investigation of news image captioning directly supports the journalistic workflow by developing systems that suggest contextually appropriate captions that maintain consistency between article content and visual elements. Through our study of news image captioning [68], we advance multimodal machine learning by addressing the technical challenges of integrating visual features with rich textual context.

Scientific figures present unique challenges that distinguish them from natural images. These figures typically contain data visualizations, graphs, and tech-

nical diagrams that require sophisticated integration of domain knowledge and research context from academic papers for proper interpretation. Our research on scientific figure captioning [69] investigates automating this interpretation process within research contexts, addressing fundamental challenges in academic communication. With the growing volume of scientific literature, clear and informative figure captions have become crucial for efficient knowledge sharing. Our work supports researchers in crafting more effective figure captions, thereby improving the clarity and accessibility of scientific communication across disciplines.

Through two systematic studies, this research contributes to understanding how contextual information enhances caption quality. We examine the unique technical challenges and methodological approaches for integrating visual understanding with contextual information through detailed investigations of news image and scientific figure captioning. These complementary studies advance both practical applications in journalism and scientific communication. The insights gained from these two studies provide a foundation for developing more advanced context-aware captioning systems across various specialized domains.

1.3 Research Overview

In news media, effective captions must synthesize visual content with broader journalistic narratives to convey news value. The challenge extends beyond describing what appears in an image to interpret visual content with news context. Scientific figure captioning presents a different but equally complex challenge, interpreting technical visualizations requires deep integration of domain knowledge and research context from academic papers. While these domains differ in how context manifests, both demonstrate the essential role of contextual understanding in generating meaningful captions.

Based on these domain-specific challenges, this research addresses two research questions

1. How can models effectively integrate information from visual and textual modalities to generate informative captions?

   This question examines the technical approaches for combining multiple contexts in caption generation.

2. What is the relative importance of textual versus visual context in context-driven caption generation?

   This question investigates:

   a) The contribution of textual context to caption quality.

   b) The contribution of visual context to caption quality.

1.3.1 News Image Captioning

Image captioning is a multimodal task automatically generating natural language descriptions for images. While conventional image captioning focuses solely on describing visible content in images, news image captioning presents unique challenges as it requires understanding visual information and associated textual context from news articles.

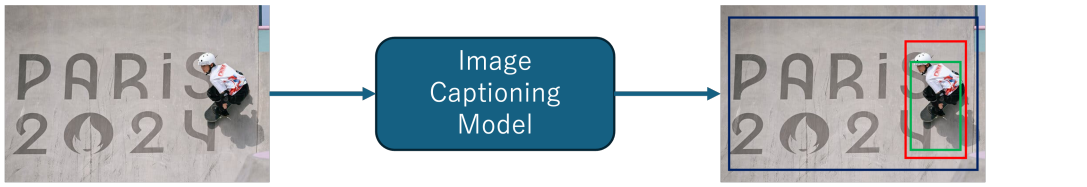


Figure 1.3: Illustration of basic image captioning without textual context from the associated news article. The captioning model takes only the image as input and generates a purely visual description ("A skateboarder descends a gray ramp"), missing contextual information from the news article.

As illustrated in Figure 1.3[1], a conventional image captioning model looking only at the image would generate a descriptive caption like "A skateboarder descends a gray ramp." Moreover, news image captions often contain contextual information that cannot be derived from the visual content alone. Figure 1.4 demonstrates how incorporating the associated news article text enables the captioning model to generate a more informative caption that includes relevant context, identifying the skateboarder as "Zheng Haohao, known as Lilibet" and highlighting her distinction as "the youngest athlete at the Paris Games."

[1]https://www.nytimes.com/2024/08/07/world/asia/china-olympics-tactics.html

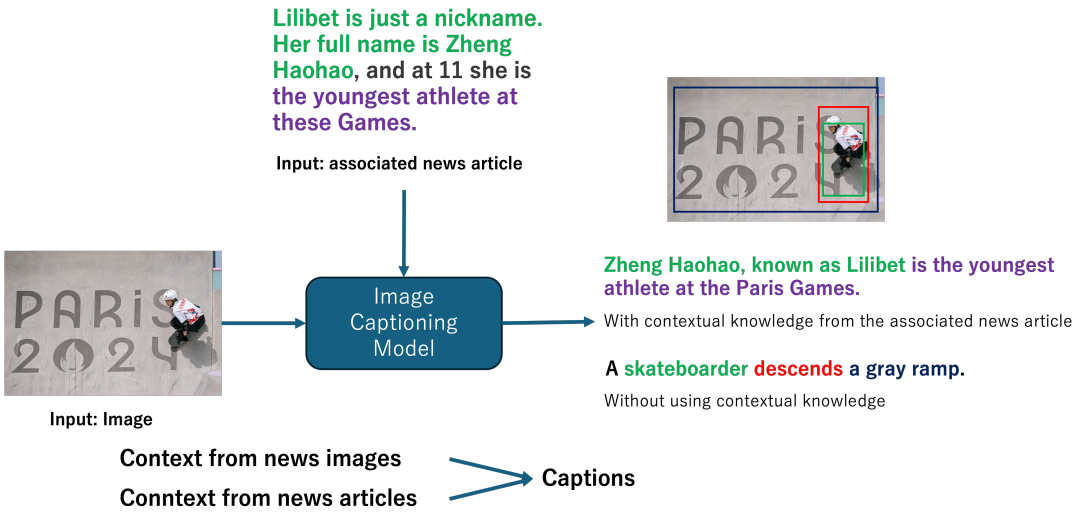


Figure 1.4: Comparison of image captioning with and without context from the associated news articles. The top caption incorporates article context to identify the skateboarder and her significance at the Paris Games. In contrast, the bottom caption shows a limited description that includes only visual information.

The traditional image captioning model aims to generate descriptions of images without background information integration. Moving from traditional image captioning to news image captioning represents a step toward more sophisticated multimodal systems. This transition introduces several key characteristics that define the news image captioning:

1. Input Composition
    - Visual content consists of photographs capturing real-world events, people, and places
    - Required contextual information comes from associated news articles that provide narrative background
    - Both modalities must be effectively integrated to generate appropriate captions
2. Caption Requirements
    - Must convey both visual descriptions and journalistic value

- Must effectively balance visual details with news significance

These characteristics necessitate specialized model architectures capable of fusing visual and textual features through advanced mechanisms such as multimodal transformers. This evolution motivates two primary research directions:

1. Developing multimodal models that effectively integrate textual and visual information.

2. Assessing the relative significance of textual context compared to visual features in generating captions for news images. Understanding this balance between visual and textual contributions advances our knowledge of multimodal learning and provides practical insights for designing more effective context-driven image captioning models.

Exploring these research directions, we take meaningful steps toward bridging the gap between AI systems and human journalists in their ability to craft news image captions that capture what is seen and what it means in the broader news context.

1.3.2 Scientific Figure Captioning

Scientific figures are a fundamental medium for communicating research findings in scholarly documents. These figures and captions provide readers with visual representations of complex information. Unlike captions for natural images, which primarily describe visible objects and scenes, scientific figure captions must interpret data, explain methodological insights, and highlight key findings within the broader context of the research.

Scientific figure captioning presents unique challenges that distinguish it from traditional image captioning, which caption nature images:

1. Scientific figures are inherently different from natural images. Instead of objects and scenes, they typically consist of data visualizations, graphs, charts, and technical diagrams.

2. Writing appropriate figure captions requires describing the visual elements and providing analysis that conveys the author's intended message to readers.

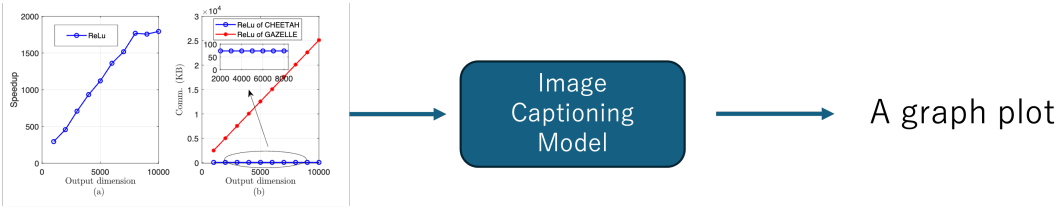


Figure 1.5: Illustration of traditional image captioning workflow for scientific figures. The left shows an example technical figure with performance comparison plots between CHEETAH and GAZELLE systems. In traditional image captioning approaches, a captioning model (center) processes the figure and outputs a generic description, "A graph plot" (right), without incorporating any technical context or deeper analysis of the data. This example demonstrates how traditional captioning models often fail to capture the technical substance of scientific figures, producing only surface-level visual descriptions.

As illustrated in Figure 1.5 and Figure 1.6, when presented with a graph comparing algorithmic performance, a caption must go beyond merely describing the visible trend lines, it needs to explain 1. what is being compared, 2. why the comparison matters, and 3. what conclusions can be drawn from the results. This requires understanding the representation of visual data and the research context.

In Figure 1.5, where a basic image captioning model only describes the presence of the graph plot, missing the critical comparison between CHEETAH and GAZELLE systems that gives the figure its meaning. Captions of scientific figures need to explain what is being compared, why the comparison matters, and what conclusions can be drawn from the results. This requires understanding the representation of visual data and the broader research context in which the figure appears.

The transition from traditional to scientific captioning represents a significant advancement toward more sophisticated multimodal systems. This progression introduces several key characteristics that define scientific figure captioning:

1. Input Composition
    - Visual content consists of data visualizations, plots, and technical di-

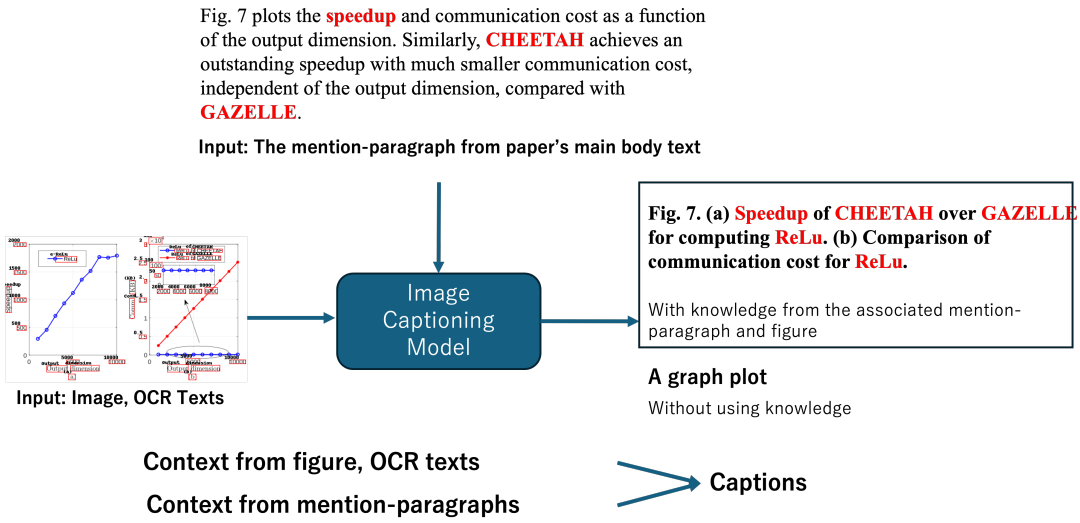


Figure 1.6: Comparison between context-aware and traditional image captioning for scientific figures. Using the same input figure comparing CHEETAH and GAZELLE systems, two different caption types are generated: With context from the mention-paragraph and OCR text, the model generates a detailed technical caption describing the speedup analysis and communication cost comparison. Without context, the model produces only a generic description, "A graph plot." This demonstrates how incorporating contextual information from the mention-paragraph, and OCR text enables more informative and technically accurate figure captions.

agrams

- Required contextual information comes from research papers, particularly mention-paragraphs
- OCR text within figures provides additional technical details
- Integration of visual, textual, and OCR information is essential

2. Caption Requirements
    - Must interpret data and explain methodological insights
    - Should highlight key findings within research context

These domain-specific requirements necessitate the development of specialized architectures capable of integrating visual features, OCR-extracted text, and research context through advanced mechanisms such as multimodal transformers. The complex nature of scientific figure captioning motivates two primary research directions:

1. Developing multimodal architectures integrating context information from multiple sources.

   As shown in Figure 1.6, modern approaches combine visual features from the figure itself, OCR-extracted texts within the figure, and relevant sections from the research paper (mention-paragraphs) to generate comprehensive captions. This multimodal integration is essential because scientific figures often contain dense, technical information that can only be properly interpreted within the broader research context.

2. Accessing the relative importance of visual and textual context in generating scientific figure captions.

   While visual elements provide the foundation for caption generation, textual context from the surrounding research paper often provides crucial technical details. For instance, in Figure 1.6, the mention-paragraph provides essential context about comparing CHEETAH and GAZELLE systems that would be difficult to infer from the visual content alone, which raises important questions about how captioning models should balance and integrate these different information modalities.

With the growth of scientific publications these days, communicating research findings in an informative and clear manner become essential to improve the sharing of scientific knowledge. Automated scientific figure caption generation has the potential to help researchers. However, developing these tools presents complex challenges. Models must learn to integrate diverse context sources meaningfully, from visual elements in figures to surrounding textual explanations, while maintaining the technical precision that scientific communication demands. By tackling these challenges, we can work toward systems supporting researchers in making their work more accessible and understandable to the broader scientific community.

1.3.3 Task Summarization

This thesis examines two types of context-driven image captioning that differ in how context manifests but share fundamental technical challenges. Table 1.1 summarizes the distinct characteristics of news image captioning and scientific figure captioning.

Characteristics	**News Image Captioning**	**Scientific Figure Captioning**
Visual Context	Visual content with journalistic context	Data visualizations (data visualizations, graphs, and plots) with research context
Textual Context	News articles providing narrative and background	Mention paragraphs and OCR text providing technical context
Caption Purpose	Visual description with news value	Technical interpretation of data and research findings

Table 1.1: Comparison between characteristics of news image captioning and scientific figure captioning

News image captioning requires integrating visual content from news images with journalistic context from news articles. The context provides a narrative background that helps convey the visual scene and its broader news significance.

Scientific figure captioning involves interpreting technical visualizations by combining figure content with domain-specific knowledge from research papers, particularly through mention-paragraphs and OCR-extracted texts. Despite their domain-specific differences, both tasks share a core technical challenge: developing multimodal architectures capable of effectively integrating context from multiple sources.

1.3.4 Contribution

The two studies presented in this thesis provide complementary perspectives on context-driven caption generation while sharing common methodological foundations. Both studies investigate:

1. The relative contributions of visual and textual information in context-driven caption generation tasks:
    - Analysis of how different modalities influence caption quality
    - Comparison of visual versus textual context importance
    - Impact of domain-specific contextual elements
2. Methods for effective multimodal integration:
    - Development of architectures for combining multiple information sources
    - Approaches for handling domain-specific challenges

Through these investigations, this thesis advances the understanding of how multimodal contextual information can be effectively leveraged for caption generation. As summarized in Table 1.2, this research makes several contributions across empirical analysis, methodological development, and resource creation. The empirical findings demonstrate the critical role of contextual information and quantify the relative impact of different modalities. The methodological contributions establish effective approaches for multimodal integration through transformer-based architectures. The enhanced datasets for news images and scientific figures also provide valuable resources that enable further research into context-driven caption generation.

Category	Details
Empirical Analysis	1. Demonstrating the critical role of context information in both news image and scientific figure caption generation through systematic experimentation
	2. Quantifying the relative impact of textual and visual modalities on caption quality through comprehensive evaluation metrics
	3. Identifying key factors that influence caption generation quality across different domains
Methodological Analysis	1. Demonstrating the effectiveness of transformer-based architectures for multimodal context integration
	2. Developing novel attention mechanisms for cross-modal feature fusion
	3. Establishing best practices for handling domain-specific challenges in context integration
	4. Providing insights into model design principles for context-driven caption generation
Resource Contributions	1. SciCap+: An enhanced scientific figure dataset incorporating paragraph mentions and OCR content, enabling research into context-driven scientific figure captioning
	2. A refined version of the news-image captioning dataset with complete article text, facilitating more comprehensive analysis of news context integration
	3. Implementation of evaluation frameworks for both tasks, supporting future research in this direction

Table 1.2: Summary of Research Contributions: Analysis of roles of contextual information, development of transformer-based architectures, and creation of enhanced datasets for both news image and scientific figure caption generation

1.4 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 provides essential background knowledge on sequence-to-sequence models, attention mechanisms, and the Transformer architecture that forms the foundation for the proposed approaches. The chapter introduces evaluation metrics, including traditional metrics like BLEU and METEOR, and specialized metrics for image captioning, such as CIDEr and CLIPScore. This technical foundation supports the methodological choices and evaluation strategies in subsequent chapters.

Chapter 3 presents a comprehensive literature review spanning image captioning, news image captioning, and scientific figure captioning. The review demonstrates how the field has evolved and the role of contextual information in news image and scientific figure captioning, positioning contributions of this dissertation within the broader research landscape.

Chapter 4 examines news image captioning, introducing a novel Transformer-based architecture that effectively integrates visual and textual features. Through extensive experiments and human evaluation, the study demonstrates that while textual context from news articles provides primary information for generating news captions, integrating visual features further improves caption quality. This work represents an advancement in news image captioning methodology and architectural design.

Chapter 5 investigates scientific figure captioning, introducing the enhanced SciCap+ dataset and reframing the task as a knowledge-augmented image captioning problem. The chapter presents evidence that effective scientific figure captioning requires sophisticated integration of multiple context sources, including mention-paragraphs and OCR-extracted text. The human evaluation reveals important insights about the challenges in scientific figure caption generation, even for domain experts, while demonstrating the effectiveness of the proposed multimodal approach.

Chapter 6 synthesizes the findings from both studies to present a unified understanding of context-driven caption generation. While textual context plays a primary role in context-driven image captioning tasks, optimal performance requires the effective integration of multiple modalities. The chapter also articulates the main contributions of this dissertation and future work.

2 Background Knowledge

2.1 Sequence-to-Sequence Model

Natural language processing tasks can be formulated as sequence-to-sequence (seq2seq) problems, where both the input and output are sequences of arbitrary length. For example, in machine translation, we map a sequence of words $x_1, ..., x_n$ from a source language to a sequence of words $y_1, ..., y_m$ in a target language. Similarly, text summarization involves transforming an input document represented as a sequence into a summary.

A key challenge in developing neural approaches for seq2seq tasks is that traditional feed-forward neural networks require fixed-dimensional inputs and outputs specified in advance. To address this limitation, Sutskever [57] introduced the encoder-decoder architecture that can handle variable-length sequences end-to-end.

The encoder-decoder framework consists of two main components: An encoder that processes the input sequence:

$$x = (x_1, ..., x_n) \tag{2.1}$$

And compresses it into a fixed-dimensional context vector c:

$$c = \text{Encoder}(x_1, ..., x_n) \tag{2.2}$$

A decoder that generates the output sequence $y = (y_1, ..., y_m)$ conditioned on c:

$$p(y|x) = \prod_{t=1}^{m} p(y_t|c, y_1, ..., y_{t-1}) \tag{2.3}$$

In the original seq2seq model [57], the encoder and decoder are implemented as Long Short-Term Memory (LSTM) networks. The encoder LSTM encodes the

input sequence and produces a fixed-dimensional representation for the decoder to decode. The decoder LSTM then generates the output sequence one token at a time, conditioned on this representation and previously generated tokens.

This architecture has since been enhanced with attention that allows the decoder to dynamically focus on different parts of the input sequence rather than rely on a single fixed context vector. Modern seq2seq models often employ Transformer architectures [60], primarily replacing LSTMs as the default choice for sequence modelling tasks.

2.2 Attention Mechanism

Human cognitive function includes the ability to selectively attend to relevant information within the current context. This biological attention mechanism has inspired the development of attention modules in modern deep-learning architectures.

Traditional sequence-to-sequence models compressed all input information into a fixed-length context vector, creating an information bottleneck that limits model performance, especially for long sequences. The attention mechanisms allow the model decoder to dynamically attend to different input parts at each generation step. When generating each output token, the attention module computes alignment scores between the current decoder state and all input elements, producing a weighted sum emphasizing the most relevant input features.

The attention mechanism provides several advantages: It helps preserve long-range dependencies and allows the model to extract relevant information based on the generation context adaptively. The success of attention in sequence modelling ultimately led to the development of self-attention and the Transformer architecture [60], revolutionizing natural language processing and computer vision tasks.

