様式3

論 文 内 容 の 要 旨

氏 名 （ 廣田 裕亮 ）	
論文題名	Evaluating and Mitigating Societal Bias in Image Captioning (画像キャプション生成システムにおける社会的バイアスの評価・緩和手法の検討)

論文内容の要旨

Societal biases in artificial intelligence, particularly in computer vision, have become a critical concern due to their potential to perpetuate and amplify harmful stereotypes. Among various computer vision tasks, image captioning exemplifies these challenges due to its interpretive nature. Image captioning models often inherit and amplify biases present in their training data, manifesting as stereotypical associations or incorrect predictions of demographic attributes such as gender and race. These biases not only compromise the fairness and reliability of generated captions but also raise ethical and societal concerns.

This thesis tackles the issue of societal bias in image captioning through three key contributions. First, it introduces a metric for measuring societal bias amplification in image captioning models. This metric quantifies how much the models amplify biases compared to their training data, providing valuable insights into the propagation of bias. Second, it proposes LIBRA, a model-agnostic framework to mitigate bias amplification in image captioning. LIBRA addresses two distinct types of gender bias-context-to-gender bias and gender-to-context bias-ensuring that efforts to reduce one do not inadvertently amplify the other. Finally, the thesis presents a dataset-level bias mitigation framework using text-guided inpainting techniques. This approach creates synthetic datasets with group-independent attribute distributions, reducing spurious correlations and enhancing fairness without sacrificing model performance.

Through these contributions, the thesis advances the understanding of societal biases in image captioning and proposes practical solutions for their quantification and mitigation. The findings pave the way for more equitable and inclusive applications of AI technologies in image captioning and beyond.

様式 7

論文審査の結果の要旨及び担当者

氏　　名　（　　廣田裕亮　　）		
	（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	教授	浦西　友樹

論文審査の結果の要旨

本学位論文は、特に画像と自然言語に関するタスクにおける社会的バイアスに関して、その評価のための指標（2章）から、自然言語テキストのデータ拡張による低減（3章）、さらにデータセットに含まれる画像の生成AIによる拡張を利用した低減（4章）の3つの研究成果に基づく。

第2章では、画像説明文生成タスクにおける社会的バイアスの評価指標を提案している。画像説明文生成における社会的バイアスの問題は、わずかなら既存研究が存在しているものの、その評価手法に関する研究は極めて限られていた。特に、画像の説明文生成で利用可能な評価手法は、画像の人物領域が出力される説明文の人物に関する記述以外の部分に与える影響を評価するものであった。本研究の評価手法は、画像中の人物領域が人物以外の記述に与える影響まで評価できる。この評価指標で既存のバイアス低減手法を評価し、既存手法はバイアスを増大させることを明らかにしている。この結果は当該領域の研究者に大きな影響を与えるものであると考える。

第3章では、同様に画像説明文生成タスクにおける社会的バイアスの低減手法を提案している。先述の通り、当該タスクに対するバイアスの低減手法はわずかながら提案されているが、既存手法は人物に関する記述を正確にすることを目的にしており、画像中の人物領域が人物以外の記述に与える影響を考慮しないものであった。この研究では、画像中の人物領域が人物以外の記述に与える影響、および人物以外の領域が人物の記述に与える影響の両方を考慮したバイアス低減手法を提案している。この手法は、生成された説明文を入力としてバイアスを低減した説明文を出力するモデルを中心とするもので、データセットに含まれる説明文に対してバイアスと類似する様な外乱を加え、これを元の説明文に戻すように学習する。これによって、任意の説明文生成モデルに対して適用可能なバイアス低減モデルの実現している。評価では、前述の2種のバイアスの両方を低減可能であることを実験的に示した。画像の説明文生成は現在広く研究されている上に、ChatGPT等の画像を入力可能な大規模言語モデル（大規模マルチモーダルモデル）で実際にサービスとして提供され始めている。このような状況で、再学習を経ることなく適用可能なバイアス低減手法は極めて有用であろうと考えられる。

第4章でも第3章と同様にバイアス低減手法を提案しているものの、適用可能なタスクが広いアプローチを採用している。バイアス低減における本質的なアプローチは、学習に利用されるデータセットに内包されるバイアス自体を低減することにある。この実現には、例えば事前に与えられたオブジェクトの集合に対して、人物の属性（性別や人種等）に対してその登場回数が等しくなるようにデータセットの再サンプルをするなどが考えられるが、オブジェクトの組み合わせを考えると実現困難となることに加えて、オブジェクトに関するラベル付けが必要であり、オブジェクトの集合に含まれない概念は考慮されないという問題があった。一方で、この研究では与えられた画像中の人物領域に対して（部分的な）画像生成を適用することにより、ぞれぞれの人物属性をもつ画像を生成する（つまり、1枚の画像に対して、例えば男性と女性の画像を生成する）。これにより、理想的には全ての属性に対して画像の任意の要素の登場回数を一致させることができる。評価では、画像に対する説明文生成とオブジェクト検出の二つのタスクで評価を行い、どちらもバイアスの低減ができることを示した。この研究は、画像生成を利用してデータセット自体のバイアス低減を実現するもので、幅広い応用可能性を持つと考えられる。

以上のとおり、本学位論文は深層学習モデルにおける社会的バイアスに関して大きく寄与するものであり、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。