	グェン ヴァン クワン
氏　　名	Nguyen Van Quang
学位の種類	博　士（情報科学）
学位記番号	情　博　第　７８４　号
学位授与年月日	令和４年　９月２６日
学位授与の要件	学位規則第４条第１項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）システム情報科学 専攻
学位論文題目	Machine Intelligence that Understands Visual and Linguistic Information and Interacts with Humans and Environments (視覚と言語情報を理解し人間や環境と作用し合う機械知能)
論文審査委員	（主査）東北大学教　授　岡谷　貴之 東北大学教　授　橋本　浩一　　東北大学教　授　乾　健太郎 東北大学准教授　鏡　慎吾

論 文 内 容 の 要 旨

第１章 Introduction

Over the past years, Artificial Intelligence has witnessed significant progress in computer vision and natural language processing thanks to deep learning advancements. Inspired by remarkable success in these two independent fields, there has been a growing interest in the problems at the intersection of visual and linguistic understanding. It is believed that advances in solving those mentioned above and related problems would open the door to many real-world applications, bringing fundamental change to society. Take virtual assistants that aid the visually impaired, automatic surveillance systems for querying over visual databases, and in-home robots that perform household tasks as examples. Thus, the integration of vision and language is a viable approach to achieving one of AI's visionary goals: building machines that can understand both the visual and linguistic worlds, communicate with humans in natural language, and further interact with environments. In this dissertation, we aim to build and improve agents endowed with such intelligence as a continuation of collective efforts by research communities. Specifically, we focus our attention on three representative vision language tasks, namely **image captioning**, **visual dialog**, and **interactive instruction following tasks**.

第２章 GRIT: Integrating Dual Visual Features for Image Captioning

In the first part of the work, we revisit how to extract and utilize visual representations, aiming to build a better and faster model for image captioning. In image captioning, understanding visual information is crucial to correctly describing its content in words. Therefore, extracting good visual representations from the input image is necessary. Current state-of-the-art methods employ region-based features extracted by high-performance object detectors, e.g., Faster R-CNN. However, they have several issues, for example, the lack of contextual information, the risk of incorrect detection, and the high computational cost. The first two could be addressed by

additionally using grid-based features. However, how to extract and integrate these two types of features was uncharted. We propose a transformer-only neural architecture, dubbed GRIT (Grid and Region-based Image captioning Transformer), that can effectively extract and integrate the two visual features to generate better captions for input images. Specifically, GRIT replaces the CNN-based detector employed in previous methods with a DETR-based one, making it computationally faster and end-to-end trainable. We find that the proposed method brings about significant performance improvement, outperforming previous methods in inference accuracy and speed.

第 3 章 **LTMI: Lightweight Transformer for Many Inputs in Visual Dialog**

In the second part of this work, we tackle the visual dialog task, which requires agents to maintain a meaningful conversation with humans about the content of input images by answering questions. Unlike image captioning, the agent must handle multiple inputs, i.e., an image, a question, a dialog history, or even its individual dialog components. Thus, the key to success lies in how to model all the interactions between these inputs effectively and efficiently. We introduce a neural architecture, LTMI (dubbed Light-weight Transformer for Many Inputs), that can efficiently deal with all the interactions between multiple inputs in the visual dialog. It has a block structure similar to the Transformer and employs the same design for attention computation. With a similar setting on visual dialog, a layer built upon the proposed attention block has less than one-tenth of the parameters compared with its counterpart, a natural Transformer extension. It has only a small number of parameters yet has sufficient representational power for the purpose. The experimental results on the VisDial dataset validate the effectiveness of our proposed method.

第 4 章 **LWIT: Improving Performance on Instruction Following Tasks**

In the last part of this work, we study interactive instruction-following tasks. An embodied AI agent is required to perform a sequence of actions to accomplish a complicated task in the interactive environment by following natural language directives. Recent studies have tackled the problem using ALFRED, a well-designed dataset for the task, but have obtained only very low accuracy. To this end, we propose a novel method based on a combination of several new ideas, which surpasses the existing methods by a large margin. One is a two-stage interpretation of the given instructions. The method first chooses and decodes an instruction without visual information, yielding a tentative sequence of object and action predictions. It then integrates this prediction with the visual information to generate the final prediction of an action and an object. It can localize the object of interest accurately from the input image.

Furthermore, the proposed method utilizes multiple egocentric views of the environment and extracts crucial information by applying hierarchical attention conditioned on the selected instruction. It leads to better accuracy in predicting navigation actions. Our proposed method attains an unseen success rate of 8.37%.

論文審査結果の要旨

視覚と言語を人と同じように駆使し，人や周囲の環境と相互に作用し合う機械知能の実現は，人工知能研究における長年の夢である。深層学習の技術的な進歩と，計算量およびデータ量の両面での学習の大規模化によって，その実現は手の届く範囲に近づきつつある。本論文は，視覚を通じて，身の回りの環境やそこで生起した事象を認識・理解しつつ，言語によって人とコミュニケーションをとり，さらに人や環境と相互に作用し合うことのできる機械知能の実現を目指したもので，全編5章からなる。

第1章は序論であり，本研究の目的と背景を述べている。

第2章では，画像理解の基本的課題の1つである画像記述（image captioning）のための手法を提案している。画像記述とは，1 つのシーンの画像を入力に，そのシーンの様子を自然言語で記述する問題である。提案手法は，従来手法と比べて，より少ない計算量でより多くの情報を画像から取り出すことができ，さらに入力から出力までを一括して学習の対象とすることができる。その結果，記述精度および計算量の2つの尺度において，従来方法を上回る性能を達成している。特に，標準的なベンチマークテストであるCOCOでの評価において，公開データのみを学習に利用する条件下での世界最高の精度を達成している。

第3章では，画像を題材に人と人工知能（AI）が対話を行う，視覚的対話（visual dialog）のための手法を提案している。視覚的対話とは， 1つのシーンの画像とそれに関する質問文，および人とAI間での質問応答の履歴という3種類の入力を元に，最新の質問に回答する問題である。提案手法は，少ない計算量でこれら3つの入力間の密な相互作用を計算でき，それによって質問に対する正確な回答を生成する。この手法は，標準的なベンチマークテストであるVisual Dialogでの評価において世界最高精度を達成している。

第4章では，言葉で与えられる指示にしたがって室内での作業をこなすAIエージェントを実現する手法を提案している。視野を広く確保することで，自己位置の推定や操作対象の物体の認識を精度よく行えるようにし，また与えられる作業指示を内部的に2度解釈することで，作業を的確に進めることができる。提案手法は，仮想空間上での家事を題材にした2020年開催の国際コンペティションAlfredにおいて，1位の成績を挙げている。

以上要するに，本論文は，視覚情報処理と言語処理を統合して操るAIのための，従来より高い性能を持つ基盤技術を提案している。この成果は，システム情報科学ならびにコンピュータビジョンの発展に寄与するところが少なくない。

よって，本論文は博士（情報科学）の学位論文として合格と認める。