Given source sequence $\boldsymbol{X} = (x_1, ..x_n)$ to a target sequence $\boldsymbol{Y} = (y_1, ...y_m)$. For an encoder-decoder model, let $\boldsymbol{E} = [\boldsymbol{e_0}, ..., \boldsymbol{e_n}]$ denote the encoder states of $\boldsymbol{X}$, and $\boldsymbol{s_{t-1}}$ represent the decoder state at time $t-1$. To generate the decoder state $\boldsymbol{s_t}$ at time t, the attention mechanism first computes an attention distribution $\alpha_{t,i}$ over $\boldsymbol{E}$, then uses it to calculate a weighted average of $\boldsymbol{E}$. The attention distribution is computed as:

$$\alpha_{t,i} = \text{softmax}\left(\text{score}\left(\boldsymbol{s}_{t-1}, \boldsymbol{e_i}\right)\right) \tag{2.4}$$

$$\text{attention}(\boldsymbol{E}, \boldsymbol{s_{t-1}}) = \sum_{i=1}^{n} \alpha_{t,i} \boldsymbol{e_i} \tag{2.5}$$

where *score* is the score function.

There are several score functions for computing attention:

1. **Additive Attention** [4] uses trainable matrices $\boldsymbol{v^T}, \boldsymbol{W}, \boldsymbol{U}$:

$$\text{score}\left(\boldsymbol{s}_{t-1}, \boldsymbol{e}i\right) = \boldsymbol{v^T} \tanh\left(\boldsymbol{W}\boldsymbol{s}t - 1 + \boldsymbol{U}\boldsymbol{e}_i\right) \tag{2.6}$$

2. **Dot-Product Attention** [40] directly computes similarity:

$$\text{score}\left(\boldsymbol{s}_{t-1}, \boldsymbol{e_i}\right) = \boldsymbol{s}_{t-1}^T \boldsymbol{e}_i \tag{2.7}$$

3. **Scaled Dot-Product Attention** [60] adds scaling factor:

$$\text{score}\left(\boldsymbol{s}_{t-1}, \boldsymbol{e}_i\right) = \frac{\boldsymbol{s}_{t-1}^T \boldsymbol{e}_i}{\sqrt{D}} \tag{2.8}$$

   where D represents input dimension

4. **General Attention** [40] introduces a trainable matrix $\boldsymbol{W}$:

$$\text{score}\left(\boldsymbol{s}_{t-1}, \boldsymbol{e}_i\right) = \boldsymbol{s}_{t-1}^T \boldsymbol{W} \boldsymbol{e}_i \tag{2.9}$$

2.3 Transformer Model

Recent advances in neural architectures have revolutionized how we model sequential data, with the Transformer [60] emerging as the de facto architecture across modern deep learning fields, from natural language processing to computer vision.

Unlike traditional recurrent architectures that process sequences token by token, the Transformer eliminates recurrence, instead relying solely on attention mechanisms to capture dependencies between tokens. The self-attention mechanism has proven helpful in capturing local and long-range dependencies, as each position can directly attend to all other positions in the sequence.

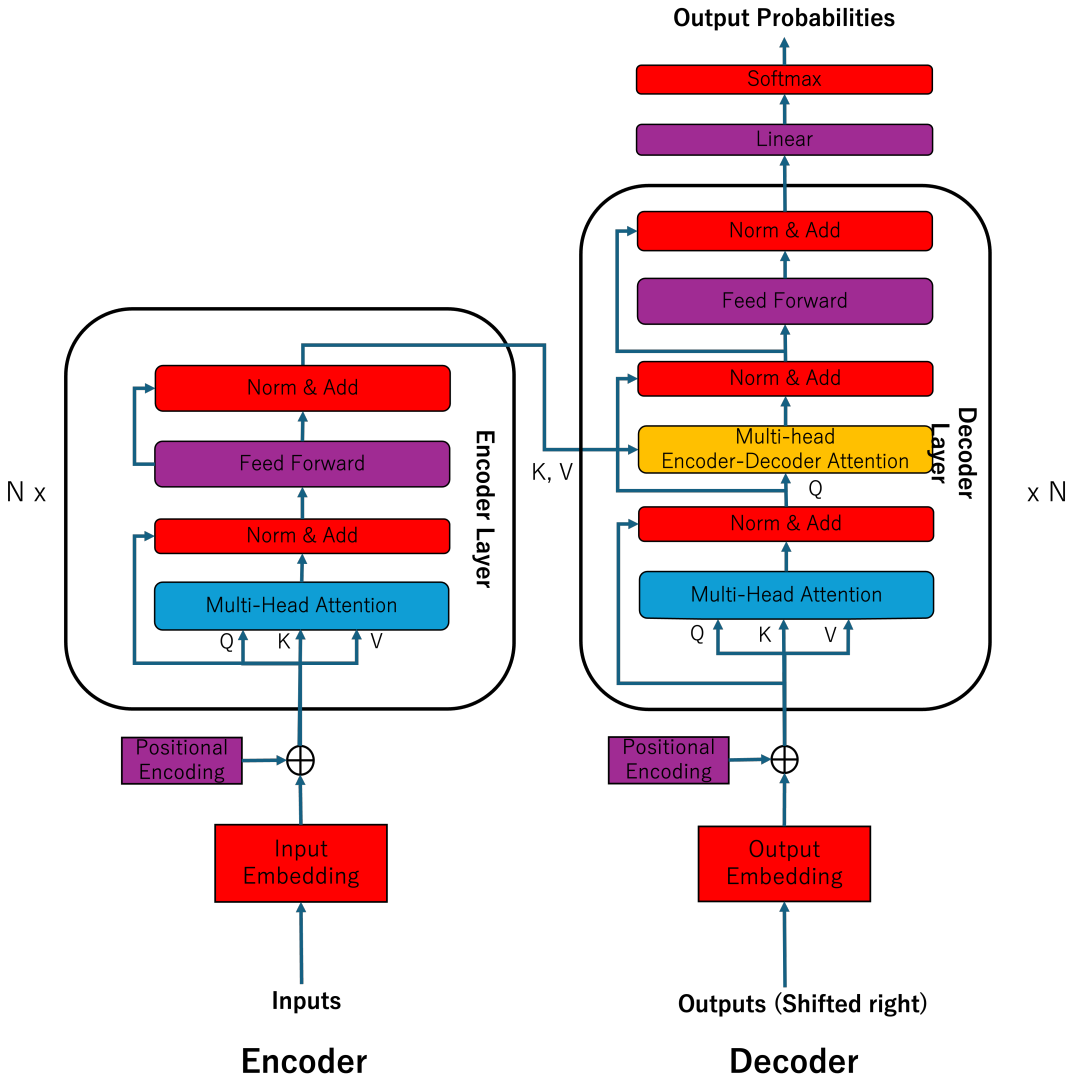


Figure 2.1: The architecture of the Transformer model [60] consists of two main components: the encoder (left) and the decoder (right). The encoder processes input sequences through a stack of modules, each comprising multi-head self-attention and feed-forward networks. In addition to similar modules, the decoder includes an encoder-decoder attention layer, enabling it to attend to output from the encoder.

The Transformer model can process entire sequences simultaneously rather than sequentially. The efficient computation and superior performances on downstream tasks of the Transformer have led to breakthrough performances, spawning numerous variants and extensions that have pushed forward deep learning research.

The remarkable success of the Transformer architecture can be attributed to several key advantages:

1. **Long-range Dependency Modeling**: Through its self-attention mechanism, the Transformer captures relationships between distant elements in a sequence. Therefore, the Transformer can learn complex dependencies without suffering from the vanishing gradient problems common in recurrent architectures.

2. **Computational Efficiency**: While RNNs must process tokens sequentially due to their recurrent nature, the Transformer can process all elements simultaneously, leveraging modern hardware accelerators effectively. This parallelization reduces training time and enables scaling to longer sequences and larger datasets.

These benefits have led to the Transformer becoming the foundation for numerous state-of-the-art models in natural language processing [20, 11], computer vision [21, 12], and multimodal learning [38, 50].

2.3.1 Positional Encoding

The Transformer architecture utilizes attention mechanisms; therefore, an explicit way to capture sequence ordering is needed. Positional encoding addresses this by incorporating absolute or relative position information of input sequence symbols directly into the embeddings.

Figure 2.1 illustrates how positional encoding combines with both input and output embeddings at the base of the encoder and decoder. Positional encoding computes as:

$$\mathrm{PE}\,(pos, 2i) = sin(pos/10000^{2i/dmodel}) \tag{2.10}$$

$$\mathrm{PE}\,(pos, 2i+1) = cos(pos/10000^{2i/dmodel}) \tag{2.11}$$

Here, *pos* represents the position and i indicates the dimension. Both positional encodings and input/output embeddings share the same dimensionality d_{model}.

2.3.2 Self-Attention Mechanism

As Figure 2.1 illustrated, The Transformer has an encoder-decoder architecture. Both encoder and decoder components are built from stacks of self-attention layers and position-wise feed-forward networks. The attention mechanism computes scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.12}$$

where $Q \in \mathbb{R}^{n \times d_k}$, $K \in \mathbb{R}^{n \times d_k}$, and $V \in \mathbb{R}^{n \times d_v}$ are queries, keys, and values respectively. The scaling factor $\sqrt{d_k}$ prevents the dot products from growing too large in magnitude.

As shown in Figure 2.2, Multi-head attention extends this by applying h parallel attention functions:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{2.13}$$

where each head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.14}$$

with learned projection matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

2.3.3 Residual Connection

As the depth of neural networks increases, accuracy tends to saturate and degrade, and performance may even degrade due to optimization challenges. even with techniques like normalized initialization and intermediate normalization that address vanishing/exploding gradients. The vanishing/exploding gradient degradation problem motivated He et al. [24] to introduce residual connections to train deep networks effectively.

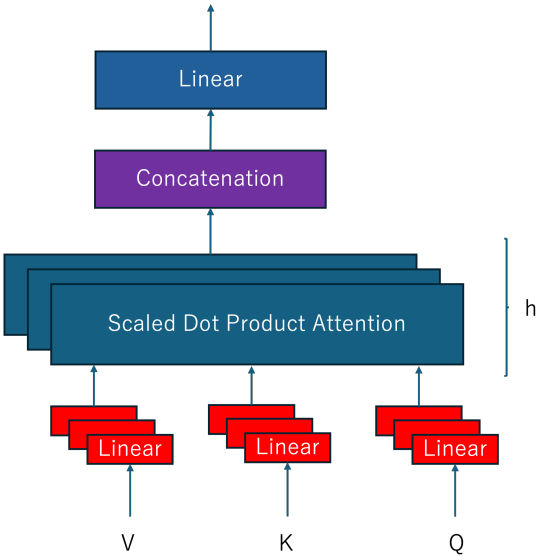


Figure 2.2: Architecture of multihead attention with h parallel attention heads. The queries (Q), keys (K), and values (V) are linearly projected h times with different learned projections. Each head computes scaled dot-product attention independently on the projected vectors. The outputs from all heads concatenate and linearly project once more to produce the final output. This allows the Transformer model to attend different representation subspaces of the input simultaneously.

As shown in Figure 2.3, each residual connection block computes:

$$y = F(x, W_i) + x \tag{2.15}$$

Here, x represents the input to the block, y represents the output, and $F(x, W_i)$ defines the residual mapping. Adding x creates a shortcut connection that allows gradients to flow directly through the network.

These skip connections make deep networks easier to optimize by providing direct paths for gradient flow during backpropagation. When the network faces degradation issues, these connections enable it to bypass layers that do not improve performance. This design makes it possible to train very deep networks with hundreds of layers while maintaining high performance.

The Transformer architecture incorporates residual connections in the sublayers of the encoder and decoder blocks, allowing the model to maintain access

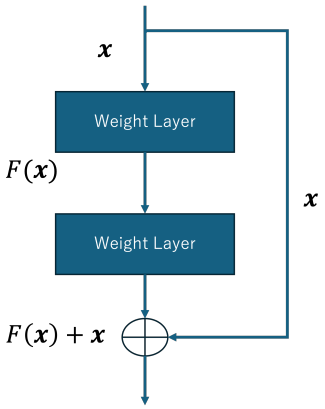


Figure 2.3: A residual connection block in a neural network. The input x flows through two weight layers that compute $F(x)$, which is then added to the original input x through a skip connection. This architecture helps address the degradation problem in deep neural networks by allowing gradients to flow directly through the network.

to lower-level features throughout the network depth. These connections help stabilize training and improve the flow of information through the multi-layer Transformer structure.

2.3.4 Feed-forward Networks

A fully connected feed-forward network appears in each encoder and decoder layer, following this formula:

$$\text{FFN}(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{2.16}$$

These feed-forward networks consist of two linear transformations with ReLU activation functions in between. Each layer uses distinct parameter matrices W_1 and W_2. The networks have the same input and output dimension of d_{model}, with an inner hidden layer dimension of 2048.

The Linear and Softmax Layer

After the stack of decoder layers, a linear and softmax layer transforms decoder outputs into token probabilities. The linear layer implements a learned fully

connected feed-forward network that projects decoder outputs into a vector with dimension of vocabulary size.

The softmax layer then converts this logits vector into probabilities. The model selects the index with the highest probability and generates its corresponding token as output.

2.4 Evaluation Metrics

While human evaluation provides the most reliable assessment of caption quality, it is time-consuming and costly. Furthermore, the subjective nature of caption evaluation means human judgments cannot be easily replicated [46]. Given the large-scale nature of our experiments, automatic evaluation metrics offer an efficient and reproducible way to assess model performance. We employ five widely used automatic evaluation metrics to evaluate caption quality quantitatively:

1. BLEU [46], METEOR [6] [19], developed initially for machine translation evaluation, measure the overlap between reference and generated text using n-gram precision and recall respectively. METEOR additionally handles synonyms and performs stemming for more flexible matching.

2. ROUGE [37], designed for evaluating text summarization, computes recall-oriented overlap statistics between generated and reference texts, with variants for different n-gram lengths.

3. CIDEr [61] and SPICE [2] were developed explicitly for image captioning evaluation. CIDEr uses TF-IDF weighted n-gram similarities to capture consensus between multiple references, while SPICE operates on scene graphs to evaluate semantic propositional content.

These metrics evaluate caption quality from various perspectives, but they all fundamentally measure the overlap between generated and reference captions. Each offers a unique focus:

1. BLEU and METEOR emphasize precision.

2. ROUGE prioritizes recall.

3. CIDEr focuses on important words and consensus

4. SPICE captures semantic accuracy

Combining multiple metrics provides a well-rounded assessment of caption quality.

A recently proposed reference-free metric, CLIPScore, leverages the pre-trained CLIP model to evaluate the semantic similarity between an image and its generated captions to complement reference-based and overlap-based evaluation metrics. CLIPScore has several advantages: it can evaluate captions without references, provides robust assessment across diverse caption styles and content, and complements traditional reference-based and n-gram overlap metrics. The approach demonstrates a strong correlation with human judgments while being more straightforward and flexible than conventional evaluation methods.

2.4.1 Bilingual Evaluation Understudy (BLEU)

BLEU (Bilingual Evaluation Understudy) [46] is a precision-based metric designed initially for machine translation evaluation that quantifies the similarity between machine-generated translations and human references. While initially developed for translation tasks, it has been widely adopted in other text generation tasks, including image captioning.

High-quality generated text should share substantial word sequences with reference texts. BLEU implements this by measuring the overlap between n-grams in the candidate and reference texts. For a candidate text c and reference text r, the n-gram precision p_n is calculated as:

$$p_n = \frac{\sum_i \sum_{ngram \in c} \text{Count}_{\text{clip}}(ngram)}{\sum_i \sum_{ngram' \in c'} \text{Count}(ngram')} \tag{2.17}$$

where $Count_{clip}(ngram)$ implements a clipping mechanism to prevent over-generation problem:

$$\text{Count}_{clip}(ngram) = \min(\text{Count}_c(ngram), max_{j \in refs}\text{Count}_{r_j}(ngram)) \tag{2.18}$$

Here, $\text{Count}_c(ngram)$ is the number of times an n-gram appears in the candidate text, and $\text{Count}_{r_j}(ngram)$ is its count in the reference text j.

The final BLEU score combines these n-gram precisions using:

$$\text{BLEU} = \text{BP} \cdot exp(\sum_{n=1}^{N} w_n log(p_n)) \tag{2.19}$$

Where BP is a brevity penalty term:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{2.20}$$

Where c is the length of the candidate translation, r is the effective reference corpus length. This penalty prevents very short candidates from receiving disproportionately high scores.

This penalty ensures that very short candidates do not receive disproportionately high scores. $w_n = 1/N$, where N is the length of n-grams (typically up to $N = 4$). The choice of N can be adjusted depending on the application.

2.4.2 Metric for Evaluation of Translation with Explicit ORdering (METEOR)

The Metric for Evaluation of Translation with Explicit ORdering (METEOR) [6, 19] calculates scores using the harmonic mean of precision and recall at the unigram level. Unlike BLEU, METEOR incorporates semantic and syntactic information, enabling a more comprehensive evaluation beyond exact n-gram precision.

METEOR addresses several limitations of the BLEU metric by providing a more comprehensive and flexible evaluation framework for machine translation:

1. **Enhanced Word Matching**: Unlike the strict requirement of BLEU for exact n-gram matches, METEOR implements a multi-level matching system. This includes exact matches, stemmed matches, synonym matches through WordNet, and paraphrase recognition, allowing for a more comprehensive evaluation of semantic equivalence.

2. **Intelligent Word Order Handling**: METEOR addresses rigid n-gram matching of BLEU with a more sophisticated approach to word order. Its fragmentation penalty mechanism can accommodate legitimate variations in word order while penalizing disordered translations that compromise meaning or readability.

3. **Balanced Evaluation Approach**: While BLEU relies exclusively on precision-based scoring, METEOR introduces a balanced approach that combines precision and recall. By emphasizing recall, METEOR better captures whether the generated text preserves all essential information from the reference text.

Enhanced Word Matching

METEOR computes alignment scores between candidate and reference sentences through a systematic matching process. The algorithm first establishes unigram alignments with the constraint that each unigram can be mapped to at most one unigram in the other sentence. To identify potential matches, METEOR employs four distinct matching stages in sequence:

1. **Exact Matching**: Identifies unigrams with identical surface forms.
2. **Stemming-based Matching**: Matches unigrams that share the same stem using the Snowball Stemmer [49].
3. **Synonym Matching**: Identifies synonymous unigrams using WordNet [42].
4. **Paraphrase Matching**: Recognizes phrase-level equivalences using language-appropriate paraphrase tables.

Score Calculation

After collecting all possible unigram mappings, METEOR selects the optimal alignment by applying the following criteria in decreasing order of priority:

1. **One-to-One Mapping**: Each unigrams has exactly one matching.
2. **Maximum Coverage**: Maximizes the total number of aligned unigrams.
3. **Minimal Fragmentation**: Minimizes the number of chunks, where a chunk represents a sequence of contiguous and identically ordered unigram mappings in both texts [19].
4. **Minimal Distance**: Minimizes the total absolute distance between the starting positions of corresponding mappings

In METEOR, scoring processing uses frequency-based classification to distinguish between content words and function words. Function words are identified as unigrams with a relative frequency exceeding 10^{-3} in the corpus. This distinction is important for weighted scoring calculations.

After establishing unigram alignments between a candidate-reference pair, METEOR categorizes matched words into:

Content words in candidate (c_c) and reference (r_c) Function words in candidate (c_f) and reference (r_f)

For each matcher i, METEOR counts:

Content word matches: $matcher_i(c_c)$ and $matcher_i(r_c)$. Function word matches: $matcher_i(c_f)$ and $matcher_i(r_f)$.

The weighted precision P and recall R are then calculated as:

$$\begin{aligned} P &= \frac{\sum_i w_i \cdot \left(\delta \cdot matcher_i(c_c) + (1-\delta) \cdot matcher_i\left(c_f\right)\right)}{\delta \cdot |c_c| + (1-\delta) \cdot \left|c_f\right|} \\ R &= \frac{\sum_i w_i \cdot \left(\delta \cdot matcher_i(r_c) + (1-\delta) \cdot matcher_i\left(r_f\right)\right)}{\delta \cdot |r_c| + (1-\delta) \cdot \left|r_f\right|} \end{aligned} \tag{2.21}$$

where: w_i represents the weight for matcher i, δ_c and δ_f are parameters controlling the relative importance of content and function words $|c_c|$, $|c_f|$, $|r_c|$, $|r_f|$ represent the total counts of content and function words

This weighted approach ensures that content words, which carry more semantic meaning, can be given higher importance in the final score calculation.

2.4.3 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [37] is a family of recall-based automatic evaluation metrics designed initially for assessing machine-generated text summaries.

The core intuition behind ROUGE is to evaluate the quality of generated text by measuring its overlap with human-written reference texts. Two of the most widely used versions are ROUGE-N and ROUGE-L. ROUGE-N measures n-gram overlap between the generated text and references, capturing the proportion of n-grams from the reference in the generated text.

Though initially developed for text summarization, ROUGE metrics have been widely adopted in other natural language generation tasks, including image cap-

tioning. These help assess how well-generated outputs match human references in content and structure.

Based on the longest common subsequence (LCS), ROUGE-L is more flexible as it automatically identifies the longest matching word sequences without requiring consecutive matches. This allows it to capture similar sentence-level word order while permitting match gaps. ROUGE-L can better handle paraphrasing and structural differences between generated and reference texts.

