論 文 内 容 の 要 旨

氏 名 (YANG ZEKUN)	
論文題名	Feature Representations for Visual and Language Task: Towards Deeper Video Understanding (視覚と言語タスクのための特徴表現：より深い映像の理解に向けて)

論文内容の要旨

This research is an attempt to promote deeper understanding of videos. It mainly focuses on feature representations for visual and language elements, and carries the research in three aspects based on two different tasks.

First, it focuses on video question answering task, which aims at answering questions about a video. Currently, most work focuses on image-based question answering, and less attention has been paid to answering questions about videos. However, video question answering presents some unique challenges that are worth studying: It not only requires modelling a sequence of visual features over time, but also needs to reason about associated subtitles. Hence, BERT, a sequential modelling technique based on Transformers, is used to encode the complex semantics from video clips. The proposed model jointly captures the visual and language information of a video scene by encoding not only the subtitles but also a sequence of visual concepts with pre-trained BERT model. In the experiments, the performance of the proposed model is studied by taking different input arrangements, showing outstanding improvements when compared against previous work on two video question answering datasets: TVQA and Pororo.

Then, a deeper study about Transformers is carried. Transformer is a novel architecture that aims at solving sequence-to-sequence tasks while handling long-range dependencies. It relies on self-attention to compute representations of the input and output without using RNNs. In recent years, many kinds of Transformers have been proposed and used extensively. They have been reported to outperform RNNs in several natural language processing tasks. Before carrying on this study, it is not clear which Transformer performs the best in video question answering task, and why it performs the best. With the aim of understanding Transformer better, the visual semantics and the subtitles are encoded with four commonly applied Transformers: BERT, XLNet, RoBERTa, and ALBERT. It is found that the accuracy is different when different Transformers are adopted on the same dataset, and the reason behind such differences is related to the different pre-training settings in different Transformers.

Finally, this research focuses on multi-modal humor prediction task. Humor provokes laughter and provides amusement, and can be induced by signals in the visual, linguistic, and vocal modalities. Previous methods mainly predict humor in the sentence level and with single modality, which often ignore humor caused by, for example, actions. In this work, a dataset for multi-modal humor prediction is proposed based on the famous sitcom the Big Bang Theory. Next, a method is introduced to find temporal segments that involve humor in videos. This method adopts a sliding window to divide the video, and model the visual modality described by pose and facial features, along with the linguistic modality given as subtitles to predict humor. Experimental results show that our method helps improve the performance of humor prediction.

様式 7

論文審査の結果の要旨及び担当者

氏　名　（Zekun Yang）			
	（職）		氏　　名
論文審査担当者	主 査	教授	竹村　治雄
	副 査	教授	八木　康史
	副 査	教授	三浦　典之
	副 査	准教授	中島　悠太

論文審査の結果の要旨

本学位論文では、映像と自然言語に関わるタスクにおいて、昨今広く利用される深層学習を利用したEnd-to-Endに学習可能なニューラルネットによるモデルに対して、人が設計した特徴量を合わせて利用することについて、実験的に論じている。特に映像と自然言語に関わるタスクでは、学習に利用可能なデータセットが比較的小さいことから、特徴量抽出に相当する部分を含む複雑なモデルの学習が必ずしも有効であるとは言えない点に着目し、映像に関する質疑応答（Video QA）タスク、及び映像中の笑いの検出タスクを例として、それぞれのタスクにデザインされた特徴量の有用性を評価している。これは、現在までに例を見ない新しい研究である。本学位論文の主な成果として次の三点が認められる。

第一に、Video QAタスクにおいて、オブジェクト検出結果のラベルを特徴とするモデルの提案が挙げられる。当該タスクは、与えられた映像に関する質問に対する回答を生成することを目的としている。現在までに広く用いられているアプローチでは、End-to-Endで学習可能なモデルで特徴量の抽出に相当する部分をデータから学習する。一方で、本学位論文のアプローチは、オブジェクト検出を行い、その結果として得られた「人」「自動車」などの検出されたオブジェクトのラベルをモデルへの入力とする。映像から直接特徴量を抽出する既存のアプローチに比べて、ラベルを入力とする本アプローチでは得られる情報が少なくなると予想されるが、実際には比較手法に比べて精度が向上するという興味深い知見を得ている。対応する国際会議論文は、2020年3月に公開され、Google Scholarによれば2021年8月の時点で29回引用されている。

第二に、同じVideo QAのタスクにおいて、ラベルをモデルに入力する方法について検討している。ラベルは自然言語で記述されたテキストであると解釈できるものであることから、自然言語処理のアプローチを援用し、事前学習されたモデルを利用する。様々な事前学習済みモデルが発表されているが、その中でも本学位論文ではBERTと呼ばれるモデルで最も高い精度が得られることを実験的に示している。また、その理由について、事前学習で用いられたタスクと、自然言語処理分野においてそれらのモデルを評価したタスク、さらにVideo QAタスクの性質の違いから比較的に考察しており、映像と自然言語に関する幅広い研究に貢献するものと考える。

第三に、映像中の笑いの検出タスクを新たに提案し、このタスクを例として映像が誘発するユーモアの検出に有用な特徴量を検討している。このタスクでは、音声中に観客の笑い声が含まれるSitcomと呼ばれる種類のテレビドラマから笑い声を検出し、その結果を真値として、映像と発話内容を表すテキストから機械学習モデルによってその区間を推定する。関連する既存のタスクでは、発話内容を表すテキストに対して、笑いを誘発するかを判定するが、本学位論文のタスクは、笑いの開始時刻と終了時刻を推定することから、発話を伴わないユーモアも推定できる。このタスクに対して、特に映像をどのように表現するかを検討し、顔の表情、体の動きなどに関する特徴量をデザインし、比較しており、今後の研究に有用な知見を与えると考える。

これらの成果は、主要な学術論文誌、及び国際会議で発表されている。

以上のように、本学位論文は現在広く研究が進められている映像と自然言語の融合分野において、End-to-End学習の有効性を検証するための一助となる重要な成果であり、情報科学の進展に寄与するところが大きい。よって、本論文は博士（情報科学）の学位論文として価値のあるものと認める。