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
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学系 知能情報コース	申請学位（専攻分野）： Academic Degree Requested	博士（学術） Doctor of
学生氏名： Student's Name	Yang Zhishen	審査員主査： Chief Examiner	岡崎直観

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

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.

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 evaluations, 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.