ROUGE-N

Unlike BLEU, which uses precision, ROUGE-N evaluates n-gram overlap using recall-based computation. The formal calculation as defined in [37] is:

$$\text{ROUGE-N} = \frac{\sum_{C \in Reference\ \ Summaries} \sum_{gram_n \in C} \text{Count}_{\text{match}}(gram_n)}{\sum_{C' \in Reference\ \ Summaries} \sum_{gram'_n \in C'} \text{Count}(gram'_n)} \tag{2.22}$$

Here, n represents the length of the gram, and $\text{Count}_{\text{match}}(gram_n)$ denotes the maximum number of times an n-gram co-occurs in a candidate text and the reference texts. When evaluating a candidate text s against multiple reference text R, ROUGE-N selects the highest score from all pair-wise comparisons:

$$\text{ROUGE-N} = argmax_i \, \text{ROUGE-N}(r_i, s), \ r_i \in R \tag{2.23}$$

ROUGE-L

ROUGE-L [37] extends the evaluation framework by incorporating the longest common subsequence. LCS offers two key advantages: it inherently captures sentence-level word ordering and structure while automatically identifying the longest matching sequences without requiring explicit n-gram size specification.

For sentence-level assessment, given a reference summary r of length m and a candidate summary c of length n, ROUGE-L computes three scores:

$$\begin{aligned} R_{lcs} &= \frac{\text{LCS}(r, c)}{m} \\ P_{lcs} &= \frac{\text{LCS}(r, c)}{n} \\ F_{lcs} &= \frac{(1 + \beta^2)\, R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \end{aligned} \tag{2.24}$$

where $\mathrm{LCS}(r, c)$ computes the length of the longest common subsequence between r and c. The parameter β is typically assigned a large value to emphasize recall.

The final ROUGE-L score is determined by the F-measure F_{lcs} combining precision P_{lcs} and recall R_{lcs}. For corpus-level evaluation with reference summaries R containing u sentences totaling m words, and candidate summaries C with v sentences totaling n words, ROUGE-L is calculated as:

$$\begin{aligned} R_{lcs} &= \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{m}, r_i \in R \\ P_{lcs} &= \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{n}, r_i \in R \\ F_{lcs} &= \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \end{aligned} \tag{2.25}$$

2.4.4 Consensus-based Image Description Evaluation (CIDEr)

Consensus-based Image Description Evaluation (CIDEr) [61] is explicitly designed to assess image caption quality by measuring how well a candidate caption aligns with the consensus of human-written reference captions. Unlike traditional metrics that rely on direct matching, CIDEr captures human consensus by weighing n-grams based on their frequency across reference captions while penalizing n-grams commonly used in the larger corpus.

The metric employs term frequency-inverse document frequency (TF-IDF) weighting to emphasize words important for describing a specific image while downweighing words frequently appearing across all image descriptions. This makes CIDEr particularly suitable for image captioning tasks as it can better distinguish informative descriptions from generic ones.

For a candidate caption c_i and a set of reference captions s_{ij} for image i, CIDEr first computes TF-IDF weighted n-gram vectors and then measures their similarity through cosine similarity. This process gives higher weight to n-grams specific to the described image while reducing the impact of commonly occurring words and phrases.

When evaluating captions that include specific words or named entities, CIDEr has demonstrated a strong correlation with human judgments. This makes it useful for assessing caption quality in specialized domains, such as scientific figure captioning and news image captioning, where specific terms frequently appear.

CIDEr metric processes candidate and reference captions by first stemming all words and representing each sentence as n-gram sequences, it then uses TF-IDF to appropriately weight n-grams based on their importance. For an n-gram w_k, its TF-IDF weight $g_k(s_{ij})$ is computed as:

$$\begin{aligned} g_k(s_{ij}) &= \mathrm{TF}(w_k)\,\mathrm{IDF}(w_k) \\ g_k(s_{ij}) &= \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)}\right) \end{aligned} \tag{2.26}$$

where Ω represents the complete n-gram vocabulary and $|I|$ denotes the total number of images in the dataset. The Term Frequency (TF) component $\frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})}$ measures how frequently an n-gram appears in a reference sentence s_{ij}, with $h_k(s_{ij})$ counting the occurrences of w_k. For candidate sentences, $h_k(c_j)$ similarly denotes n-gram appearances.

The Inverse Document Frequency (IDF) component $\log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)}\right)$ assigns higher weights to rare n-grams, where $\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)$ counts images whose reference sentences contain the n-gram w_k. The CIDEr score for a specific n-gram order n is computed as the average cosine similarity between candidate and reference sentences:

$$\mathrm{CIDEr}\, n(ci, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g^n}(c_i) \cdot \boldsymbol{g^n}(s_{ij})}{|\boldsymbol{g^n}(c_i)|\,|\boldsymbol{g^n}(s_{ij})|} \tag{2.27}$$

where $\boldsymbol{g^n(\cdot)}$ represents a vector of TF-IDF weights for all n-grams of order n, and $|\boldsymbol{g^n}(\cdot)|$ denotes its magnitude. The final CIDEr score combines n-grams of different orders ($N = 1$ to 4) with equal weights:

$$\mathrm{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n\, \mathrm{CIDEr}\, n(ci, S_i) \;\; w_n = 1/N \tag{2.28}$$

This multi-order n-gram approach allows CIDEr to capture both grammatical structures and semantic properties of the captions.

2.4.5 Semantic Propositional Image Caption Evaluation (SPICE)

Semantic Propositional Image Caption Evaluation (SPICE) [2] introduces a novel approach to caption evaluation by focusing on semantic propositional content,

which more closely aligns with human judgment criteria. Unlike traditional n-gram based metrics, SPICE transforms captions into scene graphs (structured semantic representations that capture objects, their attributes, and relationships between objects).

The conversion of a caption into a scene graph follows a two-stage process:

1. The caption undergoes syntactic parsing using a dependency parser to construct a dependency tree
2. The dependency tree is mapped to a scene graph where:
    - Nodes represent objects mentioned in the caption
    - Attributes are properties associated with these objects
    - Edges represent relationships between objects

Given an image I, its candidate caption c and reference caption set Ref, SPICE computes an F-score based on the overlap of semantic propositions:

$$\text{SPICE}(c, Ref) = \frac{2 \cdot P \cdot R}{P + R} \tag{2.29}$$

where P and R are precision and recall of the scene graph of candidate caption against scene graphs of reference caption set.

$$P = \frac{|\,\text{T}(G(c)) \otimes \text{T}(G(Ref))|}{|\,\text{T}(G(c))|} \tag{2.30}$$

$$R = \frac{|\,\text{T}(G(c)) \otimes \text{T}(G(Ref))|}{|\,\text{T}(G(R))|} \tag{2.31}$$

$G(c)$ denotes the scene graph of the candidate caption c, $G(Ref)$ denotes the scene graph of reference sentence set Ref. The $G(R)$ is the union of the scene graph of each reference sentence $s_i \in S$ by combining synonymous.

For a caption c, SPICE returns scene graph $G(c)$:

$$G(c) =< O(c), E(c), K(c) > \tag{2.32}$$

where $O(c)$: A set of objects in the caption, and $E(c)$: A set of hyper-edges representing relationships between objects. $K(c)$: A set of attributes describing the objects.

The function T returns the logical tuples from a scene graph G(c) of caption c as follows:

$$\mathrm{T}(G(c)) \triangleq O(c) \cup E(c) \cup K(c) \tag{2.33}$$

2.4.6 CLIPScore

CLIPScore [26] presents a reference-free evaluation metric that leverages CLIP [50], a pretrained vision-language model, to evaluate the semantic alignment between images and captions. While traditional metrics like BLEU and METEOR require carefully written reference captions and often fail to recognize diverse but valid descriptions, CLIPScore directly measures image-caption compatibility. The metric computes cosine similarity between image and text embeddings from CLIP, enabling more flexible and comprehensive evaluation.

The implementation demonstrates strong computational efficiency. A single consumer GPU can process approximately 4,000 image-caption pairs per minute. CLIPScore correlates more with human judgments on standard image captioning benchmarks than conventional reference-based metrics. However, the metric shows limitations when evaluating non-literal or context-dependent captions.

CLIPScore builds on CLIP (ViT-B/32), which has been pretrained on 400M image-caption pairs. The approach projects images and captions into a shared embedding space and calculates their cosine similarity. A rescaling operation maps the scores to a 0-1 range. Formally, given an image with a visual CLIP embedding v and a candidate caption with a textual CLIP embedding c, CLIPScore (CLIP-S) is computed as:

$$\text{CLIP-S}(\mathbf{c}, \mathbf{v}) = w \cdot \max\left(\cos(\mathbf{c}, \mathbf{v}), 0\right) \tag{2.34}$$

Where $w = 2.5$ serves as a scaling factor, CLIPScore averages scores across all image-caption pairs for corpus-level evaluation.

The CLIPScore also supports integration with reference captions through RefCLIP-S. This extension processes reference captions through CLIP to obtain vector representations R, then combines them with the base CLIPScore using a harmonic

mean:

$$\text{RefCLIP-S}(\mathbf{c}, \mathbf{R}, \mathbf{v}) = \text{H-Mean}\left(\text{CLIP-S}(\mathbf{c}, \mathbf{v}), \max\left(\max_{\mathbf{r}\in\mathbf{R}} \cos(\mathbf{c}, \mathbf{r}), 0\right)\right) \quad (2.35)$$

3 Related Work

3.1 Image Captioning

Image captioning aims to generate natural language descriptions from images, a representative task in multimodal machine learning. This task requires models to bridge the gap between visual understanding and language generation.

Early approaches to image captioning adopted encode-decoder architecture inspired by neural machine translation [15, 5]. One of the pioneering works in image captioning, Neural Image Caption (NIC) [63], which employs a Convolutional Neural Network (CNN) to encode images into fixed-length vectors that serve as input to a Recurrent Neural Network (RNN) decoder for caption generation. This pioneering work demonstrated the feasibility of end-to-end training for image captioning.

Xu et al. [66] proposed an attention-based model that dynamically aligns image regions with caption words. The proposed model selectively attends to relevant parts of the image when producing each caption word, establishing explicit connections between visual regions and generated words. This attention-based approach improves over previous methods by enabling more contextually appropriate caption generation.

Recent developments have shifted toward Transformer-based architectures [35, 17, 25, 39], demonstrating superior performance on image captioning tasks. The self-attention mechanism in Transformers enables more sophisticated modelling of relationships between visual elements and textual descriptions.

Traditional image captioning methods focus on describing visible content without incorporating broader contextual information. This limitation presents notable challenges in specialized domains such as news articles and scientific publications, where captioning requires understanding visual content and associated context. Our work addresses this limitation by investigating context-driven captioning approaches that explicitly integrate textual context with visual features.

3.2 News Image Captioning

News image captioning remains a relatively understudied area in image captioning. Unlike traditional image captioning tasks focusing solely on visual content, news image captioning requires models to integrate information from images and their associated news articles to generate contextually appropriate captions.

Early approaches [22, 59] adopted a two-stage pipeline: first using an annotation model to identify relevant keywords from the image, then generating caption sentences based on these extracted keywords. More recent studies departed from this staged approach, leveraging deep neural networks to directly model the implicit relationships between images and their corresponding text descriptions [7, 51, 9].

Early research in news image captioning established foundational approaches for integrating textual context with visual content. Feng and Lapata [22] introduced a pioneering two-stage framework that leveraged image-caption-document tuples from news articles. Their method addressed the challenge through two key stages: content selection and surface realization. The content selection stage employed a probabilistic image annotation model based on Latent Dirichlet Allocation (LDA) to learn joint representations across visual and textual modalities. This model assumed images and documents were generated by shared latent topics, enabling the identification of relevant content across modalities. The approach represented images using SIFT descriptors quantized into visual terms while extracting textual features from associated news articles.

For surface realization, they explored both extractive and abstractive generation approaches. The extractive model selected relevant sentences from articles based on topic similarity with image content. At the same time, the abstractive approach generated new captions through a probabilistic framework. Their experimental results demonstrated that integrating visual and textual modalities improved caption quality compared to using either modality alone. While their abstractive approach produced less grammatical output than extraction, it generated more focused descriptions that better captured image content.

Building upon this foundation, Tariq and Foroosh [59] developed an enhanced framework that systematically incorporated contextual information from multiple sources. Their work demonstrated that visual features alone were insufficient for generating informative captions in specialized domains like news media,

where captions must reflect visual content and broader contextual meaning from the news narrative. Their approach also implemented a two-stage pipeline: automatic annotation prediction and caption generation. The annotation stage predicted word annotations by jointly processing visual features and contextual information, with the key innovation of projecting all information sources into a common probability space. This enabled an effective combination of heterogeneous information types through unified processing and fusion. The generation stage then employed an extractive approach to select appropriate sentences based on predicted annotations and contextual relevance.

Batra et al. [7] proposes a deep neural network architecture that addresses the news image captioning task. Their methodology begins by encoding sentences from news articles using a pre-trained order-embedding model while simultaneously processing images through a pre-trained VGGNet CNN. These encodings are projected into a common semantic space, allowing for direct comparison between textual and visual elements. The encoded information is then processed through an LSTM network, which generates a vector representation capturing both textual and visual semantics. The final caption is selected by identifying the sentence from the original article with the highest cosine similarity with this generated vector.

The effectiveness of this approach was demonstrated through both automated metrics and human evaluation on a BBC News dataset [22]. The system surpassed traditional LDA-based methods in automated metrics. For human evaluation, evaluators preferred captions produced by the proposed model over baselines. Additionally, human evaluators did not prefer 32.91% of captions evaluated, underscoring the need for further improvement in news image caption generation.

Although this research represents the possibility of using encoder-decoder architecture in news image captioning, it is limited to extractive summarization, which involves selecting existing sentences rather than generating novel captions.

Move beyond extraction methods, Ramisa et al. [51] investigated two main architectural approaches for caption generation. Their first approach utilized an LSTM with fixed features, combining Word2Vec representations of article text with visual features extracted from VGG19 and Places networks. The second approach implemented an end-to-end learning system that integrated CNN-based image processing with LSTM-based text generation. Their experiments demonstrated that combining textual and visual features improved performance, with

the best results (BLEU-3 and BLEU-4) achieved when using both VGG19 and Places features alongside textual information.

Their proposed model achieved relatively modest performance compared to similar models applied to conventional image captioning tasks, highlighting the challenge of the news image captioning task. This work indicates we need further research to bridge the gap between visual content in news images and contextually rich news captions.

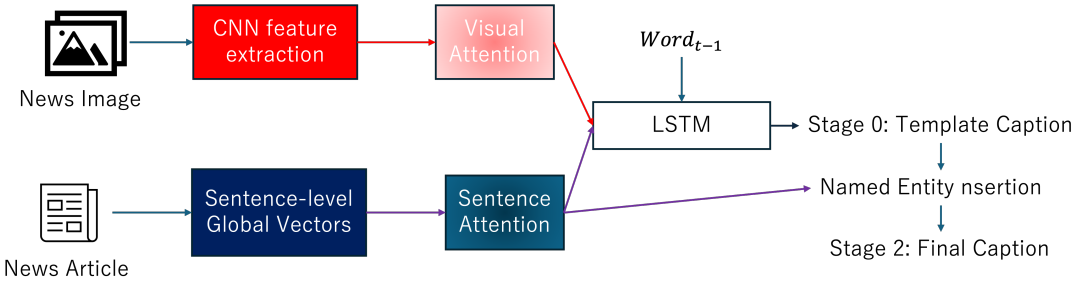


Figure 3.1: Overall architecture of the news-image captioning proposed by Biten et al. [9]

Biten et al. [9] advanced the field by introducing a two-stage model utilizing template completion for news image caption generation. Their architecture incorporated two encoders and an LSTM component: a CNN-based image encoder extracted visual features while a text encoder processed article text at the sentence level. The first stage employed a long short-term memory (LSTM) network to combine textual and visual features, producing template captions with placeholders for named entities. The second stage completed these captions by replacing placeholders with named entities extracted from the article text.

Previous approaches to news image captioning have primarily relied on two-stage architectures or LSTM-based models. [22] and [59] introduced foundational two-stage pipelines that extract keywords before generating captions. More recent work by [9] also employed a two-stage approach using an LSTM to generate template captions before filling in named entities. While these methods demonstrated the importance of integrating textual and visual information, they were limited by their staged architectures, which could propagate errors between stages. In contrast, our work [68] presents an end-to-end Transformer model that seamlessly integrates textual and visual features through a novel image-

attending module. This unified architecture enables direct caption generation without requiring intermediate templates or stages. Through extensive experiments, we demonstrate that the proposed method significantly outperforms previous state-of-the-art methods [9] across multiple automatic evaluation metrics. The automatic and human evaluation further reveals that while textual features provide primary information for caption generation, integrating visual features through our model contributes meaningfully to generating more descriptive and appropriate captions.

3.3 Scientific Figure Captioning

Compared with natural images, scientific figures differ from two perspectives: 1. Visual content: scientific figures contain complex combinations of data visualizations. 2. Captions: Scientific figure captions communicate specific analytical findings rather than merely describe visual scenes. This inherent complexity has made figure captioning a particularly challenging research area that remains underexplored compared to the extensive work done in natural image captioning.

The evolution of scientific figure understanding research can be traced through several significant contributions. Siegel et al. [53] presents an end-to-end framework for automatically parsing and analyzing figures in research papers. The framework first extracts figures from PDF documents and handles sub-figure separation using an iterative decomposition method. For classification, they leverage CNNs (ResNet-50) to categorize figures into types like graphs, flowcharts, and algorithms, achieving 86% accuracy. The core technical contribution is their approach to figure content analysis, which uses a novel graph-based reasoning system with CNN-based similarity metrics to parse figure elements like axes, legends, and plot data. They formulate plot data extraction as an optimal path-finding problem and train their model using a Siamese network architecture to learn robust feature representations.

To evaluate their system, the authors created a large dataset: Figureseer, consisting of 60,000 figures extracted from research papers across multiple disciplines, with detailed annotations for over 600 graph figures. While individual components like axis detection achieved high accuracy (>90%), the overall end-to-end figure parsing had lower performance (17.3%), highlighting the complexity of the task. They demonstrated the practical utility of their framework through

a query-answering application that allows users to search and analyze results across multiple papers. The work represents an important step toward automatic understanding of scholarly figures, though challenges remain in areas like OCR performance and handling complex plot arrangements.

Kahou et al. [29] introduced FigureQA, a comprehensive dataset designed to advance machine learning capabilities in scientific figure comprehension. The FigureQA dataset comprises over one million question-answer pairs grounded in more than 100,000 synthetic scientific figures. These figures span five common types: line plots, dot-line plots, vertical and horizontal bar graphs, and pie charts. The authors prepared 15 templates to generate questions concerning various relationships between plot elements, such as the maximum and the minimum.

The authors of FigureQA also carefully controlled the data generation process. In FigureQA, the training set includes all 100 colours, but the validation and test sets use new colour-plot combinations not seen during training. They also maintained a balanced distribution of yes/no answers to avoid biases. Additionally, each figure comes with rich annotations, including the underlying numerical data and bounding boxes for all plot elements.

The dataset can be extensible and support curriculum learning, allowing for incremental increases in task complexity as model performance improves. The authors position FigureQA as a crucial first step toward developing models that can understand visual representations of data, similar to how humans can grasp patterns and relationships in figures at a glance.

The authors tested four neural network models on the dataset, with the Relation Network (RN) performed the best, reaching 72.40% accuracy on the alternated colour scheme test set. However, human annotators achieved 91.21% accuracy, showing a significant gap and indicating that models still have much room for improvement. Interestingly, humans performed better on bar graphs than line plots and found questions about smoothness and medians the most challenging.

Based on FigureQA [29], the FigCAP dataset [13] has two variants: FigCAP-H and FigCAP-D. FigCAP-H contains high-level descriptions of figures, while FigCAP-D provides detailed descriptions, including relationships between elements. This research introduced three attention mechanisms. Feature Maps Attention processes basic visual features, while Label Maps Attention specifically handles text labels within figures. The Relation Maps Attention mecha-

nism enables the model to understand and describe relationships between elements in a figure. The model architecture combines these attention mechanisms with a ResNet encoder and LSTM decoder, implementing sequence-level training through reinforcement learning to address the long sequence generation and the exposure bias problems. The experimental results demonstrate the effectiveness of this approach, as the model combining all three attention mechanisms achieved superior performance across multiple evaluation metrics, including CIDEr, BLEU, METEOR, and ROUGE scores.

SciCap [27] is a large-scale dataset designed for scientific figure captioning. The authors constructed this dataset by collecting computer science papers from arXiv, spanning 10 years of collections (2010 to 2020), resulting in over 2 million figures extracted from 295,028 papers. The dataset focuses exclusively on graph plots. Its development involved subfigure identification, text normalization, and caption type classification. The final dataset consists of three caption types: first-sentence captions, single-sentence captions, and captions limited to a maximum of 100 words.

