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)	YANGZhishen
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,徳永 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)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		Yang Zhishen	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	岡崎 直観	教授	審査員	荒瀬 由紀	教授
	審査員	徳永 健伸	教授			
		井上 中順	准教授			
		篠田 浩一	教授			

論文審査の要旨（2000字程度）

本論文は「An Investigation of Context-Driven Caption Generation」（文脈駆動型のキャプション生成に関する研究）と題し、英文6章で構成されている。本研究はマルチモーダル統合における基本的な課題、すなわち視覚的特徴とテキスト的特徴を効果的に組み合わせる方法と、それらがキャプションの品質に与える影響を評価している。このとき、画像のみに基づくキャプション生成手法と比較して、文脈情報がキャプションの品質を向上させることを示している。

第1章「Introduction」（序論）では、従来の画像キャプション生成と対比しながら、文脈駆動型キャプション生成を説明している。従来手法は視覚的内容を記述することのみに焦点を当てるのに対し、本研究の文脈駆動型キャプション生成では、より広範な情報を統合する必要がある。本章では、視覚情報とテキスト情報の統合、それらのキャプション生成における重要性、およびニュースメディアと科学文献における特有の課題に言及し、文脈駆動型キャプション生成の重要性を述べている。

第2章「Background Knowledge」（背景知識）では、本研究で提案する手法の基盤となる系列変換モデル（sequence-to-sequence models)、注意機構、Transformerアーキテクチャについて、基礎事項をまとめている。また、評価指標の詳細として、BLEUやMETEORなど自然言語処理で旧来から用いられてきた指標に加え、CIDErやCLIPScoreといった画像キャプション向けの専門的な評価指標を紹介している。この基礎知識は、後続の章の基盤となるものである。

第3章「Related Work」（関連研究）では、画像キャプション生成、ニュース画像キャプション生成、および科学文献の図のキャプション生成に関する包括的なサーベイを行っている。このサーベイでは、当該分野の発展の経過を説明しながら、ニュース画像および科学文献の図のキャプション生成において文脈が果たす役割を明らかにし、本論文の貢献をより広い研究領域の中で位置づけている。

第4章「News Image Caption Generation」（ニュース画像キャプション生成）では、視覚情報とテキスト情報を統合するTransformerベースのアーキテクチャを提案している。提案手法は、Transformerアーキテクチャのキーとバリューに画像エンコーダの出力を与え、画像の情報をテキスト処理のTransformerに取り込む。自動評価指標と人手評価の両方を用いた包括的な実験を通じて、①ニュース記事のテキスト文脈は、適切なキャプションを生成する上で不可欠な情報を含む、②提案手法はテキストのみのアプローチよりも品質のよいキャプションを生成し、当時の最先端手法を大幅に上回る性能を示すことを報告している。人手評価によると、提案手法は従来手法よりも情報量が多く文脈的に適切なキャプションを生成するが、報道における微妙な意図を捉えることに課題が残ることが明らかになった。本章は、ニュースのキャプション生成におけるテキストと画像の相互作用に関する理解を深め、ニュースにおけるキャプションの自動生成に向けた実践的な解決策を提供している。

第5章「Scientific Figure Caption Generation」（科学文献の図のキャプション生成）では、学術文献の図に関するデータセットSciCapに対して、論文中で図に言及している箇所と図中のテキストのOCR結果を追加したSciCap+データセットを導入している。このデータセットを用いて、科学文献の図のキャプション生成において、論文から得られるテキスト情報や図中のテキスト情報の統合が必要であるか検証している。実験の結果、視覚情報のみを使用するベースラインと比較して、図に言及している段落やOCRテキストを組み込むことで、キャプションの品質が大幅に向上することが明らかになった。また、アブレーション実験では、異なる情報源の寄与を体系的に分析している。人手評価の結果から、提案するマルチモーダル手法の有効性が示されたが、科学文献の図のキャプション生成にはドメイン知識が必要であるため、人間にとっても難しいタスクである、という

課題も明らかになった。本章は、科学文献の図のキャプション生成のリソースと手法の両方の開発に貢献し、特に、構築された SciCap+データセットは研究コミュニティにとって貴重なデータである。

第 6 章「Conclusion」（結論）では、本論文で得られた知見を総括し、文脈駆動型キャプション生成においては複数のモダリティの効果的な統合が不可欠であると結論付けている。また、文脈駆動型キャプション生成の将来展望を述べている。

本論文は、ニュース画像キャプション生成と科学文献の図のキャプション生成という 2 つのマルチモーダル言語生成タスクに取り組み、複数のモダリティを統合する新しい Transformer アーキテクチャ、テキストと画像の文脈の重要性に関する知見、後続の研究に貢献するデータセット(SciCap+) を提供している。マルチモーダル基盤モデルの研究だけでなく、自動ジャーナリズム、自動研究など、今後有望視されている人工知能の実現にも寄与する。よって、本論文は博士（学術）の学位論文として十分価値あるものと認める。