論 文 内 容 の 要 旨

氏 名 （ CHEN TIANWEI ）	
論文題名	Knowledge Transferability in Vision-and-language Models and Its Applications (Vision-and-languageモデルにおける知識転移とその応用)

論文内容の要旨

In this paper, we explore the knowledge transferability in recent vision-and-language models. Recently, transferring knowledge from pre-trained vision-and-language models to handle a new task has become a common idea in solving tasks related to both visual and linguistic data. However, the knowledge transfer strategy of current vision-and-language models is not always effective. On the one hand, some knowledge may not be helpful for knowledge transfer. On the other hand, harmful knowledge, such as social bias, may be involved in the pre-trained models. For both cases, the knowledge transferability in vision-and-language models is limited, as these models may not surely solve new tasks with their knowledge and may provide unfair performance to different social groups. To explore the limitations of the current knowledge transfer strategy, analyze the reason, and further improve the models' performance in solving vision-and-language tasks, I choose the exploration of knowledge transferability in vision-and-language tasks as my PhD topic and make an exhaustive analysis on this topic.

First, we explored the knowledge transferability between 12 vision-and-language tasks to verify that some knowledge in one task may not always be helpful for other tasks. We conduct an exhaustive analysis based on hundreds of cross-experiments on twelve vision-and-language tasks categorized into four groups. We further evaluate four factors that may affect the knowledge transferability, which are the random seeds, the data scale, the training stage, and the dataset similarity.

We then explore how recent large pre-trained models (e.g., VilBERT and CLIP) can be directly applied to challenging tasks. We first propose a task and annotate an evaluation dataset to detect the artwork regions that provoke certain emotions, and this task requires both knowledge of artwork and emotions. We then evaluate eight baseline models on this task, including a weakly-supervised model that we proposed for this task. Furthermore, we explore how the recent deep generative model, Stable Diffusion, can understand emotional stimuli.

At last, we explore how social bias, a kind of harmful knowledge, can affect future models. Recent studies show that deep generative models (e.g., Stable Diffusion) can generate biased images without intention. However, the generated images are increasing on the internet and may become training data for future models.

To explore how deep generative models will affect future models, we conduct simulation experiments of dataset contamination by replacing the original image with the generated images. We then evaluate both the bias changes and model performance changes to evaluate how the future models are affected by the deep generative models. Furthermore, we make an analysis during the experiments and point out some factors that may affect the bias changes of data contamination.

In conclusion, my PhD research focuses on knowledge transferability in vision-and-language tasks in three different situations: (1) the knowledge transferability between twelve vision-and-language tasks;(2) the knowledge transferability from large pre-trained vision-and-language models toward a emotional stimuli detection task;(3) the effect of harmful knowledge related to social bias in deep generative models toward future vision-and-language models.

We hope our work and our insights through the experiments can bring inspiration to the fields of knowledge transferability in vision-and-language tasks.

様式 7

論文審査の結果の要旨及び担当者

氏　名　（　CHEN TIANWEI　）		
	（職）	氏　名
論文審査担当者	主 査　教授	中島　悠太
	副 査　教授	長原　一
	副 査　准教授	大倉　史生

論文審査の結果の要旨

本学位論文は、近年広く研究が進められており、実際のサービスとしての利用も始まっている視覚言語モデルにおいて、他のタスクへの転移学習の可能性（2章）、転移学習の応用（3章）、また社会的バイアスの転移可能性（4章）に関する3つの研究成果に基づく。

第2章では、視覚言語タスクにおける知識の転移可能性について、実験的に示している。視覚言語タスクとは、画像とテキストを入力とする、もしくは画像を入力としてテキストを出力とするタスクであり、現在広く研究が進められている。深層学習モデルでは、あるタスクで学習したモデルを別のタスク（目的タスク）でファインチューニングすることにより、目的タスクのみで学習した場合に比べて性能が高くなる可能性があることが知られている。一方で、どの程度の性能向上が得られるかは実験的に示すことしかできない。既存研究では、画像のみを入力とする様々なタスクに対して転移学習の効果を示したものがあったが、本研究は同様に視覚言語タスクにおいて典型的な画像に関する質疑応答から3種、画像検索から2種、指示表現理解から5種、マルチモーダル検証から2種の12種のタスクの中から、任意の2つの間で転移学習の効果を示した。同じグループのタスクで効果が高い傾向などを確認しており、後続の研究に一つの指針を提供するとともに、タスク感の類似性に関する示唆を与えている。

第3章では、転移学習を利用することにより、利絵画が想起する感情の検出を行った。絵画は見る人によって異なる感情を想起させる可能性があると考えられるが、本研究ではこれを画像内で注目する領域の違いによるものと考え、画像中の異なる領域がどのような感情を想起させるかを予測するモデルを提案している。絵画の分析に利用可能なデータセットは極めて限定的であり、データセットに含まれる絵画の数も少ないことから、この研究では絵画を含む大規模データセットで事前学習された大規模視覚言語モデルを転移学習することにより、この予測を実現している。前述の通り、絵画に関するデータセットは少なく、領域ごとに想起させる感情のラベルが付与されたデータセットは存在しないことから、本研究では既存のデータセットを基に、感情ラベルが付与された評価のためのデータセットを構築し、公開している。本研究は絵画からの感情予測に新しい切り口から取り組むものであると言える。また、データセットの構築は後続の研究の発展に大きく資するものであると考える。

第4章では、生成AIによる画像がインターネット上で多くみられるようになった現状を踏まえ、将来的にこれらの生成AIによる画像を含むデータセットで学習されたモデルを利用して目的タスクに転移学習した場合における社会的バイアスの影響を実験的に示した。生成AIは（社会的）バイアスを持つことが知られており、従って生成AIから得られた画像で学習したモデルもこの影響を受けると予想される。本研究では、画像と自然言語テキストを対応付ける大規模視覚言語モデルについて、その学習のためのデータセットの一部の画像を生成AIによる画像に置き換えたデータセットで学習し、そのモデルを更に画像の説明文生成と検索のタスクで転移学習して影響を調査した。実験の結果、本研究における実験設定では社会的バイアスの影響はほとんど見られなかった。この研究は、生成AIが持つ社会的バイアスが将来的に他のモデルに与える影響を明らかにしている点で非常に有用なものであると考える。

以上のように、本学位論文で得られた研究成果は視覚言語モデルにおける転移学習の基本的な性質を明らかにするとともに、その応用可能性も思索しており、視覚言語モデルの今後の発展に寄与する成果が得られていると考える。よって、本学位論文は博士（情報科学）の学位論文として価値のあるものと認める。