In baseline experiments, the authors developed a CNN-LSTM architecture with three variations: vision-only, text-only (utilizing text extracted from figures), and multimodality (vision and textual features). Across all variations, the baseline models achieved relatively low BLEU-4 scores (approximately 0.02 to 0.03), highlighting the challenges of replicating human-written captions for scientific figures. The experimental results also revealed comparable performance between vision-only and text-only features, with no significant improvement observed when combining the two.

Building upon these foundations, particularly SciCap [27], we reframe scientific figure captioning as a knowledge-augmented task through SciCap+ [69]. SciCap+ incorporates mention-paragraphs and OCR tokens to address several limitations in previous works. Previous datasets like FigureQA [29] and FigCAP [13] rely on synthetic figures, while SciCap+ utilizes real-world scientific figures to represent actual scientific communication better. SciCap defines figure captioning as a figure-to-caption task, but SciCap+ recognizes that generated figure captions require contextual knowledge from both the surrounding text and the figure itself. Our experiments using a multimodal transformer architecture demonstrate that integrating mention paragraphs and OCR tokens significantly improves caption generation performance compared to vision-only baselines. The

human evaluation reveals that even expert annotators struggle to write captions without access to contextual information, validating our knowledge-augmented approach. These findings reinforce that scientific figure captioning fundamentally differs from natural image captioning and requires deeper integration of visual and textual knowledge.

3.3.1 Chart/Table-to-Text

Generating text from tables is closely related to generating captions for figures. A recently proposed benchmark, Chart-to-Text [30], addresses table-to-text generation. It offers two problem settings: one where the underlying data table of the chart is provided and another where data needs to be extracted from chart images. The experimental findings suggest that utilizing the underlying data table results in table captions with fewer factual errors. Another study [56] introduced a table-to-text generation dataset comprising pairs of tables and paragraphs describing the tables. The authors employed pre-trained language models with copy mechanisms to enhance numerical reasoning.

Chart-to-text [30]: a large-scale benchmark for automatically generating natural language descriptions of charts and graphs. The benchmark consists of two datasets totalling 44,096 charts collected from Statista and Pew Research, covering diverse chart types (bar, line, pie, etc.) and topics. The authors presented two variations of the chart-to-text task: one where the underlying data table is available, and another more challenging scenario where only the chart image is available.

The authors implemented several baseline approaches using state-of-the-art neural models, including image captioning models (using Show, Attend and Tell with a ResNet50 encoder), data-to-text models (BART and T5), and hybrid models that combine computer vision and text generation. For the image-only scenario, they use OCR to extract text from charts before generating descriptions. Their evaluation using both automatic metrics and human assessment shows that while models can generate fluent summaries, they struggle with complex patterns, trends, and factual accuracy. The best performing models were T5 and BART variants, particularly when given access to the underlying data tables.

The work identifies several key challenges in chart summarization: handling perceptual and reasoning aspects of charts, avoiding hallucinations in generated

text, maintaining factual accuracy (especially for OCR-based approaches), and computer vision challenges like associating data values with chart elements.

Suadaa et al. [56] introduced numericNLG, a dataset and framework focused on table-to-text generation with numerical reasoning capabilities. Unlike previous table-to-text datasets that mainly focused on descriptive text generation, numericNLG was specifically designed to support the generation of analytical text that requires mathematical reasoning from tabular data. The dataset consists of pairs of numerical tables and their corresponding descriptions collected from scientific papers, where the descriptions naturally contain richer numerical inferences written by domain experts.

The authors proposed a novel framework combining pre-trained language models with a copy mechanism. Their approach used template-guided text generation with pre-executed numerical operations to guide the generation of text containing numerical reasoning. To improve faithfulness to the source data, they incorporated a copy mechanism using general placeholders during fine-tuning to avoid hallucinated content while maintaining fluency. The framework considered different types of table representations, including data-based templates and reasoning-based templates that captured explicit facts and numerical operations.

Their experimental results showed that while pre-trained models like T5 could generate fluent text, they still struggled with maintaining fidelity to the source table contents. Adding their proposed copy mechanism helped improve performance, particularly for encoder-decoder architectures like T5, though decoder-only models like GPT-2 had difficulty incorporating the copy mechanism effectively. Through automatic and human evaluation, they demonstrated that their approach could generate more accurate descriptions while preserving readability, though challenges remained in balancing fluency with faithful numerical reasoning.

Compared to these studies, our work on SciCap+ [69] addresses unique challenges in scientific figure captioning. While Chart-to-Text [30] and numericNLG [56] focus on describing data in tables, SciCap+ tackles the broader challenge of explaining scientific figures. While tables are with fixed structures, figures have diverse visual representations, making caption generation more challenging. The table-to-text research primarily concentrates on data-to-text generation, whereas SciCap+ frames figure captioning as a knowledge-augmented task that integrates information across multiple modalities: the figure itself, OCR-

extracted text, and contextual paragraphs that mention the figure. We adopt the M4C-Captioner architecture incorporating a pointer network, enabling dynamic text selection from OCR tokens or a predefined dictionary during caption generation. Through comprehensive experiments, we demonstrate that this multimodal approach with pointer-based text selection significantly improves caption generation quality compared to methods that rely on visual features alone. Our human evaluation reveals that even with access to mention-paragraphs, writing informative figure captions remains challenging for humans, highlighting the complexity of scientific figure captioning as a distinct task from chart description or table-to-text generation. This indicates that future work should focus on better integrating domain knowledge and visual understanding for scientific figure comprehension.

4 News Image Caption Generation

4.1 Introduction

Image captioning is a task that automatically generate natural language descriptions for images, has emerged as a significant research area at the intersection of Computer Vision and Natural Language Processing. This task has drawn substantial attention from both research communities due to its dual importance: 1. as a practical tool for applications such as automatic image indexing and accessibility enhancement, and 2. as a fundamental challenge in advancing image understanding capabilities, including object recognition, relationship detection between objects, and scene comprehension [62, 67, 31].

This study investigates a context-driven variant of image captioning called *news image captioning*. This task extends beyond conventional image captioning by incorporating the image and accompanying article text as input to generate appropriate descriptions, unlike traditional image captioning systems, which operate solely on visual input without referring to textual context. News image captioning requires the model to integrate information from multimodal sources while maintaining contextual relevance to the news article.

Research in news image captioning has evolved through two main approaches. Early work by Feng and Lapata [22] and Tariq and Foroosh [59] introduced a two-stage methodology: first annotating images and text with relevant keywords, then generating descriptions based on these annotations. More recent approaches have shifted toward end-to-end architectures that directly integrate image and text features using deep neural networks [51, 7, 9]. However, the previous studies did not focus on the usefulness of text in the news image captioning task, extending the conventional models for image captioning to incorporate text features.

Figure 4.1 shows an example of a news image caption. It may be challenging

The Silent-Era Shostakovich

When prominent conductors visit New York for important engagements with local institutions, their schedules are usually very full. So it says good things about the priorities of the Russian maestro Vladimir Jurowski that while in New York to conduct a six-performance run of Strauss's "Die Frau Ohne Schatten" at the Metropolitan Opera, he made time to work with the Juilliard Orchestra on an adventurous Shostakovich program. That concert took place on Monday at Alice Tully Hall, the night before Mr. Jurowski's final "Frau" at the Met. … …

Juilliard Orchestra Vladimir Jurowski conducting at Alice Tully Hall on Monday evening.

Figure 4.1: An example demonstrating the importance of news article context in image captioning. While the image shows only a conductor and orchestra in performance, the article text reveals crucial details: this is Vladimir Jurowski leading the Juilliard Orchestra at Alice Tully Hall during his Metropolitan Opera engagement.

to recognize the central object in the image, for example, people, violins, and a stick (a bow). Also, the caption includes much information (e.g., *Juilliard Orchestra*, *Vladimir Jurowsky*, and *Alice Tully Hall*) that may be hard to tell only from the image. This kind of example is relatively common in news articles, where text is the primary medium of information, and an image and its caption provide additional explanations that support the text. However, no previous work explored a method that seamlessly integrated an article text with an image in news image captioning.

This example demonstrates a common pattern in news media: textual context carries major information, and the synergy between text and visual contexts in news articles creates a comprehensive information package. Images and captions are complementary elements that enhance and reinforce the news narrative in news articles. Despite the apparent importance of multimodal integration, existing research has not fully addressed the challenge of seamlessly integrating article text with image content in news image captioning models.

In this study, we propose a method for news image captioning based on the Transformer architecture [60], which has demonstrated success across various natural language processing tasks, including machine translation, abstract summarization, and contextualized word embeddings. Our proposed model extends the Transformer framework to integrate textual and visual modalities, enabling

attention mechanisms to draw upon textual features informed by visual elements during caption generation.

Our experimental findings yield two insights: first, that article text is the primary source of information for generating journalist-style image captions, and second, that our proposed model outperforms the current state-of-the-art approach [9]. We complement these quantitative results with human evaluation studies, comprehensively analyzing the challenges inherent in news image captioning.

4.2 Problem Definition

News image captioning presents a unique challenge in multimodal machine learning, extending beyond traditional image captioning tasks. Given an input pair consisting of an image i and its associated news article text represented as a sequence of n tokens $(x_1, x_2, \ldots x_n)$, the task aims to generate a contextually appropriate caption as a sequence of m tokens $(y_1, y_2, \ldots y_m)$. This can be formally expressed as learning the conditional probability distribution:

$$P(y_1, y_2, \ldots y_m | i, x_1, x_2, \ldots x_n) \tag{4.1}$$

4.3 Method

In this section, we detail the design of the proposed model for news image captioning. The proposed news image captioning model is based on Transformer [60], a proven successful model in many natural language processing tasks. Since the input modalities are across text and vision, we utilize the attention mechanism in the Transformer model and adapt it from monomodal attention to multimodal attention.

Drawing from the architecture illustrated in Figure 4.2, the proposed news image captioning system employs a three-stage architecture to effectively combine visual and textual information from news articles to generate contextually relevant captions. In the first stage, the image encoder processes the input news image, extracting relevant visual features using deep convolutional neural networks. These visual features are fed into the image-article encoder and the corresponding

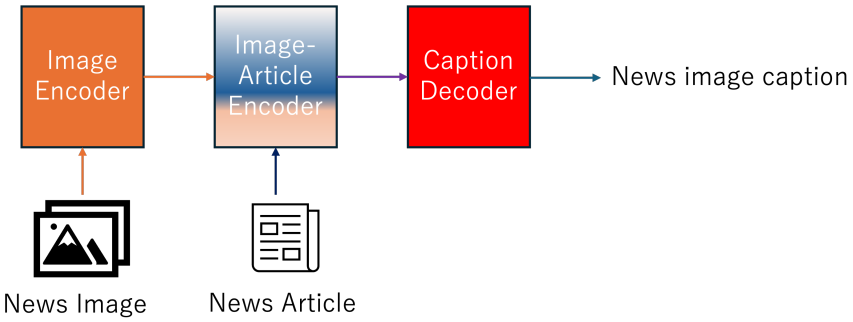


Figure 4.2: Overall architecture of the news image captioning system. The model consists of three main components: an image encoder that processes the input news image, an image-article encoder that jointly processes visual and textual features from both the image and news article, and a caption decoder that generates the final news image caption. The architecture demonstrates how visual and textual information are combined to generate contextually relevant image captions for news articles.

news article text. This second component is a crucial bridge, performing joint encoding of visual and textual modalities to create rich, multimodal representations that capture the relationships between the image content and the article context. Finally, the caption decoder takes these integrated representations. It generates a news image caption that not only describes the visual content of the image but also incorporates relevant contextual information from the news article.

The end-to-end design allows for seamless information flow between components. At the same time, the dedicated encoders for both image and text ensure that essential features from each modality are adequately captured and utilized in the final caption generation. This unified architecture enables the model to produce captions incorporating important context and maintain news value beyond simple visual descriptions, which is essential for news image captioning tasks.

4.3.1 Multimdoal Transformer Model

Unlike conventional image captioning, which focuses solely on describing visual content, news image captioning must integrate information from visual and textual modalities. We extend the popular Transformer model [60] to multimodal

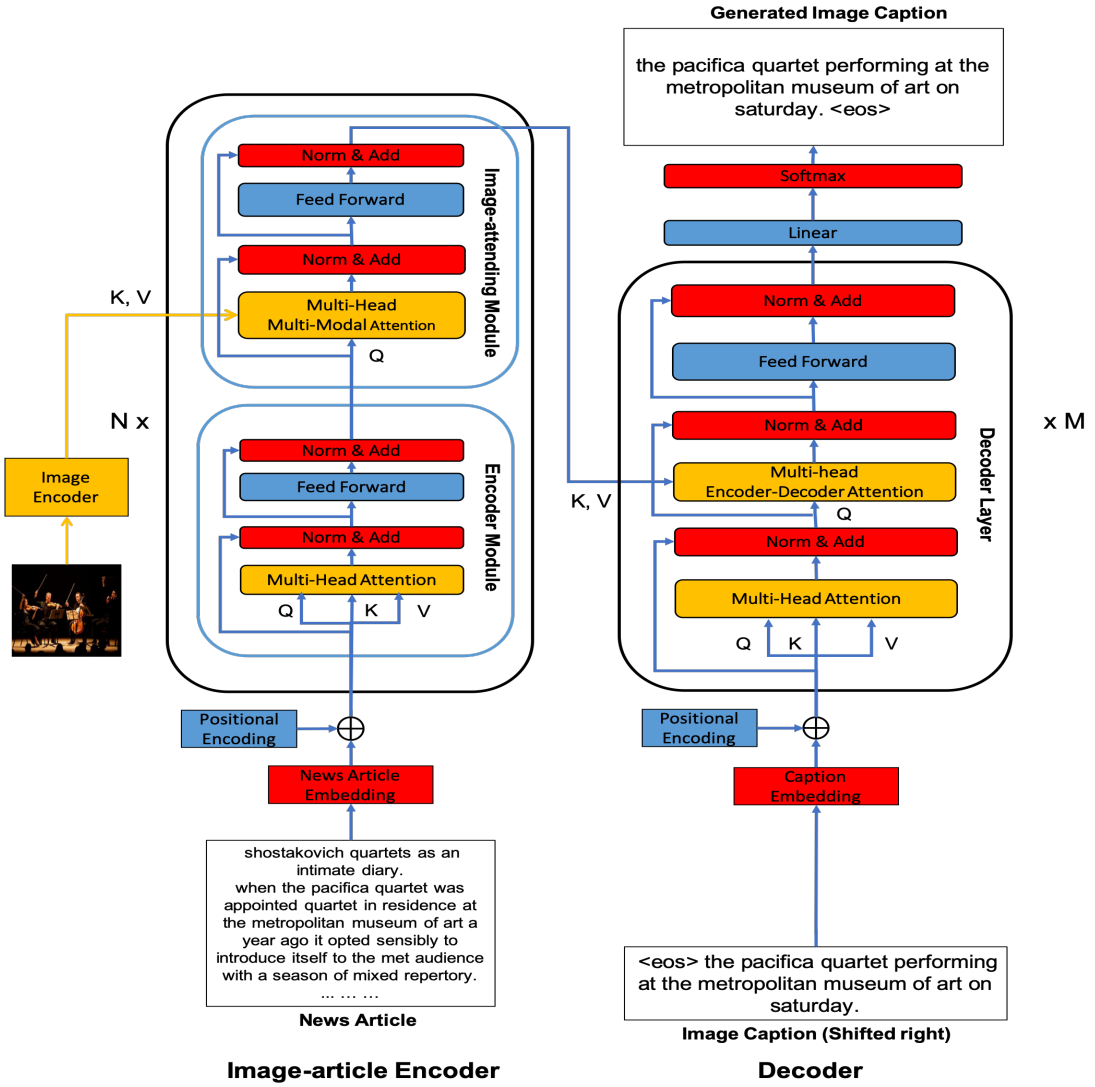


Figure 4.3: Overall architecture of the proposed multimodal Transformer model.

fashion. Our proposed multimodal Transformer architecture illustrated in 4.3, consists of three main components: **an image encoder**, **an image-article encoder**, and **a decoder**. Each component plays a distinct role in integrating visual and textual information to generate news image captions.

Image Encoder

The image encoder, shown in the left branch of Figure 4.3, transforms an input image i into a feature vector $p_i \in \mathbb{R}^d$ through:

$$p_i = \text{CNN}(i) \tag{4.2}$$

where d represents the dimensionality of hidden feature vectors, and CNN(.) denotes a convolutional neural network. In this study, we experiment with two variants of CNN architectures: one trained for object recognition and another for scene recognition, allowing us to capture different aspects of visual information.

Image-Article Encoder

The image-article encoder, depicted in Figure 4.4, processes visual and textual inputs to compute their joint representations. Its first layer receives a $d \times n$ matrix formed by combining token embeddings of texts with positional encodings. This encoder consists of N identical blocks, each containing two key sub-modules:

1. The encoder module (lower box in each block) implements the standard Transformer encoder architecture, mapping $\mathbb{R}^{d \times n}$ to $\mathbb{R}^{d \times n}$ through multi-head self-attention followed by a feed-forward layer. This allows the model to capture relationships between different parts of the input text.

2. The image-attending module (upper box in each block) extends the standard Transformer decoder architecture, also mapping $\mathbb{R}^{d \times n}$ to $\mathbb{R}^{d \times n}$. It employs multi-head target-source attention using the image vector p_i as both keys and values. Through residual connections [24], it combines the outputs from both the encoder module and the target-source attention, enabling the model to incorporate visual context when processing textual features.

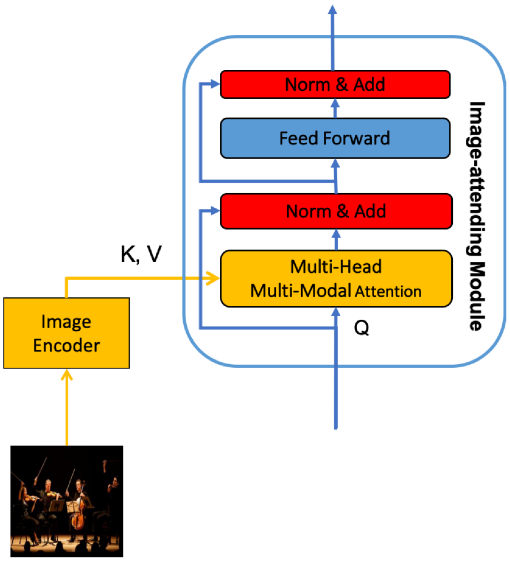


Figure 4.4: Overall architecture of the proposed image-article encoder.

The target-source attention in the image-attending module is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.3}$$

where $K = p_i$, $V = p_i$.

This module leverages the attention mechanism to facilitate joint text-vision attention. The target-source attention effectively captures implicit contextual relationships between visual and textual features, enabling the modification of input representations based on the provided image.

Decoder

The decoder, shown in the right portion of Figure 4.3, comprises M layers that follow the original Transformer decoder architecture. Each layer contains:

1. Multi-head masked self-attention over previously generated tokens
2. Target-source attention receiving encoder outputs as keys and values

Each decoder layer first applies multi-head masked self-attention to the previously generated tokens, followed by target-source attention that uses the output from the encoder as keys and values. This structure allows the decoder to create captions by attending to textual and visual information processed by the image-article encoder.

4.3.2 Two-staged Template Model

Previous state-of-the-art work Biten et al. [9] presented a two-stage approach for news image captioning based on template completion. Their model architecture combines two encoders with an LSTM component: a CNN encoder that extracts features from images, and a text encoder that processes article text at the sentence level. We adopted this approach as a baseline in our study.

In the first stage, a long short-term memory (LSTM) network combines textual and visual features to produce template captions with placeholders for named entities. In the second stage, these placeholders are replaced with named entities extracted from the article text, completing the caption.

Image Encoding

The image encoder computes attended image features at each time-step t as:

$$\begin{aligned} I_f &= \text{CNN}(I) \\ I_t &= \text{Att}(h_{t-1}, I_f) \end{aligned} \tag{4.4}$$

Here, I_f represents visual features extracted from input image I using a CNN. The attention model Att is a multilayer perceptron conditioned on the previous LSTM hidden state h_{t-1} [67]. The model extracts object-level visual features using ResNet152 [24] pretrained on ImageNet [18].

Article Encoding

Encoding articles at the sentence-level instead of word-level offers two advantages: reduced dimensionality and preserved sentence-level context. For an article $A_i = sen_0, ..., sen_m$ with sentences using GloVe [48] word vectors $sen_j = w_0, ..w_{nj}$, the model implements four encoding methods:

1. **Average (AVG)** Averaging the GloVe vectors of each word creates a single vector representation for each sentence:

$$A_{f_j}^{avg} = \frac{1}{n_j} \sum_{i=0}^{n_j} w_i, \text{where } j = 0, 1, \ldots, m \tag{4.5}$$

2. **Weighted average (WAVG)** Computing a weighted average of word vectors produces a single vector for each sentence:

$$\begin{aligned} A_{f_j}^{wAvg} &= \frac{1}{n_j} \sum_{i=0}^{n_j} p\left(w_i\right) w_i \\ p(w) &= \frac{a}{a + tf(w)} \end{aligned} \tag{4.6}$$

   where $p(w)$ denotes the smoothed inverse frequency and $tf(\cdot)$ indicates the term frequency.

