Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Doctoral Consortium Track

Culturally-aware Image Captioning

Youngsik Yun
Department of Artificial Intelligence, Dongguk University
yys3606@dgu.ac.kr

Abstract

The primary research challenge lies in mitigating and measuring geographical and demographic biases in generative models, which is crucial for ensuring fairness in AI applications. Existing models trained on web-crawled datasets like LAION-400M often perpetuate harmful stereotypes and biases, especially concerning minority groups or less-represented regions. To address this, I proposed a framework called CIC (Culturally-aware Image Caption) to generate culturally-aware image captions. This framework leverages visual question answering (VQA) to extract cultural visual elements from images. It prompts both caption prompts and cultural visual elements to generate culturally-aware captions using large language models (LLMs). Human evaluations confirm the effectiveness of our approach in depicting cultural information accurately.

Two key future directions are outlined. First, current image caption evaluation methods are inadequate for assessing culturally-aware captions, necessitating the development of new evaluation metrics leveraging cultural datasets and representations. Second, ethical considerations, particularly concerning stereotypes embedded in existing models, demand consensus and standards development through diverse cultural perspectives. Addressing these challenges is vital for the responsible deployment of AI technologies in diverse real-world contexts.

1 Main Research Directions or Challenges

My main research challenges are mitigating and measuring geographical and demographic biases in generative models. With the increasing usage of AI applications in the real world, developing applications that ensure fairness is becoming increasingly important. It is important for AI applications to avoid exhibiting discriminatory behavior toward specific groups or populations [Mehrabi *et al.*, 2021].

However, existing generative models have been trained on extensive web-crawled datasets like LAION-400M [Schuhmann *et al.*, 2021]. Large-scale datasets collected by crawling the internet without filtering contain malicious stereotypes and biases [Garcia *et al.*, 2023]. Text-to-image generation models such as Stable Diffusion [Rombach *et al.*, 2022] tend to generate images with a Western-centric focus or occasionally produce harmful stereotypes or inaccurate representations of minority countries [Liu *et al.*, 2024].

My research focuses on mitigating and measuring cultural biases within text generation models such as image description and large language models (LLMs). To mitigate cultural bias, some research conducts fine-tuning utilizing language and description data collection from specific countries [Liu *et al.*, 2021]. However, existing approaches rely on substantial data for training, making it challenging for minority ethnic groups or countries with scarce data availability. Other research utilizing prompting in LLMs inherently shows performance degradation for non-Western cultures lacking information [Tao *et al.*, 2023]. Current bias measurement methods are limitations in accurately evaluating cultural bias by only measuring specific factors such as gender and race [Hirota *et al.*, 2023; Zhao *et al.*, 2021].

2 The Specific Contributions

Many Vision-Language Pre-trained models (VLPs) demonstrate remarkable advancements in image captioning [Li *et al.*, 2023; Wang *et al.*, 2022]. However, current state-of-the-art VLPs generate captions that overlook cultural information in various culture-related images. For example, in an image of a Korean person wearing traditional attire, existing models fail to describe the traditional Korean clothing known as ***hanbok***. This issue arises because image-text pair datasets for image captioning often fail to capture and mostly omit cultural elements within the images.

To address this challenge, I have proposed a framework called CIC (Culturally-aware image caption) for generating culturally-aware image captions [Yun and Kim, 2024]. Collecting image-text datasets for various cultural regions is challenging, so I devised a method to obtain cultural visual elements through visual question answering (VQA). I generated questions for five cultural categories defined in the [Halpern, 1955]: architecture, clothing, food & drinks, dance & music, and religion. For the architecture category, a generated question is *'What is the architecture style of the buildings in this image?'* Using all generated questions may

extract cultural visual elements not depicted in the image. Therefore, I utilized VQA only when the caption prompt from BLIP2 [Li *et al.*, 2023] contained words related to cultural categories. Finally, generating culturally-aware captions using LLM model by prompting both caption prompts and cultural visual elements.

Our framework adeptly portrays cultural information (highlighted in red). Quantitative evaluation through word matching-based metrics such as BLEU and Rouge is limited to culturally-aware image captions. Human evaluations were conducted by recruiting participants from each cultural region for quantitative evaluation, and our framework was selected as the best in cultural depiction. Through this approach, I introduce a new task of culturally-aware image captioning and consider it the first step in generating image-text pair datasets for cultural images.

3 The Directions for the Remaining Work

There are two main future directions. First, existing image caption evaluation methods are unsuitable for assessing culturally-aware image captions. The CLIPScore [Hessel *et al.*, 2021], as a method for evaluating image captions, is susceptible to cultural biases since it is also trained on existing image caption datasets. Hence, the evaluation of culturally-aware captions relies on human surveys. With the recent increase in research on collecting cultural datasets [Liu *et al.*, 2021] and learning representations between cultural images and texts [Yin *et al.*, 2023; Ignat *et al.*, 2024], we plan to leverage such studies to develop new evaluation methods.

Secondly, there is a need for consensus on ethical issues such as stereotypes. The framework proposed earlier may generate inappropriate captions as it does not consider the stereotypes inherent in existing models. Ethical standards are needed to address these issues, and social consensus involving diverse cultural perspectives must be achieved.

References

[Garcia *et al.*, 2023] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966, 2023.

[Halpern, 1955] Ben Halpern. The dynamic elements of culture. *Ethics*, 65(4):235–249, 1955.

[Hessel *et al.*, 2021] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[Hirota *et al.*, 2023] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Model-agnostic gender debiased image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15191–15200, 2023.

[Ignat *et al.*, 2024] Oana Ignat, Longju Bai, Joan Nwatu, and Rada Mihalcea. Annotations on a budget: Leveraging geo-data similarity to balance model performance and annotation cost. *arXiv preprint arXiv:2403.07687*, 2024.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[Liu *et al.*, 2021] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*, 2021.

[Liu *et al.*, 2024] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. *arXiv preprint arXiv:2401.08053*, 2024.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Schuhmann *et al.*, 2021] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[Tao *et al.*, 2023] Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*, 2023.

[Wang *et al.*, 2022] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[Yin *et al.*, 2023] Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. Givl: Improving geographical inclusivity of vision-language models with pre-training methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10961, 2023.

[Yun and Kim, 2024] Youngsik Yun and Jihie Kim. Cic: A framework for culturally-aware image captioning. *arXiv preprint arXiv:2402.05374*, 2024.

[Zhao *et al.*, 2021] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14830–14840, 2021.