COSE474-2024F: Final Project Proposal
"Few-shot Image Captioning with BLIP: Exploring Bootstrapping Language-Image Pre-training for Niche Domains"

June Hee Lee

1. Introduction

Image captioning, the task of automatically generating textual descriptions for images, plays a crucial role in a wide range of applications, from improving accessibility for the visually impaired to enhancing content generation in social media. However, traditional models require large datasets, which is challenging in niche domains like medical imaging, where annotated data is scarce. This project aims to explore how few-shot learning can be applied to image captioning using BLIP, a pre-trained vision-language model. The goal is to generate high-quality captions with minimal training data.

2. Problem Definition & Challenges

Few-shot learning for image captioning is challenging due to the limited availability of labeled data, especially in niche domains. The challenge lies in generating meaningful captions from just a handful of examples while ensuring effective transfer of knowledge from general datasets to specialized fields, such as medical or wildlife imagery. This project investigates how effectively BLIP can generalize in such low-resource environments.

3. Related Works

Traditional image captioning models, such as CNN-LSTM architectures, have relied on large datasets for training. Recent models like OSCAR and UNITER use multimodal transformers for improved performance in vision-language tasks. BLIP is designed for joint vision-language tasks and pre-trained on large-scale datasets, positioning it as a strong candidate for few-shot learning tasks.

4. Datasets

COCO (Subset): A small subset (50-100 images) will simulate few-shot conditions with diverse objects and scenes.
Domain-specific dataset: We will use a niche domain (e.g., medical or wildlife images) to test BLIP's ability to generalize in low-resource settings.

5. Goals to Achieve

Few-shot Image Captioning: Test BLIP's performance in generating captions from minimal data.
Generalization to Niche Domains: Evaluate BLIP's effectiveness in generating captions in specialized fields with few labeled examples.
Performance Comparison: Compare BLIP's few-shot performance on general datasets and niche domains, using metrics such as BLEU, CIDEr, and SPICE.

6. State-of-the-art methods and baselines

State-of-the-art models like OSCAR and UNITER have demonstrated significant success in image captioning tasks by leveraging large-scale datasets and complex architectures. However, these models tend to require extensive data and computational resources, making them less ideal for low-resource environments.
In contrast, BLIP, a pre-trained vision-language model, offers a more efficient approach, particularly in few-shot learning scenarios where only minimal training data is available. This project aims to evaluate whether BLIP can generate competitive image captions with limited labeled data, especially in niche domains, providing insights into BLIP's ability to match or even surpass the performance of more complex models in low-data environments.

7. Schedule

Week 1: Relevant literature review
Week 2: Dataset preparation
Week 3: Implement BLIP, test on COCO subset
Week 4: Evaluate BLIP on niche domain dataset
Week 5: Compare results, analyze performance
Week 6: Fine-tune, if necessary, to improve niche domain caption accuracy
Week 7: Prepare final report