3. **Tough-to-beat baseline (TBB)** This method follows the tough-to-beat baseline (TBB) [3], which subtracts the first PCA component from the weighted average encoding matrix:

$$A_{f_j}^{TBB} = A_f^{wAvg} - X \tag{4.7}$$

   where X denotes the 1^{st} component of A_f^{wAvg}

4. **Article encoding with attention** The model computes the final article representation by multiplying the article encoding matrix $A_f \in R^{M \times D_w}$ with an attention vector $\beta_t \in R^M$:

$$\begin{aligned} A_f &= GloVe\left(A_i\right) \\ A_t &= \beta_t * A_f \end{aligned} \tag{4.8}$$

   A fully connected layer learns the attention vector β_t using the LSTM state h_{t-1} and article matrix A_f:

$$\begin{aligned} \theta_t &= \text{FFN}\left(h_{t-1}, A_f\right) \\ \beta_t &= \text{softmax}\left(\theta_t\right) \end{aligned} \tag{4.9}$$

Template Caption Generation

The first stage generates template captions by combining textual and visual features. The template caption generation process for sequence $s_i = w_0, \ldots, w_N$ follows:

$$\begin{aligned} x_t &= W_e w_t, \text{ where } t \in 0, 1, \ldots, N-1, \\ o_t &= \text{LSTM}(\text{concat}(x_t, I_t, A_t)) \\ w_{t+1} &= \text{softmax}(W_{ie} o_t) \end{aligned} \tag{4.10}$$

The model predicts the next word based on previous words and attended visual features at each time step.

Named Entity Insertion The second stage inserts named entities into the template captions using two approaches:

1. **Context insertion (CtxIns)** inserts named entities based on cosine similarity between article sentences and template captions using GloVe embeddings.

2. **Attention insertion (AttIns)** leverages the article attention vector β_t from template generation to guide entity insertion.

4.4 Experiments

The experiments in this section address two fundamental research questions:

- How effectively can the proposed multimodal Transformer model integrate visual and textual features for caption generation?

- What is the relative contribution and importance of textual versus visual features in generating news image captions?

We conducted experiments to investigate these research questions using an enhanced version of the GoodNews dataset, a large-scale news image captioning dataset. Section 4.4.1 details our improvements to address quality issues in the original dataset. After explaining the experimental settings (in Setions 4.4.2 to 4.4.5), we present evaluation results using both automatic metrics (Section 4.4.6) and human evaluation (Section 4.4.7) to provide a thorough assessment of model performance.

4.4.1 Dataset for News image Captioning

Biten et al. [9] introduced the GoodNews dataset, the largest dataset available for news image captioning in 2020. However, our analysis revealed several critical issues that needed addressing. First, many instances contain incomplete article text, particularly missing leading sentences that typically provide essential context and information. This omission is particularly problematic for our research, which aims to study the interplay between visual and textual information in caption generation. Additionally, the original dataset by Biten et al. [9] used image-level splits between training and test sets, potentially allowing the same news article text to appear in both sets, which could lead to data leakage and unreliable evaluation results.

To address these issues, we reconstructed the dataset by crawling complete text for each news article and implementing article-level random splits. The resulting dataset comprises 269K articles containing 489K images in total. Statistical analysis of the dataset reveals that articles contain 1.8 images on average, with 59% of articles containing a single image; regarding text length, articles average 963.29 words for body text, 8.57 words for headlines, and 17.55 words for image captions. We divided the dataset into three splits: 245K articles for training, 10K for validation, and 13K for testing.

4.4.2 Data Preprocessing

Due to GPU memory constraints and the observation that most essential information appears in headlines and leading paragraphs, we truncated each article to a maximum of 416 words (including the headline). For text preprocessing, we removed all punctuation marks except for periods ('.'), which were retained as sentence delimiters, and filtered out non-ASCII characters. We then constructed a vocabulary of 32,000 subwords by applying Byte-Pair-Encoding (BPE) to the preprocessed training articles using SentencePiece[1] [34].

4.4.3 Baselines and Model Variants

Biten et al. [9] established the state-of-the-art performance on the GoodNews dataset. We evaluated their approach using their publicly-available implementa-

[1]https://github.com/google/sentencepiece

tion[2] on our dataset across six configurations:

- Avg + AttIns
- Avg + CtxIns
- TBB + AttIns
- TBB + CtxIns
- Wavg + AttIns
- Wavg + CtxIns

These configurations are detailed in 4.3.2.

To investigate the relative importance of visual and textual features in news image captioning, we developed several variants of our proposed model:

- **Transformer (Text)**: The original Transformer model without image features is a baseline for evaluating caption generation using only textual information.
- **Transformer (ImageNet)**: A decoder-only Transformer model using image features from an object recognition model trained on ImageNet[3] for the multi-head target-source attention keys and values. This baseline generates captions using only visual information.
- **Transformer (Places 365)**: Similar to Transformer (ImageNet), but utilizing image features from a scene recognition model trained on Places 365[4] [72].
- **Multimodal Transformer (ImageNet)**: Our proposed model incorporates image features from the ImageNet-trained object recognition model.
- **Multimodal Transformer (Places 365)**: Our proposed model uses image features from the Places 365-trained scene recognition model.

[2]https://github.com/furkanbiten/GoodNews/
[3]http://www.image-net.org/
[4]http://places2.csail.mit.edu/

- **Multimodal Transformer (ImageNet & Places 365)**: Our proposed model combines object and scene recognition features through averaging.

For all models requiring visual features, we employed ResNet-18 [24] pre-trained on either ImageNet or Places365 to extract image features. We focused on these two pre-training datasets to evaluate the impact of both object-centric and scene-centric visual features. Additionally, we implemented two simple baselines: **Lead** and **Headline**, which use the article's lead sentences and headlines as image captions, allowing us to assess the similarity between image captions and these standard article components.

4.4.4 Implementation and Training

We implemented all Transformer models using PyTorch [47] based on Fairseq [44].

Hyper-parameters All Transformer-based models in our experiments used the same hyper-parameters:

- the number of dimensions of hidden vectors $d = 512$;
- the number of attention heads $H = 8$;
- the number of encoder blocks $N = 3$;
- the number of decoder blocks $M = 6$.

Table 4.1 presents the number of trainable parameters for each model variant.

Model	**# Parameters**
Transformer (Text)	93M
Transformer (Image)	57M
Multimodal Transformer	93M

Table 4.1: Number of parameters trained in the Transformer-based models.

Training For parameter optimization, we employed Adam [33] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$. The learning rate schedule consisted of:

- Linear warmup from 10^{-7} to 5×10^{-4} over 4,000 steps
- Decay proportional to the inverse square root of step number
- Minimum learning rate of 10^{-9}

We linearly increased the learning rate from the initial rate 10^{-7} until 0.0005 in the 4,000 warm-up steps and then decreased it proportionally to the inverse of the square root of the step number with the minimum learning rate of 10^{-9}. The objective function is the cross-entropy loss with label smoothing of 0.1 [58].

In each layer of the model, we applied dropout [55] with the rate of 0.3 after the layer normalization and before the residual connection. In both the encoder and decoder, we applied dropout with the rate of 0.3 after taking the sum of token embeddings and positional encoding. We also applied dropout with the 0.1 rate to the attention weights.

We trained all models for 50 epochs, saving the parameters that achieved the minimum loss. Training a single Multimodal Transformer model required approximately (1.2) days using four NVIDIA Tesla V100 GPUs with NVLink ((16) GiB HBM2).

4.4.5 Evaluation Metrics

We evaluated our model using five standard automatic metrics: BLEU [46], METEOR [19], ROUGE [37], CIDEr [61], and SPICE [2]. Among these metrics, we identified CIDEr as particularly informative for our task, given the prevalence of named entities in news image captions. For all metrics, we utilized the MS-COCO caption evaluation tool[5], preprocessing the captions by converting to lowercase and removing punctuation.

To address the dense presence of named entities in news articles and their image captions, where these entities often carry crucial contextual information, we introduced a specialized metric called $\text{Coverage}_{\text{NE}}$. This metric measures how effectively generated captions preserve named entities from the ground truth, defined as:

[5]https://github.com/tylin/coco-caption

$$\text{Coverage}_{\text{NE}} = \frac{|E_{generated} \cap E_{gold}|}{|E_{gold}|}. \tag{4.11}$$

Here, $E_{generated}$ and E_{gold} represent the sets of named entities in the generated and ground-truth captions, respectively. To compute this metric, we employed SpaCy[6] for named entity recognition in ground truth captions and used regular expressions to identify exact matches in generated captions.

4.4.6 Results (Automatic Evaluation)

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Multimodal Transformer (ImageNet)	18.62	10.81	6.70	4.46	8.38
Multimodal Transformer (Places 365)	**18.78**	**10.90**	**6.76**	**4.52**	8.42
Multimodal Transformer (ImageNet & Places 365)	18.10	10.53	6.56	4.38	8.27
Transformer (Text)	15.10	8.72	5.37	3.55	7.27
Transformer (Image) (ImageNet)	12.13	4.66	2.20	1.23	3.49
Transformer (Image) (Places 365)	11.72	4.43	2.06	1.16	3.40
Lead	14.54	7.57	4.47	2.92	**8.62**
Headline	6.86	3.16	1.52	0.81	4.69
Biten et al. [9] (Avg + AttIns)	6.73	2.53	1.17	0.63	3.56
Biten et al. [9] (Avg + CtxIns)	6.95	2.52	1.14	0.62	3.58
Biten et al. [9] (TBB + AttIns)	5.22	1.74	0.72	0.36	2.95
Biten et al. [9] (TBB + CtxIns)	5.85	2.08	0.92	0.48	3.43
Biten et al. [9] (Wavg + AttIns)	6.31	2.34	1.05	0.56	3.61
Biten et al. [9] (Wavg + CtxIns)	6.52	2.33	1.02	0.52	3.67

Table 4.2: Performance of news image caption generation measured by BLEU and METEOR.

Table 4.2 and 4.3 present the automatic evaluation results comparing our baseline models, Transformer-based variants, and the state-of-the-art model [9] on our dataset. The Multimodal Transformer (ImageNet) achieved the highest CIDEr score among all models. All Transformer-based models incorporating textual features demonstrated substantial performance improvements over the state-of-the-art model [9].

Interestingly, the simple Headline baseline performed comparably to the state-of-the-art model [9], while the Lead method emerged as a particularly strong

[6]https://spacy.io/

Model	ROUGE-L	CIDEr	SPICE
Multimodal Transformer (ImageNet)	**20.56**	**44.16**	**10.19**
Multimodal Transformer (Places 365)	20.41	44.04	10.15
Multimodal Transformer (ImageNet & Places 365)	20.29	43.01	10.07
Transformer (Text)	17.83	37.02	8.96
Transformer (Image) (ImageNet)	11.70	9.37	2.44
Transformer (Image) (Places 365)	11.45	8.70	2.29
Lead	12.91	12.79	6.96
Headline	10.31	16.80	5.46
Biten et al. [9] (Avg + AttIns)	11.81	12.38	3.27
Biten et al. [9] (Avg + CtxIns)	11.72	11.40	2.94
Biten et al. [9] (TBB + AttIns)	10.80	8.43	2.62
Biten et al. [9] (TBB + CtxIns)	11.52	10.85	2.92
Biten et al. [9] (Wavg + AttIns)	11.72	11.92	3.18
Biten et al. [9] (Wavg + CtxIns)	11.72	11.36	2.92

Table 4.3: Performance of news image caption generation measured by ROUGE-L, CIDEr, and SPICE.

baseline for news image captioning. We attribute this strong baseline performance to the dual role of images and captions in news articles: they often act as indicative summaries that capture reader attention and provide entry points to the full article. This suggests that journalists strategically craft image captions as article summaries, positioning the image-caption pair as an alternative beginning point for readers engaging with news articles.

Another noteworthy finding is that the Transformer (Text) model outperforms Transformer (Image) models even without access to the actual images and proved to be a strong baseline for generating news image captions. This finding suggests that news image captions describe visual content and incorporate substantial information from the article text. The significant gap in CIDEr scores between Transformer (Text) and Transformer (Image) particularly highlights the limitations of vision-only models in handling named entities.

The Multimodal Transformer models demonstrated further performance improvements over the Transformer (Text), confirming the value of incorporating visual features in caption generation. Among these models, Multimodal Transformer (Places 365) achieved the highest BLEU scores, while Multimodal Trans-

former (ImageNet) led in ROUGE-L, CIDEr, and SPICE metrics. Combining both visual features in Multimodal Transformer (ImageNet & Places 365) resulted in a slight performance decrease. However, it's worth noting that the performance variations among different Multimodal Transformer models were relatively minor. A detailed case study comparing captions generated by Multimodal Transformer models and the Transformer (Text) model is presented in Section 4.5.

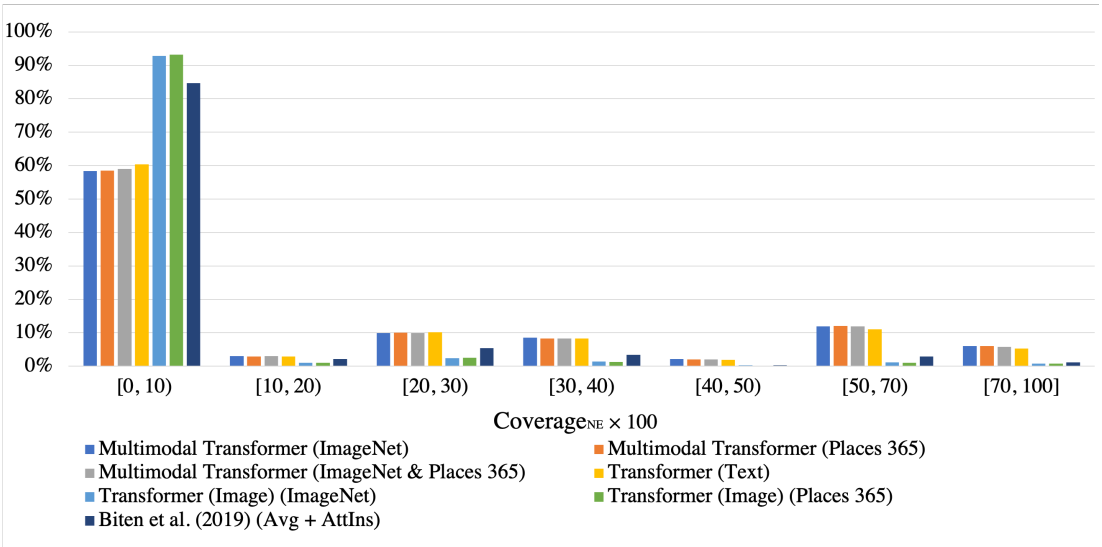


Figure 4.5: Distributions of $\text{Coverage}_{\text{NE}}$ scores for seven representative models.

Figure 4.5 illustrates the distribution of named entity coverage scores ($\text{Coverage}_{\text{NE}}$) across different models. The graph maps the ($100 \times \text{Coverage}_{\text{NE}}$) scores on the x-axis against the proportion of test instances falling within each score range on the y-axis.

The named entity coverage score distribution reveals a clear pattern: Transformer models with access to article text demonstrated substantially higher coverage of correct named entities than other models. Since Transformer (Image) models cannot access article text, they naturally struggled to incorporate correct entities. The previous state-of-the-art model [9] achieved intermediate performance, falling between the Transformer (Image) models and the Multimodal Transformer models regarding named entity coverage.

Model	Grammaticality	Faithfulness	Descriptiveness	Overlap
Multimodal Transformer (Places 365)	4.51	2.29	**1.77**	**1.60**
Transformer (Text)	**4.58**	**2.37**	1.69	1.47
[9] (Avg + AttIns)	2.62	2.08	1.41	1.23

Table 4.4: Average scores of human evaluation for three representative models.

4.4.7 Results (Human Evaluation)

Because we were unsure of the appropriateness of the automatic evaluation in this task, we also conducted a comprehensive human evaluation. We asked three native English speakers to evaluate generated captions from three models: Multimodal Transformer (Places 365), Transformer (Text), and Biten et al. [9] (Att + AttIns).

We randomly chose an evaluation set with 136 images. Each instance in this evaluation set included the image, news article, ground truth caption, and generated captions from the three models. To prevent order bias, three generated captions were presented to human subjects in random order so they could not guess the quality of a caption from the appearance order.

We designed four criteria for rating generated captions: grammaticality, faithfulness, descriptiveness, and overlap.

- **Grammaticality**: The caption has no error (5), has one error (4), has two errors (3), is understandable (2), or is incomprehensible (1).
- **Faithfulness**: The caption has: no unfaithful fact (5), one unfaithful fact (4), a few (but less than 50%) unfaithful facts (3), more than 50% unfaithful facts (2), or the content that is totally unrelated to the article and image (1).
- **Descriptiveness**: The caption explains the image (3); the caption does not explain the image but describes something related to the image (2), or the caption is totally unrelated to the image (1).
- **Overlap**: The overlap between generated and ground-truth captions, 100% overlap (5), 80% overlap (4), 50% overlap (3), 20% overlap (2), or no overlap (1).

Note that human evaluation is not easy. Human evaluators are not guaranteed to be familiar with objects and scenes (e.g., people, buildings, locations) appearing in the news. Although a news article (text) provides a hint for interpreting an image, they may find it hard to search for evidence of the image on the Internet. Therefore, we always presented ground-truth captions to human subjects to help them understand images.

Table 4.4 shows the average score assigned to each model and criterion. The two Transformer models significantly outperformed the previous state-of-the-art model across all four evaluation criteria. Notably, Biten et al. [9] (Avg + AttIns) received particularly low scores in grammaticality, suggesting issues with the linguistic quality of its generated captions.

However, neither emerged as a clear winner between the two Transformer models. The ranking pattern observed in the overlap criterion aligns with the results from automatic evaluation metrics discussed in Section 4.4.6, which is expected given that the overlap criterion essentially serves as a manual counterpart to these automatic metrics. While Multimodal Transformer (Places 365) showed slightly better performance in the descriptiveness criterion, its margin over Transformer (Text) was minimal. The generally low scores across all models highlight the inherent challenges in news image captioning.

4.5 Case Study

Figure 4.6 showcases examples of captions generated by different Transformer models. In example (a), the three Multimodal Transformer variants successfully captured both the subject and the visual context by describing the ”news conference” setting in their generated captions. In contrast, while accurately identifying the person, the Transformer (Text) model was limited to generating only the subject's name without capturing the visual scene. Example (b) highlights a common challenge in news image captioning: the disconnect between literal visual content and journalistic intent. The ground truth caption (”a little campari before dancing”) demonstrates how journalists often write captions that support news narratives rather than merely describing visual elements in images. While the Multimodal Transformer models generated factually accurate descriptions of the visual scene (focusing on the apartment setting), they failed to capture the broader narrative context the journalist intended to convey. This illustrates the

complexity of news image captioning, where success requires visual accuracy and understanding and conveying the broader journalistic message.

4.6 Conclusion

In this study, we presented a method for news image captioning based on the Transformer model that integrates text and image modalities and attends to textual features from visual features when generating captions. We demonstrated that our proposed model successfully integrates visual and textual information in caption generation through extensive experiments using both automatic metrics and human evaluation. Our findings reveal that news image captioning functions primarily as a context-driven task, where text from news articles fundamentally contributes to generating context-coherent captions. While textual features provide major information, incorporating visual features enhances caption quality.

Our work contributes to multimodal learning research and provides insights into the unique challenges of generating captions for news images. The findings from this study have broader implications for understanding how to effectively combine textual and visual information in the context-driven image captioning task.

We are interested in several research directions for future work to advance context-driven image captioning. First, developing specialized visual encoders trained specifically on news images would enhance the ability to recognize and interpret journalistic visual content. Current approaches rely on visual encoders pretrained on general image datasets, which may not effectively capture the unique characteristics of news photography.

Second, we aim to explore more sophisticated attention mechanisms beyond the basic image-article encoder in our news image captioning model: attention architectures could better handle the complex relationships between news text and images, improving caption quality and relevance.

(a)

Ground Truth
yoshihiko noda japan s prime minister speaking at a news conference in tokyo on monday.

Multimodal Transformer (ImageNet & Places 365)
prime minister yoshihiko noda of japan at a news conference in tokyo on monday.

Multimodal Transformer (ImageNet)
prime minister yoshihiko noda at a news conference in tokyo on monday.

Multimodal Transformer (Places 365)
prime minister yoshihiko noda of japan at a news conference in tokyo on monday.

Transformer (Text)
prime minister yoshihiko noda of japan left and prime minister yoshihiko noda of japan in tokyo on monday.

(b)

Ground Truth
a little campari before dancing.

Multimodal Transformer (ImageNet & Places 365)
ms weinra weinra at her apartment in the west village.

Multimodal Transformer (ImageNet)
ms weinrauch in her manhattan apartment.

