Image Captioning with Vision-Language Models

Computer Vision (CS6350)

TPA-19

1. Problem Statement

Image captioning is the task of generating natural language descriptions for visual content. It requires understanding objects, their attributes, and relationships within an image, and expressing this understanding in coherent text.

This project involves implementing and fine-tuning a vision-language model for image captioning using publicly available datasets. The model must learn to generate descriptive captions conditioned on image features and, optionally, auxiliary signals such as object tags or prompts. Use **transformer-based architectures** that support image-text modeling.

2. Input

- A set of images from a captioning dataset (e.g., COCO, Flickr30k)
- Optional auxiliary inputs such as object tags, prompts, or bounding boxes

3. Output

- A natural language caption for each input image
- Captions must be fluent, relevant, and descriptive of the image content

4. Dataset

Models must be initialized from publicly available pretrained checkpoints. Training from scratch is not required.

Pretraining (already completed by model authors): Models pretrained on large-scale image-text datasets such as **Conceptual Captions**, **LAION-400M**, or **Visual Genome** may be used. Examples include BLIP, ViLT, OFA, or similar transformer-based architectures.

Fine-tuning and Evaluation: Use **COCO dataset** for training and validation. Alternatively, **Flickr30k** or **NoCaps** may be used for diversity or comparative analysis.

5. Objectives

- Select a pretrained **transformer-based** vision-language model suitable for image captioning
- Fine-tune the selected model on a curated captioning dataset
- Evaluate caption quality using both quantitative metrics and qualitative analysis
- Optionally, incorporate auxiliary inputs such as object tags or prompts to enhance caption generation

6. Evaluation Metrics

- BLEU, CIDEr, METEOR, ROUGE-L
- Include qualitative analysis of caption relevance, fluency, and diversity.

7. Optional Extensions

- **Captioning for Fashion Images**

  Replace general image captioning with a focused task: generating **fine-grained fashion descriptions** for clothing items in fashion photography. This task emphasizes attribute-level detail (e.g., color, texture, style, garment type) and is highly relevant for e-commerce and visual search applications.

  Recommended Dataset: FACAD (FAshion CAptioning Dataset)

- Compare performance across multiple model variants or configurations.
- Visualize attention maps or intermediate representations to interpret model behavior.
- Evaluate zero-shot captioning performance on out-of-domain images.

8. References

- Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." *International conference on machine learning*. PMLR, 2022.
- Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." *International conference on machine learning*. PMLR, 2023.
- Wang, Jianfeng, et al. "Git: A generative image-to-text transformer for vision and language." *arXiv preprint arXiv:2205.14100* (2022).
- Kim, Wonjae, Bokyung Son, and Ildoo Kim. "Vilt: Vision-and-language transformer without convolution or region supervision." *International conference on machine learning*. PMLR, 2021.

- Wang, Peng, et al. "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework." *International conference on machine learning*. PMLR, 2022.
- Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *European conference on computer vision*. Cham: Springer International Publishing, 2014.
- Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." *Proceedings of the IEEE international conference on computer vision*. 2015.
- Sharma, Piyush, et al. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.
- Yang, Xuewen, et al. "Fashion captioning: Towards generating accurate descriptions with semantic rewards." *European conference on computer vision*. Cham: Springer International Publishing, 2020.