Multimodal Transformer (Places 365)
ms weinuch in her apartment in manhattan.

Transformer (Text)
helena weinrauch with her daughter arlene weinberg at the museum of jewish heritage in manhattan.

Figure 4.6: Captions generated by the Transformer models. In (a), the Transformer (Text) made the correct prediction for the person in the image (the prime minister of Japan). The Multimodal Transformer models injected the correct visual information (news conference) into the caption. In (b), all four models failed to generate the correct caption. The transformer (Text) predicted the correct name but the wrong contextual information. The Multimodal Transformer models generated captions with a different focus.

5 Scientific Figure Caption Generation

5.1 Introduction

Researchers share their work through various types of scholarly documentation, from peer-reviewed journal articles to conference proceedings and book chapters. These publications form the backbone of knowledge sharing in academia. These scientific documents primarily consist of text enhanced by figures and tables to accelerate knowledge communication. Figures, in particular, offer powerful visual representations that can distill complex scientific findings into clear, interpretable formats.

In scientific documents, figures are supported by two text components: 1. Captions interpret the visual content of figures 2. Supporting paragraphs provide deeper research context. A fundamental principle of scientific figure caption writing is that figures and their captions should stand alone, conveying their core message without requiring readers to consult the main text. This independence enables efficient knowledge transfer, supporting quick comprehension during initial review and deeper understanding during careful study.

Writing effective figure captions is challenging for authors, as it needs to balance brevity with informativeness, provide necessary context, and articulate key findings from figures. Nonetheless, the quality of figure captions directly impacts the overall effectiveness of scientific communication and knowledge sharing.

This study addresses this challenge by developing automated approaches for generating captions for scientific figures. We aim to assist authors in creating more informative figure captions, ultimately enhancing the clarity and accessibility of scientific documents and accelerating knowledge sharing within the scientific community.

While scientific figure captioning shares the fundamental goal of image cap-

tioning, generating descriptive text for figures, it presents distinct challenges that set it apart from traditional image captioning tasks:

1. The visual language of scientific figures stands apart from that of natural images. Instead of depicting familiar objects and scenes from the physical world, scientific figures communicate through abstract visual elements: from data plots and graphs to mathematical notation and textual annotations. This fundamental difference in structure and visual representation requires moving beyond conventional computer vision approaches to develop methods tailored explicitly for interpreting scientific figures.

2. The role of scientific figure captions transcends simple visual descriptions. While traditional image captions focus on what can be seen, scientific captions must explain not just what the data shows but what it means in the research context (analytical insights). These scientific figure captions serve as interpreters, helping readers connect visual elements to scientific conclusions by integrating figure content with broader research context.

Previous research in scientific figure captioning, notably the SciCap [27], approached this as a direct figure-to-caption task, where models generate captions using only the figure as input. However, their relatively low performance on automatic evaluation metrics suggested significant room for improvement. This limitation aligns with intuitive understanding: even human experts struggle to interpret scientific figures and write appropriate captions without sufficient context and background knowledge.

These observations and previous research lead to a key insight: generating informative scientific captions requires contextual information beyond figures. This context primarily exists in two forms: background knowledge from the body text that discusses figures and the textual elements embedded within the figure (OCR text). Both sources of information are crucial for reducing the complexity of the captioning task and improving output quality. For example, as demonstrated in Figure 5.1, the abbreviation "comm.(KB)" in isolation provides insufficient information for caption writing. Nonetheless, the mention-paragraph provides a crucial context in the caption: "communication cost." This example illustrates how context information is essential for generating accurate and informative captions.

Building on this insight, we present SciCap+, an enhanced version of the SciCap dataset that augments the original figures with two critical elements: mention paragraphs (text segments referencing the figures in the body text) and OCR-extracted text from figures. We reframe scientific figure captioning as a knowledge-augmented image-captioning task, where the caption generator integrates information from visual and textual sources. Using the M4C captioner model [52] as our baseline, we demonstrate how leveraging multimodal knowledge significantly improves captioning performance. Our experimental results show that incorporating information from different modalities, particularly from mention-paragraphs and OCR tokens, leads to substantial improvements in caption quality as measured by automatic evaluation metrics.

Beyond automatic evaluation metrics, we conducted human generation and evaluation tasks to investigate the inherent challenges of scientific figure captioning. This human evaluation yielded three significant findings:

1. Models leveraging multimodal knowledge outperform humans in caption generation tasks.

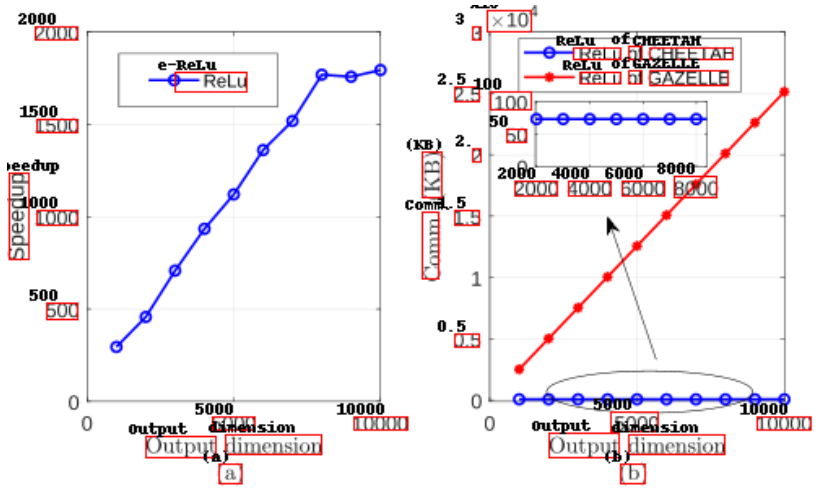
2. The informative quality of model-generated captions approaches that of ground-truth captions, with human evaluators showing no significant preference between the two.

3. Human caption writers face substantial challenges in producing captions that align closely with ground truth, even when provided with relevant paragraphs for reference.

To the best of our knowledge, this work presents the first approach to frame scientific figure captioning as a knowledge-augmented image-captioning task, demonstrating that integrating mention paragraphs and OCR tokens significantly improves the quality of generated scientific figure captions.

5.2 Problem Formulation

Scientific figure captioning was initially formulated by [27] as a standard image-captioning task, where given a figure I, a model generates a caption $C = [c_0, c_1, ..., c_N]$ as a sequence of subword tokens. As scientific figure captioning is a context-driven



Caption:
Fig. 7. (a) Speedup of CHEETAH over GAZELLE for computing ReLu. (b) Comparison of communication cost for ReLu.

Mention-paragraph:
Fig. 7 plots the speedup and communication cost as a function of the output dimension. Similarly, CHEETAH achieves an outstanding speedup with much smaller communication cost, independent of the output dimension, compared with GAZELLE.
……

Figure 5.1: Example figure [71] with its captions and mention-paragraph and the texts recognized via OCR. This example demonstrates a crucial point: without the contextual information provided by the mention-paragraph and OCR text to connect the figure with its textual references, it becomes difficult to properly interpret the presented data, specifically the communication cost comparison and speed-up metrics between the CHEETAH and GAZELLE systems.

captioning task, this formulation does not consider the necessary context needed for composing figure captions.

Therefore, we propose reframing scientific figure captioning as a knowledge-augmented image-captioning task that explicitly incorporates knowledge from textual and visual contexts. In our formulation, we identify two key modalities:

1. **Textual modalities**: These comprise the paragraph that mentions the figure (mention-paragraph) and any text embedded within the figure itself (extracted via OCR)

2. **Visual modalities**: Figures and visual appearance/layout of OCR texts are visual modalities.

Formally, given a scientific figure I and information extracted from text (K_{text}) and vision (K_{vision}) modalities, we define the figure caption generation task as finding the optimal caption C that maximizes the conditional probability $P(C|I, K_{text}, K_{vision})$. This problem formulation acknowledges that generating informative scientific figure captions requires integrating contextual knowledge from textual and visual sources.

5.3 SciCap+ Dataset

SciCap [27] is a large-scale figure-caption dataset comprising graphplots. These plots were extracted from ten years of collections from computer science sections of arXiv under computer science (cs) and machine learning (stat.ML) topics.

Building upon this foundation, we augmented approximately 414k figures from SciCap by incorporating two critical types of contextual information for each figure: its mention-paragraph and OCR tokens (both OCR texts and their corresponding bounding boxes). This augmentation process transforms SciCap into SciCap+, creating a richer dataset that captures the textual and visual context of figures. This section details the dataset creation and data augmentation processes. Figure 5.2 illustrates the complete workflow behind the creation of SciCap+.

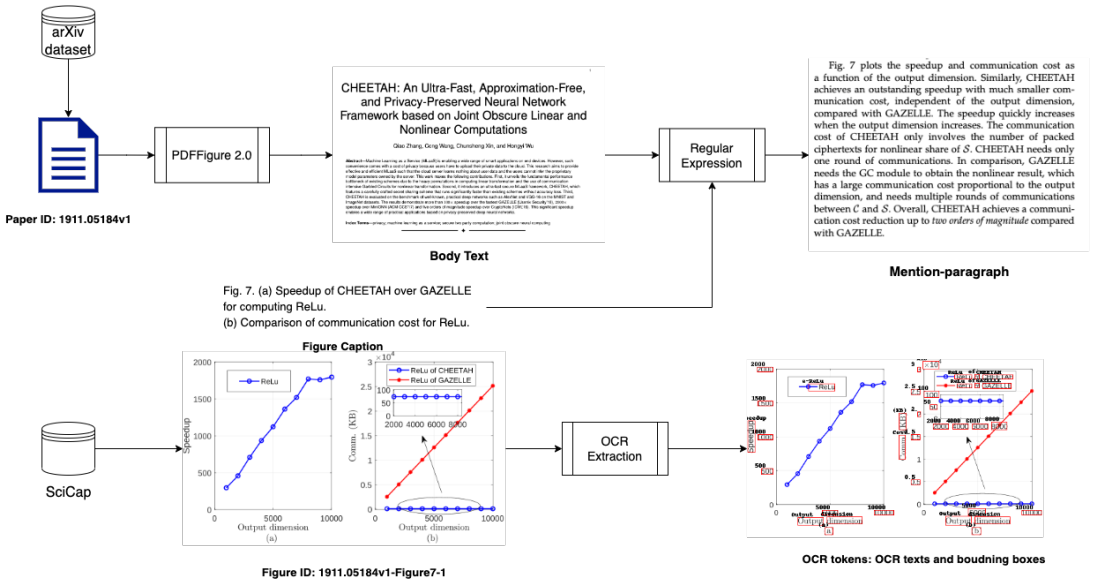


Figure 5.2: Overall workflow of the data augmentation for creating SciCap+ dataset. For each figure in SciCap+, we extracted its mention-paragraphs and OCR tokens (OCR texts and bounding boxes).

5.3.1 Mention-paragraph Extraction

We first obtained papers in PDF format from the Kaggle arXiv dataset[1]. While arXiv provides source files for some papers, we chose to work with PDFs since they offer a more consistent format, many papers either lack source files or have source files that are complex and inconsistent to parse.

For text extraction, we employed PDFFigures 2.0[2] [16], a specialized tool designed for extracting figures, captions, tables, and text from scholarly PDFs in computer science. After extracting the body text, we developed a systematic approach to locate mention-paragraphs. Since scholarly documents follow the convention of labelling figures with standardized numbering (e.g., "Figure 1" or "Fig. 1"), we implemented a regular expression pattern matching system that uses these figure numbers to identify and extract the corresponding paragraphs that reference each figure.

[1]https://www.kaggle.com/datasets/Cornell-University/arxiv
[2]https://github.com/allenai/pdffigures2

5.3.2 OCR Extraction

The original SciCap dataset includes text extracted from figures as metadata, but it does not provide spatial information on these texts. This spatial information is essential for understanding the relationship between textual elements in figures (For example, to recognize title, legend, and units).

We employed the Google Vision OCR API to perform a comprehensive OCR text extraction to address this limitation. This process captures the text content and locations (coordinates of their bounding boxes) of OCR tokens within each figure. Therefore, this process provides richer information about the spatial layout and organization of textual elements in scientific figures.

5.3.3 Data Statistics

The original SciCap dataset was split at the figure level, which meant figures from the same research paper could appear in different splits (train/validation/test). This created potential evaluation bias since information from related figures might leak across splits. To address this issue, we reorganized the dataset to split at the document level, ensuring all figures from the same paper remain in the same split.

Following the findings of [27] that text normalization and figure filtering do not improve model performance, we maintained several key design choices from the original dataset:

1. Retained original captions without normalization
2. Included all figures, regardless of whether they contain subfigures
3. Preserved the complete set of graphplots

For each figure, we augmented the original data with:

1. The first mention-paragraph from the body text
2. Complete bounding box information for OCR text

Table 5.1 positions SciCap+ in relation to previous figure-captioning datasets. Our dataset stands out through its use of real-world scientific papers and the inclusion of rich contextual knowledge.

Dataset	Images	In-context knowledge	Real-world Data
SciCap [27]	Figures	OCR(texts)	Yes
SciCap+	Figures	OCR(texts, bounding boxes) & mention-paragraphs	Yes
FigCAP [14, 13]	Bar, Line and Pie Charts	N/A	No
FigureQA [29]	Bar, Line and Pie Charts	N/A	No

Table 5.1: Comparison with the previous figure captioning datasets. The proposed SciCap+ dataset builds upon the SciCap dataset by incorporating additional in-context information and utilizing data from real-world scientific papers.

Split	Figures	Words
Training	394,005	12,336,511
Test	10,336	323,382
Validation	10,468	329,072

Table 5.2: Statistics of the SciCap+ dataset showing the distribution of figures and total word counts across training, test, and validation splits

Table 5.2 provides detailed statistics for SciCap+ (all figures are graphplots), showing the distribution across training, validation, and test splits. In all three splits, approximately 90% of the captions contained fewer than 66 words.

5.3.4 Dataset Quality Evaluation

Before conducting our main experiments, we performed a comprehensive quality evaluation of the SciCap+ dataset to verify the extraction quality of mention-paragraphs and OCR tokens. Our primary goal was to assess whether these extracted elements were accurate and relevant to their corresponding figures and captions. We randomly sampled 200 figures from the training set for detailed manual evaluation.

Two expert annotators independently evaluated each figure using a 5-point relevance scale:

- 1: No relevance
- 2: Low relevance
- 3: Moderate relevance

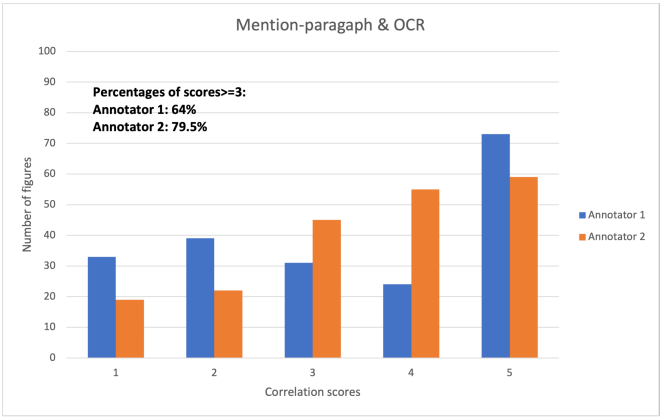


Figure 5.3: Score distribution on correlations between mention-paragraph, OCR tokens and figure captions. Both evaluators judged most figures and captions with at least moderate correlations with their mention-pargraphs and OCR tokens.

- 4: High relevance
- 5: Perfect relevance

We designed a two-step evaluation process, as illustrated below:

1. First, annotators examined each figure and assessed the relevance of the mention-paragraphs and OCR tokens separately.
2. Then, they provided a joint relevance score considering how well the mention-paragraphs and OCR tokens related to the figures and their captions.

Unlike natural image captioning, expert knowledge is crucial for evaluating scientific figures, where general visual comprehension might be insufficient. Therefore, we carefully selected two qualified annotators for this task: computer science Ph.D. holders currently working as researchers in natural language processing (NLP). Their academic and professional background ensured they possessed the necessary expertise to evaluate technical figures and their captions effectively.

Figure 5.3 shows the distribution of relevance scores. Both evaluators rated most figures positively, with 64% of evaluator 1 scores and 79.5% of evaluator 2 scores exceeding a relevance score of three. The inter-annotator agreement yielded a Cohen's kappa score of 0.28. These evaluation results demonstrate two

important findings: first, that the extraction quality of the mention-paragraphs and OCR tokens is satisfactory, as evidenced by the positive scores; and second, that these extracted elements maintained relevance to their corresponding figures and captions.

The relatively low Cohen's kappa score (0.28) reveals notable differences between annotators in assessing the relevance between the extracted mention paragraphs, OCR tokens and figures. We attributed this lower agreement to the inherently subjective nature of scientific figure captions evaluation. The task requires annotators to manually assess how well the text aligns with the visual content, an evaluation that heavily depends on:

1. Individual interpretation of technical content in figures and texts
2. Personal understanding of the relationship between figures and texts
3. Different thresholds for what constitutes high versus moderate relevance

The subjectivity in evaluation underscores a broader challenge in scientific figure captioning: even expert annotators may differ significantly in assessing figure-text relationships, highlighting the complexity of automating this task.

Figure 5.4 presents a case study illustrating significant scoring discrepancies between the two annotators. This example demonstrates how different evaluators can arrive at substantially different conclusions when assessing the same figure-text relationship.

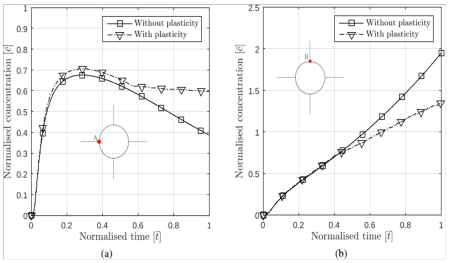
The evaluation by Annotator 1 focused on two main concerns:

1. The mention-paragraph contained excessive text, making it challenging to isolate the portions specifically relevant to the caption
2. The presence of incorrectly recognized Japanese words in the OCR tokens raised concerns about extraction quality

Based on these observations, Annotator 1 concluded there was only moderate relevance between the caption and the mention-paragraph, ultimately assigning a total score of 2 to reflect these limitations.

In contrast, Annotator 2 took a markedly different approach to evaluation:

1. Identified the presence of a critical keyword "two-way coupling" in the mention-paragraph



OCR tokens: {1.} {2.5} {Without} {plasticity} {-7:-} {With} {plasticity} {E} {Without} {plasticity} {-V-} {With} {plasticity} {0.9} {O} {0.8} {B} {0.7} {ヤ—} {マ—} {マ—マ—} {0.6} {1.5} {0.5} {0.4} {中} {0.3} {0.2} {0.5} {0.1} {0.6} {Normalised} {time} {[t]} {0.2} {0.4} {0.6} {0.8} {0.2} {0.4} {0.8} {1.} {Normalised} {time} {[t} {(a)} {(b)} {Normalised} {concentration} {C} {Normalised} {concentration} {[C} {2.}

Mention: In order to illustrate the effect of plasticity on the stress-diffusion interactions, we compare the results of pure elastic and elastoplastic material and two-way coupled model. Figure 9 shows the hydrostatic stress evolution at different points as a function of normalized time for pure elastic as well as elastoplastic material. The state of stress at a point A is initially tensile but changes to compressive due to continuous pulling and concentration evolution in the domain. Moreover, the state of stress changes from compressive to tensile at point B. The plastic yielding significantly reduces the stress level in the tensile site (point B) and makes it less compressive in the compressive site (point A). The level of concentration in the domain depends on the gradient of concentration as well as the localized state of stress. Tensile sites are capable to hold more concentration whereas compressed sites tries to push the concentration to near by sites. Figure 10a shows the concentration evolution at the point A. Initially the state of stress at this point is tensile and gradient of concentration is also present, the concentration of species increases with respect to time. But as the state of stress changes from tensile to compressive (at normalized time 0.3 please refer Figure 9a), the point A is no longer able to take/hold the concentration of the species and hence magnitude of concentration of

Caption: Figure 10: Two way coupling- Normalized concentration vs Normalized time at point (a) A (b) B for a plate with a hole with and without plasticity

Scores: Annotator 1: 2 , **Annotator 2:** 5

Figure 5.4: Case study on dataset quality evaluation. Two annotators subjectively weigh the contributions of mention-paragraphs and OCR tokens, resulting in significant differences in scores.

2. Valued the detailed explanations surrounding this key concept

3. Considered the incorrect OCR recognition of Japanese text as a minor issue that did not significantly impair overall comprehension

These factors led Annotator 2 to assign a perfect score of 5, highlighting a dramatically different interpretation of the same content.

This contrast in scoring (2 versus 5) illuminates a fundamental challenge in scientific figure caption evaluation: subjectiveness. Even with predefined scoring criteria, determining relevance remains highly subjective. Different annotators may:

1. Weight various aspects of the content differently

2. Have varying tolerances for imperfections in the extracted text

3. Place different emphasis on the presence of key technical terms

4. Assess the impact of errors differently based on their perceived importance to the overall understanding

5.4 Figure Captioning Model

The figure depicted in Figure 5.5 outlines the overall workflow of a scientific figure-captioning task. Our approach leverages the M4C-Captioner [52] as the baseline model to study the challenges of the task.

The M4C-Captioner is based on a multimodal multicopy mesh (M4C) [28], which jointly learns representations across input modalities. It incorporates a pointer network capable of selecting text from OCR tokens or a predefined fixed dictionary to address out-of-vocabulary issues during caption generation. This architecture provides two key advantages:

- Dynamic text selection between OCR-extracted tokens and a predefined vocabulary

- Effective handling of out-of-vocabulary terms, which is particularly crucial for scientific content where domain-specific terminology is common

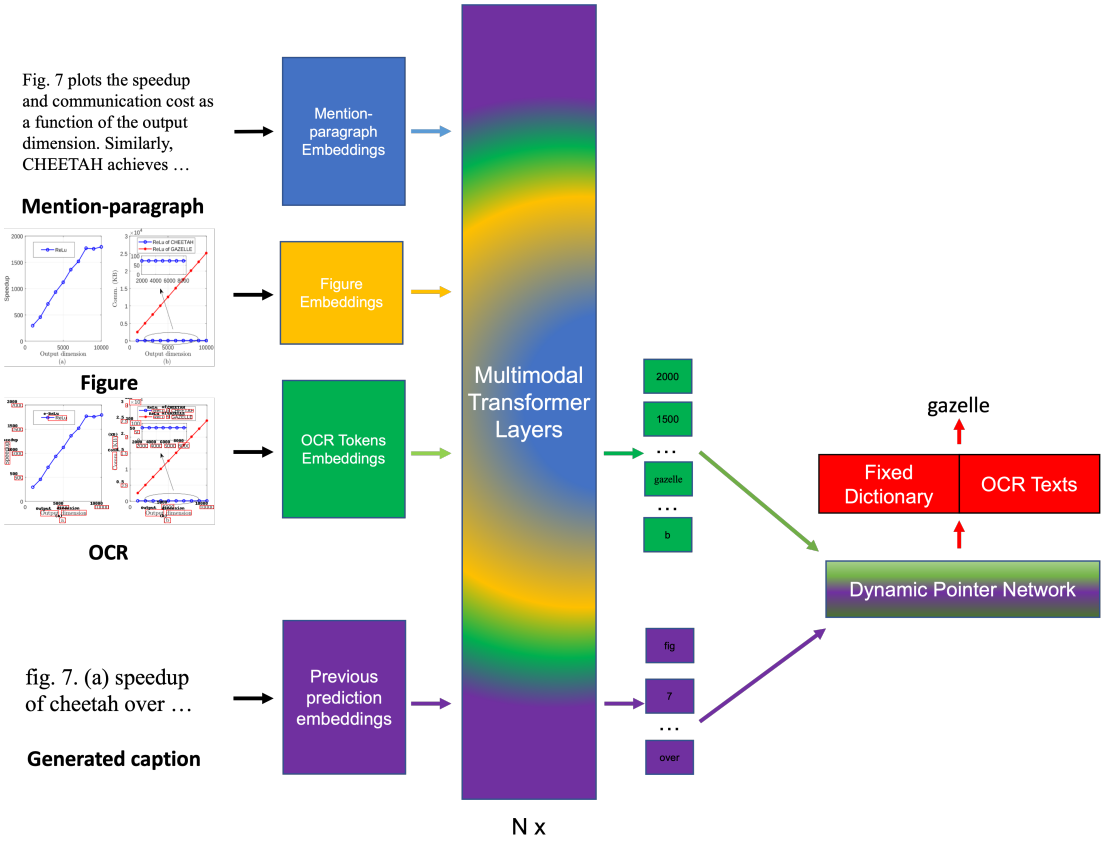


Figure 5.5: Overall framework for scientific figure captioning is centred around the architecture derived from M4C-Captioner [52]. This core component is designed to learn representations collaboratively from various input modalities. It incorporates a pointed network to choose text from OCR tokens or a predefined dictionary dynamically.

The multimodal nature of this model is particularly well-suited for scientific figure captioning tasks due to its capacity to simultaneously process visual features and textual information while preserving domain-specific terminology through its pointer-generator mechanism. This capability is essential for dealing with out-of-vocabulary issues during caption generation.

5.5 Method

The proposed multimodal approach for scientific figure captioning builds upon the M4C-captioner architecture. Our model jointly processes three input modalities to generate captions: mention-paragraphs, figures, and OCR tokens.

Given a figure I, its mention-paragraph P, and OCR tokens O, our model generates a caption $C = [c_1, c_2, ..., c_n]$ that describes the figure content.

5.5.1 Input Representations

Mention-Paragraph Encoding

For a mention-paragraph containing K words, we obtain a sequence of d-dimensional feature vectors $\{x_1^p, x_2^p, ..., x_K^p\}$ using a pretrained language model (LM):

$$x_i^p = \text{LM}(p_i) \in \mathbb{R}^d \tag{5.1}$$

Figure Visual Encoding

For the input figure I, we extract its visual feature using a pretrained vision encoder. For each figure, we applied a 2D adaptive average pooling over the outputs from layer 5 to obtain a global visual feature vector x^{fig}.

OCR Token Representation

For N OCR tokens detected in the figure, each token n is represented by:

1. Text embedding $x_n^{ft} \in \mathbb{R}^{text}$
2. Appearance feature $x_{apperance}^{fr}$ from a vision encoder

3. Symbol embedding $x_n^p \in \mathbb{R}^{symbole}$

4. Location feature $x_n^b = [x_{min}/W, y_{min}/H, x_{max}/W, y_{max}/H]$

These features are projected and combined:

$$x_n^{ocr} = \text{LN}(W_3 x_n^{ft} + W_4 x_n^{fr} + W_5 x_n^p) + \text{LN}(W_6 x_n^b) \tag{5.2}$$

5.5.2 Multimodal Transformer

The model applies L transformer layers over the concatenated sequence of all entity embeddings $[x^p; x^{fig}; x^{ocr}]$. Each transformer layer contains:

1. Multi-head self-attention

2. Feed-forward networks

3. Layer normalization and residual connections

Through self-attention, each entity can attend to all other entities regardless of their modality:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{5.3}$$

5.6 Experiments

We conducted a series of experiments using the SciCap+ dataset to empirically demonstrate that scientific figure captioning is a context-driven image captioning task that leverages information from textual and visual modalities. Our experimental design aimed to evaluate how different types of contexts (mention paragraphs, OCR-extracted text, and visual features) contribute to generating accurate and informative figure captions. Through these experiments, we sought to validate our hypothesis that effective scientific figure captioning requires the integration of multiple contexts rather than relying on visual information alone. The experimental framework was structured to address several key research questions:

1. How much does textual context from mention-paragraphs contribute to caption quality?
2. What role do OCR-extracted texts play in figure caption generation?
3. How do visual features enhance the caption generation process?
4. What is the relative importance of each type of context in producing informative captions?

5.6.1 Implementation and Training

Our implementation of the M4C-Captioner was based on the MMF framework [54] and Pytorch. The implementation allows users to specify diverse pre-trained encoders for each modality, which can be fine-tuned or frozen during training. The M4C-captioner has $D = 768$ hidden dimension size, $K = 4$ transformer layers and 12 attention heads. We used sentencepiece [34] to obtain a dictionary of 32000 subwords built from mention-paragraphs and OCR tokens. This is used as the vocabulary for M4C-captioner. We followed the BERT-BASE hyperparameter [20] settings and trained from scratch

Encoder Architecture

For the encoders feeding features to the M4C-captioner:

- **Vision Encoder:** We used a pre-trained Resnet-152 as the figure encoder. We applied a 2D adaptive average pooling over the outputs from layer 5 to obtain a 2048-dimensional global visual feature vector. During training, layers 2, 3, and 4 were fine-tuned to acquire specialized features from the figures [70].

- **Mention-paragraph Encoder:** SciBERT [8] was used to encode[3] text into 758-dimensional feature vectors. The number of vectors equals the number of sub-word tokens in the mention-paragraph, which we limit to 192 (around 90% mention-paragraphs have less than 190 words), including additional start and end sentence symbols. The encoder was fine-tuned during training.

- **OCR Token Encoder:** We used both text and visual features. For text, we employed FastText [1] as the word encoder and PHOC [10] as the character encoder. For visual features, we extracted Faster R-CNN fc6 features and applied fc7 weights to obtain 2048-dimensional appearance features for OCR token bounding boxes. The fc7 weights were fine-tuned during training. We retained a maximum of 95 OCR tokens per figure.

Training Details

We trained the model on eight Nvidia Tesla V100 GPUs, with training requiring approximately 13 hours for a complete feature set. We used a batch size of 128 and selected CIDEr as the primary evaluation metric. The training was evaluated every 2000 iterations and stopped if the CIDEr score did not improve for four consecutive intervals. We used Adam with a learning rate of 0.001 and $\epsilon = 1.0\text{E}{-}08$ for optimization. We implemented a multistep learning rate schedule with 1000 warmup iterations and a warmup factor 0.2. The maximum number of steps was set to 67 during decoding, including start and end sentence symbols.

Evaluation Metrics

We implemented a comprehensive set of metrics for evaluation to assess caption quality from multiple perspectives. The evaluation framework comprised five

[3]We only used the first 3 layers of SciBERT for lightweightness.

established text-based metrics: BLEU-4 [45] for assessing n-gram precision, METEOR [6] for handling synonyms and paraphrases, ROUGE-L [36] for measuring the longest common subsequence, CIDEr [61] for capturing consensus in human references, and SPICE [2] for evaluating semantic propositional content. Given that figure captions contain scientific terms which can be seen as uncommon words, we mainly focused on CIDEr since it emphasizes such terms.

To complement these text-only evaluations, we incorporated vision-and-language metrics CLIPScore and RefCLIPScore [26], which evaluate semantic similarities between generated captions and images.

Model	BLEU-4	METEOR	ROUGE-L	SPICE	CIDEr
1. MC (Figure Only)	1.5	5.6	15.4	4.3	4.6
2. MC (Mention Only)	5.3	11.0	27.4	14.3	49.0
3. MC (Figure and OCR)	2.6	7.6	20.5	10.1	22.2
4. MC (Figure and Mention)	**6.4**	11.5	27.9	14.6	50.5
5. MC (Figure, Mention and OCR)	6.3	**12.0**	29.2	15.8	55.8
Ablation Study on Figures					
6. MC (Mention and OCR)	6.3	**12.0**	**29.3**	**16.1**	**56.4**
Ablation Study on OCR features					
7. MC (Figure, Mention and Bounding boxes)	5.8	11.1	27.3	14.1	48.0
8. MC (Figure, Mention and OCR visual and OCR text features)	**6.4**	**12.0**	29.1	15.7	54.6
9. MC (Figure, Mention and OCR text features and Bounding boxes)	6.2	11.9	28.9	15.6	54.1

Table 5.3: Experimental results of different M4C-Captioner model configurations on the SciCap+ dataset. The main results section evaluates the effectiveness of incorporating different modalities. The ablation studies examine the impact of visual features and OCR information. The evaluation uses five standard image captioning metrics: BLEU-4, METEOR, ROUGE-L, SPICE and CIDEr. The results demonstrate that models leveraging textual and visual modalities consistently outperform single-modality (Figure-only) baselines. MC: M4C-Captioner

5.7 Results

5.7.1 Main Result

Exact-Matching Metrics

Experimental results in table 5.3 demonstrate that incorporating both mention-paragraph and OCR tokens (row #6) leads to substantial improvements across

Model	RefCLIPScore	CLIPScore
1. M4C-Captioner (Figure Only)	71.2	67.9
2. M4C-Captioner (Mention Only)	75.3	70.4
3. M4C-Captioner (Figure and OCR)	75.3	**72.2**
4. M4C-Captioner (Figure and Mention)	75.4	70.7
5. M4C-Captioner (Figure, Mention and OCR)	76.4	71.7
Ablation Study on Figures		
6. M4C-Captioner (Mention and OCR)	**76.5**	71.8
Ablation Study on OCR features		
7. M4C-Captioner (Figure, Mention and Bounding boxes)	75.2	70.4
8. M4C-Captioner (Figure, Mention and OCR visual and OCR text features)	76.3	71.7
9. M4C-Captioner (Figure, Mention OCR text and Bounding boxes)	76.3	71.6

Table 5.4: Performance comparison of M4C-Captioner variants using soft-matching metrics RefCLIPScore and CLIPScore on the SciCap+ dataset. Results demonstrate that incorporating knowledge from mention-paragraphs and OCR tokens substantially improves model performance compared to using visual features alone. The ablation analysis examines the impact of visual features and OCR components on caption generation. The highest RefCLIPScore achieved by the model without visual features (# 6) indicates that textual knowledge from mention-paragraphs and OCR tokens effectively captures semantic relationships for caption generation.

all five exact-matching metrics compared to the figure-only baseline (row #1). These findings strongly support our hypothesis that scientific figure captioning is inherently a context-driven image captioning task, where both OCR tokens and contextual knowledge from the mention-paragraphs play crucial roles in generating informative captions.

To establish a foundational understanding, we first implemented a baseline M4C-Captioner (Figure only) that uses figures as the sole input modality (row #1). Operating in a non-context setting, this baseline achieved relatively low scores across all metrics, highlighting the need for additional contextual information. The mention-only configuration in row #2 revealed the rich information content within mentions, demonstrated by a marked increase in performance. When we enhanced the figure input with OCR features (row #3), we observed notable improvements compared to the figure-only baseline (row #1), though these scores remained below those achieved with mentions alone (row #2). Given that mentions appeared to encode more valuable information than OCR, we combined mentions with figures as model inputs (row #4). This combination yielded further improvements across all metrics compared to the mention-only approach (row #2).

These promising results motivated us to explore the full integration of mentions and OCR tokens (row #5). This configuration outperformed all previous baselines: figure-only (row #1), mention-only (row #2), figure-OCR-only (row #3), and figure-mention-only (row #4). These comprehensive results demonstrate that explicitly leveraging multimodal contexts significantly enhances the quality of generated captions.

Through comprehensive ablation studies, we investigated the contribution of each feature to the overall performance. Removing visual feature vectors (row #6) led to a slight increase in CIDEr score, suggesting that visual features might introduce noise when the text modality provides sufficient context. When OCR information was withheld (row #4), the model could still leverage figure information to improve scores, though marginally compared to using mention-paragraphs alone (row #2). However, these scores remained inferior to configurations utilizing both OCR and mention-paragraphs (row #5 and row #6). This performance gap likely comes from the ResNet-152 visual encoder not being pretrained on scientific figures, resulting in less informative visual features when abundant textual information is available.

We incorporated text, visual, and spatial features to enhance OCR token representation. Our ablation analysis examined the impact of each OCR token feature, using row #5 as the primary comparison point, despite row #6 showing slightly superior performance. The complete removal of OCR features (row #4) resulted in a CIDEr score decrease of 5.3. Utilizing only OCR spatial features (row #7) led to a more substantial CIDEr score reduction of 7.8. The removal of OCR spatial features (row #8) caused a modest CIDEr score decline of 1.2, while eliminating OCR visual features (row #9) produced comparable results to the spatial feature removal.

These ablation findings highlight the significant contribution of enriched OCR token features to caption informativeness. Interestingly, while OCR token appearance features proved beneficial, removing figure visual features improved CIDEr scores. This result emphasizes the need for a specialized vision encoder trained specifically for scientific figures to extract meaningful visual features.

Soft-Matching Metrics

Using CLIPScore and RefCLIPScore to evaluate scientific figure caption generation provides a more comprehensive assessment than traditional text-based metrics like BLEU, METEOR, SPICE and CIDEr. While these classic metrics measure n-gram overlap, they struggle with semantic understanding and fail to consider whether a caption is relevant to the figure. CLIPScore and RefCLIPScore address this by directly evaluating how well a generated caption aligns with the visual content of a figure, which is particularly important in scientific figures, where accurate descriptions of trends, structures, or experimental results are essential.

RefCLIPScore enhances the CLIPScore by incorporating semantic similarity between the generated caption and reference captions, balancing image-text alignment (CLIPScore) with text-text relevance. This is crucial in scientific figure captioning, where scientific terminology matters. A caption might be visually grounded but fails to capture key scientific details. Combining CLIPScore for multimodal alignment and RefCLIPScore for textual accuracy ensures that captions are visually relevant and scientifically meaningful. This dual approach mitigates the weaknesses of purely lexical metrics and provides a more holistic evaluation framework for scientific figure captioning.

Table 5.4 presents the performance of different model variants evaluated using RefCLIPScore and CLIPScore. The figure-only baseline (row #1) achieves the lowest scores on both metrics (RefCLIPScore: 71.2, CLIPScore: 67.9), indicating that figures alone provide insufficient information for generating informative captions. This finding aligns with the observations from exact-matching metrics.

When incorporating mention-paragraphs (row #2), the model demonstrates substantial improvement in both RefCLIPScore (75.3) and CLIPScore (70.4). A notable discrepancy between the two metrics emerges in the model variant (row #3), which integrates OCR with visual features. This variant achieves the highest CLIPScore (72.2) among all models while maintaining a moderate RefCLIPScore (75.3). The disparity suggests that while the model excels at generating captions that align well with image content (high CLIPScore), these captions may deviate from the reference captions provided by human authors (lower RefCLIPScore). This observation reveals an important tension between image-text alignment and adherence to human-written reference captions.

The ablation studies reveal several important patterns. First, removing visual features (row #6) leads to the highest RefCLIPScore (76.5) but a lower CLIPScore (71.8) compared to the model variant (row #3). This inverse relationship further emphasizes the trade-off between semantic accuracy and image-text alignment. The model without visual features better captures the semantic content of reference captions but may generate text that is less grounded in the visual content.

The OCR ablation experiments (row #7-9) reveal the distinct contributions of different OCR components. Model variant (row #7), which uses only bounding boxes for OCR information, achieves lower scores on both metrics (RefCLIPScore: 75.2, CLIPScore: 70.4). Model variant (row #8), which incorporates both OCR visual features (appearance features from bounding boxes) and OCR text features, demonstrates improved performance (RefCLIPScore: 76.3, CLIPScore: 71.7). Similarly, model variant (row #9,) which uses OCR text features with bounding box locations but excludes OCR visual features, achieves comparable scores (RefCLIPScore: 76.3, CLIPScore: 71.6). These results reveal two key findings: first, textual information extracted from OCR contributes more significantly to caption generation quality than spatial information from bounding boxes alone; second, the OCR visual features (appearance features) do not provide substantial additional benefits when OCR text features are already present.

These results complement the exact-matching metric findings and further support the hypothesis that scientific figure captioning benefits significantly from multimodal knowledge integration. The systematic discrepancies between RefCLIPScore and CLIPScore across different model variants highlight a fundamental challenge in scientific figure captioning: scientific figures often require extensive research context and domain knowledge to be interpreted appropriately, making it difficult to simultaneously optimize for both semantic accuracy (RefCLIPScore) and visual grounding (CLIPScore).

A key characteristic of scientific figure captioning is that it goes beyond simply describing visual content. For example, a figure caption may include research methods that are not directly visible but are essential for interpreting the results presented in the figure. This highlights that, unlike conventional image captioning, scientific figure captions must incorporate research context to ensure clarity and accuracy. The soft-machine metric evaluation results demonstrate the importance of evaluating scientific figure captions from multiple perspectives to ensure they meet the visual description requirements and the need for precise scientific communication.

5.8 Human Evaluation

We conducted a human caption generation study to complement our automatic evaluation metrics. Our primary research question was whether humans could outperform models in writing figure captions when given the same input conditions.

Interpreting scientific figures requires technical knowledge and familiarity with academic writing conventions. Selecting annotators with domain expertise was particularly crucial for these evaluations. Therefore, we selected two expert annotators with Ph.D. degrees in computer science and research experience in natural language processing to perform both tasks.

5.8.1 Figure Caption Generation Task

Evaluation Step

We designed the caption generation task with two distinct experimental conditions to systematically evaluate human performance against our model:

1. **Figure-only condition:** Annotators generated captions using only the figures as input, paralleling the basic configuration of the M4C-Captioner with access to figures and OCR features.

2. **Figure-Mention condition:** Annotators wrote captions with access to both figures and their corresponding mention-paragraphs, allowing us to assess the impact of additional contextual information.

For this evaluation, we randomly sampled 100 figures from the test set and compared human-generated captions with those produced by the M4C-Captioner under various configurations.

Evaluation Results and Analysis

Annotator	Inputs	BLEU-4	METEOR	ROUGE-L	SPICE	CIDEr
1. Annotator 1	Figure-only	2.4	8.3	13.2	9.4	14.6
2. Annotator 2	Figure-only	**3.8**	**10.1**	**21.5**	8.9	**23.8**
3. M4C-Captioner	Figure and OCR	3.6	7.6	20.5	**11.5**	18.7
4. Annotator 1	Figure-Mention	**7.7**	13.4	19.1	15.9	11.3
5. Annotator 2	Figure-Mention	7.5	**14.8**	24.8	14.3	18.8
6. M4C-Captioner	Mention, Figure and OCR	5.5	11.6	**28.1**	**16.1**	**47.7**

Table 5.5: Automatic exact-matching evaluation scores on human-generated captions. The model has similar performances when the figure is the only available source. Using information from vision and text modality, the model gains more on CIDEr scores.

Table 5.5 and Table 5.6 present evaluation results comparing human annotators and the M4C-Captioner model using exact-matching and soft-matching metrics. The results reveal distinct patterns in how humans and machines leverage different information modalities for caption generation. Human annotators demonstrated comparable or superior performance to the M4C-Captioner model

Annotator	Inputs	RefCLIPScore	CLIPScore
1. Annotator 1	Figure-only	77.0	**76.8**
2. Annotator 2	Figure-only	**77.7**	76.4
3. M4C-Captioner	Figure and OCR	74.4	71.1
4. Annotator 1	Figure-Mention	75.5	71.4
5. Annotator 2	Figure-Mention	**78.7**	**75.7**
6. M4C-Captioner	Mention, Figure and OCR	76.3	72.0

Table 5.6: Automatic soft-matching evaluation scores on human-generated captions. Humans obtained higher RefCLIPscore and CLIPScore than models.

across most metrics in the figure-only condition. With access to only visual information, Annotator 2 achieved the highest scores in BLEU-4 (3.8), METEOR (10.1), ROUGE-L (21.5), and CIDEr (23.8), while the M4C-Captioner model performed best in SPICE (11.5). The soft-matching metrics further reinforce this pattern, where Annotator 1 and Annotator 2 achieved RefCLIPScores of 77.0 and 77.7, respectively, surpassing the model score of 74.4. These results indicate that humans are more capable of interpreting and describing scientific figures without additional context.

The introduction of mention-paragraphs produced notable performance shifts across all participants. The M4C-Captioner model showed substantial improvement in exact-matching metrics, particularly in the CIDEr score, which increased from 18.7 to 47.7. Human annotators also demonstrated improvements in most metrics, with Annotator 1 achieving the highest BLEU-4 score (7.7) and Annotator 2 obtaining the highest METEOR score (14.8). The soft-matching metrics reveal a similar trend, with Annotator 2 achieving the highest scores (RefCLIPScore: 78.7, CLIPScore: 75.7).

A notable pattern emerges when comparing RefCLIPScore and CLIPScore values. The disparity between these metrics becomes particularly pronounced under the mention-paragraph condition, where Annotator 2 achieved a RefCLIPScore of 78.7 but a CLIPScore of 75.7, a 3-point difference. Given that RefCLIPScore evaluates caption quality against reference captions. At the same time, CLIPScore focuses on image-caption alignment. This widening gap suggests that captions generated with mention-paragraphs contain more contextual informa-

tion not directly depicted in the figures. This observation is further supported by the substantial increase in CIDEr scores when mention-paragraphs are provided, particularly for the M4C-Captioner model. These findings highlight the complex nature of scientific figure captioning, where success depends on accurate visual interpretation and effective integration of contextual knowledge.

Case Studies and Analysis

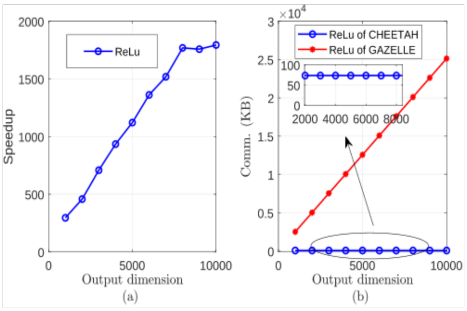
This section presents a case study (Figure 5.6) comparing human-generated and machine-generated figure captions. Our analysis reveals that the mention-paragraphs serve as a critical source of contextual information, enabling the creation of self-contained captions that combine research context with visual interpretation. Furthermore, we note the subjective nature of caption writing styles, which can vary significantly among authors.

Without the mention-paragraphs, annotators interpreted figures solely based on visual content. This leads to human annotators writing captions that merely enumerate visual components without providing a deeper contextual understanding or scientific interpretation.

Mention-paragraphs significantly improved the quality of human-generated captions by providing essential contextual information. With access to this contextual information, annotators could correctly interpret domain-specific terminology and abbreviations - for example, successfully translating “Comm. (KB)” to “communication cost.” The challenge of handling structural elements remained, as annotators consistently struggled to appropriately address subfigures despite having access to the additional context.

The model exhibited similar but distinct behaviours across different input configurations. It demonstrated an interesting trade-off when limited to the figure and its OCR tokens without mention-paragraphs. While successfully identifying and separating subfigures (generating distinct captions for (a) and (b)), it struggled to meaningfully integrate OCR text content (such as ”CHEETAH” and ”GAZELLE”) into these captions. This suggests that the model learned structural visual patterns but lacked the contextual understanding to utilize textual elements effectively.

The model defaulted to an extractive approach with access to only the mention-paragraphs, directly copying the first relevant sentence. When provided with both



Mention-paragraph:
Fig. 7 plots the speedup and communication cost as a function of the output dimension. Similarly, CHEETAH achieves an outstanding speedup with much smaller communication cost, independent of the output dimension, compared with GAZELLE. The speedup quickly increases when the output dimension increases. The communication cost of CHEETAH only involves the number of packed ciphertexts for nonlinear share of S. CHEETAH needs only one round of communications. In comparison, GAZELLE needs the GC module to obtain the nonlinear result, which has a large communication cost proportional to the output dimension, and needs multiple rounds of communications between C and S. Overall, CHEETAH achieves a communication cost reduction up to two orders of magnitude compared with GAZELLE.

Caption:
Fig. 7. (a) Speedup of CHEETAH over GAZELLE for computing ReLu. (b) Comparison of communication cost for ReLu.

Human-generated captions

Without mention-paragraphs:
Annotator 1:
The left graph shows the relationship between the output dimension and the speed up in the case of ReLu. The speed increases as the dimension size increases and hit the plateau at the dimension size of 8000. The right graph shows the relationship between Comm, and the dimension size in the cases of ReLu of CHEETAH and ReLu of GAZELL. The Comm, of ReLu of CHEETAH stays about 75 over all dimension sizes whereas that of ReLu of GAZELLE increases linearly to the dimension size.

Annotator 2:
Speedup vs output dimenson of the ReLu (left) and Communication speed in KB vs output dimension of the ReLU of CHEETAH and GAZELLE. The cost of GAZELLE increses as the output dimension increases but the cost of CHEETAH remains the same regardless of the output dimension.

With mention-paragraphs:
Annotator 1:
Speedup and communication cost as a function of the output dimension are shown. CHEETAH achieves an outstanding speedup with much smaller communication cost, independent of the output dimension, compared with GAZELLE. The speedup quickly increases when the output dimension increases.

Annotator 2:
Speedup and comminication costs of ReLu as a function of its output dimension. Since CHEETAH only uses packed ciphertexts, it is less costly than GAZELLE which relies on the GC modeule.

Model-generated captions

M4C-captioner (Figure, OCR)
fig. 3. (a) speedup of relu and relu with different dimensions. (b) comparison of the computation time of relu and relu with different dimensions.

M4C-captioner (Mention-paragraph)
figure 7: speedup and communication cost as a function of the output dimension.

M4C-captioner (Mention-paragraph, OCR)
fig. 7. the speedup and communication cost of chetah and gazelle with different output dimension.

M4C-captioner (Mention-paragraph, Figure, OCR)
fig. 7. speedup and communication cost as a function of the output dimension.

Figure 5.6: Case study on human-generated and model-generated captions. The mention-paragraph provides major information for the model and human annotators to compose informative captions.

the mention-paragraphs and OCR tokens, it demonstrated an improved ability to incorporate textual information and generate accurate summaries. When integrating with full features (mention-paragraphs, OCR tokens, and figures), the model failed to handle subfigure structure while incorporating textual contexts, suggesting challenges in balancing multiple competing objectives.

Annotators employed different caption writing styles. Annotator 1 wrote captions with details including all available information, while Annotator 2 favoured a concise writing style. These phenomenons revealed significant variation in caption writing styles.

We also found that neither annotator naturally followed the ground-truth caption style, particularly in treating subfigures. This divergence in writing style illustrates the inherently subjective nature of scientific caption writing and raises important questions about standardization in scientific communication.

These observations suggest that effective scientific figure captioning requires access to multiple context sources and sophisticated mechanisms for integrating them coherently. The challenge lies in extracting relevant information from each modality and combining them in a way that maintains structural consistency while conveying accurate scientific content. The result also indicates that we need to standardize scientific figure caption writing style to facilitate effective knowledge sharing.

5.9 Conclusion

This study investigated the challenges of scientific figure captioning by reframing it as a knowledge-augmented image-captioning task. Building upon prior work [27], we introduced SciCap+, an enhanced version of the SciCap dataset that incorporates mention-paragraphs and OCR tokens alongside figures.

Using the M4C-captioner model as a baseline, we demonstrated the effectiveness of integrating knowledge from three key modalities: mention-paragraphs, figures, and OCR tokens. Our automatic evaluation experiments revealed significant improvements in performance metrics when utilizing this multimodal knowledge approach. Notably, the model-generated captions outperformed human-generated ones according to automatic evaluation metrics.

Through human evaluation, we found that generating informative and accurate scientific figure captions remains challenging for humans and machines. The

release of SciCap+ represents an important contribution to the field, providing a foundation for advancing scientific figure captioning research and development.

For future work, we aim to expand our pool of annotators and conduct preliminary studies to select high-quality figure captions that provide detailed in-context and visual interpretations for human evaluation and generation tasks. In this study, we performed pre-extraction of mention-paragraphs from papers and OCR tokens from figures. This preprocessing step represents another limitation of the paper. Another of our future work aims to develop methods to automatically reference knowledge from entire papers and perform automatic OCR token extractions.

6 Conclusion

This thesis has investigated context-driven caption generation through systematic empirical analysis across two context-driven image captioning tasks: news images and scientific figures captioning. Our research has comprehensively answered the core research questions while making methodological, analytical and resource contributions to the field.

Impact of Contextual Information on Caption Generation Two studies have revealed the critical role of multimodal contexts, especially textual contexts, in context-driven caption generation tasks. For the first study, news image captioning, our study demonstrates: 1. Textual context from news articles provides fundamental information for generating news image captions. 2. Incorporating article context significantly improves caption quality compared to vision-only approaches 3. News captions require understanding visual content and its relationship to the broader textual context.

For the second study, scientific figure captioning, our study showed that 1. Domain-specific textual context from the mention-paragraphs is essential for accurate figure interpretation 2. OCR texts from figures also provide technical information needed for scientific figure captioning

Relative Importance of Textual versus Visual Context Our research also provides clear insights into the contribution of textual versus visual contexts. In the news image captioning task, we found that 1. Textual context from news articles proved to be the primary driver of caption quality. 2. Visual features provide complementary information but are secondary to textual context. 3. Integrating both modalities produces optimal results, though textual features have a more significant impact.

Our study on the scientific figure captioning task reveals that 1. The mention-paragraphs from papers provide essential domain knowledge and technical con-

text. 2. OCR content provides additional contextual information. 3. Visual features of OCR texts contribute to figure caption generation, not figures.

Overall Contributions Our studies are based on transformer-based models, demonstrating the effectiveness of transformer-based architectures for multimodal context integration. We also found the need to improve visual feature extraction for news images and scientific figures.

In both studies, we developed comprehensive approaches combining automatic and human evaluation, guaranteeing automatic metrics validation. The developed evaluation approach can be used to evaluate similar context-driven image captioning tasks.

Despite the contribution of the methodological and evaluation framework, we also contribute to the resource for future research in context-driven image captioning tasks. We refined the existing dataset with complete news articles and fair train/test splits for news image captioning. For scientific figure captioning, we enhanced the existing SciCap dataset with important contextual information: the mention-paragraphs from main body texts and OCR texts recognized from figures. Therefore, we reframe the definition of the scientific figure captioning task from image-to-text to a knowledge-augmented image-captioning task.

Overall Impacts, Future Work and Applications Our study on context-driven image captioning tasks impacts multimodal learning research. It demonstrates the importance of integrating contextual information for context-driven caption generation tasks, the model must go beyond image-to-text translation. Our findings on textual modality weights more in news image captioning and scientific figure captioning align with the recent research on multimodal large language model, where language model encoder contains significantly more parameters than vision encoder [64].

Several promising directions exist for future research. For scientific figure captioning, we can make improvements in developing methods for extracting and utilizing contextual information from research papers. Rather than relying on pre-extracted mention-paragraphs and OCR text, future work should develop end-to-end approaches that automatically identify and integrate relevant contextual information from entire papers.

Another direction is improving visual feature extraction for news images and

scientific figures. In this study, we used vision models pretrained on natural images, which may not effectively extract domain-specific visual features from news images and figures. Training vision encoders on news images and scientific figures could enhance caption generation quality.

The application of this research is to assist humans in news and scientific publishing. In journalism, the news image captioning system can assist editors and journalists in instantly generating contextually aware captions when journalists upload images with articles, saving time in fast-paced news publications. For scientific publication support, assists authors in generating comprehensive figure captions that accurately describe their methods and results.

Bibliography

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. URL https://arxiv.org/pdf/1409.0473.pdf.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016. URL https://arxiv.org/abs/1409.0473.

[6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[7] Vishwash Batra, Yulan He, and George Vogiatzis. Neural caption generation for news images. In *International Conference on Language Resources and Evaluation (LREC)*, 2018. URL https://www.aclweb.org/anthology/L18-1273.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.

[9] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12466–12475, 2019. doi: 10.1109/CVPR.2019.01275.

[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017. URL http://dblp.uni-trier.de/db/journals/tacl/tacl5.html#BojanowskiGJM17.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[13] Chen Charles, Zhang Ruiyi, Koh Eunyee, Kim Sungchul, Cohen Scott, and Rossi Ryan. Figure captioning with relation maps for reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[14] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850*, 2019.

[15] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179.

[16] Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 143–152. IEEE, 2016.

[17] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014. doi: 10.3115/v1/W14-3348.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[21] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[22] Yansong Feng and Mirella Lapata. Automatic caption generation for news

images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812, 2013. doi: 10.1109/TPAMI.2012.118.

[23] Christopher R Fetsch and Uta Noppeney. How the brain controls decision making in a multisensory world, 2023.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[26] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595.

[27] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.277. URL https://aclanthology.org/2021.findings-emnlp.277.

[28] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[29] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

[30] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.277. URL https://aclanthology.org/2022.acl-long.277.

[31] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015. doi: 10.1109/CVPR.2015.7298932.

[32] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[33] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980.

[34] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

[35] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8928–8937, 2019.

[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[37] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *42nd Annual Meeting on Association for Computational Linguis-*

tics (ACL), pages 605–612, 2004. doi: 10.3115/1218955.1219032. URL https://www.aclweb.org/anthology/P04-1077.

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[39] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2286–2293, 2021.

[40] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[41] Manuel R Mercier and Celine Cappe. The interplay between multisensory integration and perceptual decision making. *NeuroImage*, 222:116970, 2020.

[42] George A Miller and Christiane Fellbaum. Wordnet then and now. *Language Resources and Evaluation*, 41(2):209–214, 2007. URL http://wordnet.princeton.edu/.

[43] Jahangir Moini, Anthony LoGalbo, and Raheleh Ahangari. Chapter 7 - sensation and perception. In Jahangir Moini, Anthony LoGalbo, and Raheleh Ahangari, editors, *Foundations of the Mind, Brain, and Behavioral Relationships*, pages 115–123. Academic Press, 2024. ISBN 978-0-323-95975-9. doi: https://doi.org/10.1016/B978-0-323-95975-9.00003-2. URL https://www.sciencedirect.com/science/article/pii/B9780323959759000032.

[44] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, 2019. doi: 10.18653/v1/N19-4009. URL https://www.aclweb.org/anthology/N19-4009.

[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of*

the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, 2002. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[48] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[49] Martin F Porter. Snowball: A language for stemming algorithms, 2001. URL http://snowball.tartarus.org/texts/.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[51] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5): 1072–1085, 2018. doi: 10.1109/TPAMI.2017.2721945.

[52] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, 2020.

[53] Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. Figureseer: Parsing result-figures in research papers. In *European Conference on Computer Vision*, pages 664–680. Springer, 2016.

[54] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020.

[55] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. URL http://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

[56] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.115. URL https://aclanthology.org/2021.acl-long.115.

[57] I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[58] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

[59] Amara Tariq and Hassan Foroosh. A context-driven extractive framework

for generating realistic image descriptions. *IEEE Transactions on Image Processing*, 26(2):619–632, 2017. doi: 10.1109/TIP.2016.2628585.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[61] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[62] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015. doi: 10.1109/CVPR.2015.7298935.

[63] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[65] Wei Wei, R Austin Benn, Robert Scholz, Victoria Shevchenko, Ulysse Klatzmann, Francesco Alberti, Rocco Chiou, Demian Wassermann, Tamara Vanderwal, Jonathan Smallwood, et al. A function-based mapping of sensory integration along the cortical hierarchy. *Communications Biology*, 7(1):1593, 2024.

[66] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*, pages 2048–2057, 2015.

[67] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015. URL http://proceedings.mlr.press/v37/xuc15.html.

[68] Zhishen Yang and Naoaki Okazaki. Image caption generation for news articles. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1941–1951, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.176. URL https://aclanthology.org/2020.coling-main.176.

[69] Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing*, 31(3):1140–1165, 2024.

[70] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[71] Qiao Zhang, Cong Wang, Chunsheng Xin, and Hongyi Wu. Cheetah: An ultra-fast, approximation-free, and privacy-preserved neural network framework based on joint obscure linear and nonlinear computations. *arXiv preprint arXiv:1911.05184*, 2019.

[72] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. doi: 10.1109/TPAMI.2017.2723009.

Publication List

Journals (Peer-Reviewed)

1. Zhishen Yang, Tosho Hirasawa, Mamoru Komachi, Naoaki Okazaki. Why videos do not guide translations in video-guided machine translation? An empirical evaluation of video-guided machine translation dataset, *Journal of Information Processing*, Vol. 30, pp. 388-396, May 2022.

2. Zhishen Yang, Raj Dabre, Hideki Tanaka, Naoaki Okazaki, SciCap+: A Knowledge Augmented Dataset to Study the Challenges of Scientific Figure Captioning, *Journal of Natural Language Processing*, Volume 31, Issue 3, pp. 1140-1165, September 2024.

International Conferences (Peer-Reviewed)

1. Zhishen Yang, Raj Dabre, Hideki Tanaka, Naoaki Okazaki. SciCap+: A Knowledge Augmented Dataset to Study the Challenges of Scientific Figure Captioning, In *Proceedings of the Workshop on Scientific Document Understanding, co-located with 37th AAAI Conference on Artificial Intelligence (CEUR Workshop)*, Feb. 2023.

2. Zhishen Yang, Lars Wolfsteller, Naoaki Okazaki. TextLearner at SemEval-2020 Task 10: A Contextualized Ranking System in Solving Emphasis Selection in Text, In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval 2020)*, pp. 1691-1697, Dec. 2020.

3. Zhishen Yang, Naoaki Okazaki. Image caption generation for news articles, In *Proceedings of the 28th International Conference on Computational Linguistics (COLING2020)*, pp. 1941-1951, Dec. 2020.

4. Zhishen Yang, Sam Vijlbrief, Naoaki Okazaki. Emotion-related Symbols in Emotion Detection, In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pp. 350-354, June 2019.

International Conferences (Non Peer-Reviewed)

1. Tosho Hirasawa*, Zhishen Yang*, Naoaki Okazaki, Mamoru Komachi. Keyframe Segmentation and Positional Encoding for Video-guided Machine Translation Challenge 2020, *First Workshop on Advances in Language and Vision Research (ALVR 2020)*, July 2020. (*Equal Contribution)

Domestic Conferences (Non Peer-Reviewed)

1. Zhishen Yang, Tosho Hirasawa, Edison Marrese-Tyalor, Naoaki Okazaki. Large Language Models as Manga Translators: A Case Study, 言語処理学会第 *30* 回年次大会 *(NLP2024)*, pp. 2012-2017, Mar. 2024.

2. Zhishen Yang, Raj Dabre, Hideki Tanaka, Naoaki Okazaki. Knowledge-Augmented Figure Caption Generation, 言語処理学会第 *29* 回年次大会 *(NLP2023)*, pp. 460-465, Mar. 2023.

3. 遠藤洸亮, Zhishen Yang, 岡崎直観. 画像キャプション生成における JPEG 圧縮への頑健性の改善, 言語処理学会第 *29* 回年次大会 *(NLP2023)*, pp. 419-424, Mar. 2023.

4. Zhishen Yang, Naoaki Okazaki. News Image Caption Generation, 第 *34* 回人工知能学会全国大会 *(JSAI2020)*, June 2020.