Enhancing Image Captioning with Text Augmentation Techniques

DI725 Term Project Phase 1 Report

Onat Zeybek Kuşkonmaz
Student Number: 2237634

GitHub: https://github.com/Onatparagus/DI725_TermProject
WANDB: https://wandb.ai/onatzk-metu-middle-east-technical-university/DI725_TermProject

***Abstract*—This project proposes a lightweight method to improve caption quality for remote sensing images by applying text augmentation techniques. Specifically, I'll be enhancing the training data of the PaliGemma vision-language model with paraphrasing and back-translation. This approach aims to increase caption diversity, leading to improved generalization and caption quality.**

I. Introduction

Vision-language models (VLMs) such as PaliGemma combine powerful visual encoders and language decoders to perform image captioning. Despite their strength, caption quality can suffer when trained on datasets with redundant or simplistic text descriptions, as is the case with remote sensing datasets like RISC. This proposal explores using simple text augmentation methods to improve model performance.

II. Literature Review

Text augmentation has proven to be an effective tool in various natural language processing tasks. I reviewed the following relevant works that inform my approach:

A. Back-Translation for Data Augmentation

Sennrich et al. (2016) demonstrated that back-translation could be used to synthesize high-quality data for machine translation tasks. This technique has since been generalized to augment datasets for various NLP tasks. In image captioning, it helps by generating natural sentence variants that preserve semantic meaning, which helps diversify the dataset and improve training quality.

B. Paraphrasing with Pretrained Transformers

Raffel et al. (2020) introduced the T5 model, a versatile text-to-text transformer that includes paraphrasing capabilities. Similarly, Pegasus (Zhang et al., 2020) pre-trains on sentence gaps for summarization and paraphrase generation. These models enable the generation of caption variants, promoting greater linguistic richness in training.

C. Image Captioning Metrics

Zhao et al. (2021) detailed the metrics commonly used to evaluate image captioning tasks such as BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. These methods proved to be more reliable than using human ratings, being less subjective and easier to use on a large scale.

D. Opportunities in Captioning Models

While architectural improvements like LoRA (Hu et al., 2022) allow efficient transformer fine-tuning, simple preprocessing-based enhancements remain underexplored in the context of vision-language models. Text augmentation provides a promising, low-cost opportunity to improve training without increasing computational complexity.

These papers indicate that while complex fine-tuning methods dominate the research, there is a research gap in leveraging simple augmentation strategies for improving captioning diversity, especially in the context of remote sensing.

III. Project Proposal

A. Research Question

How can simple text augmentation methods like paraphrasing and back-translation improve the diversity and quality of captions generated by a transformer-based vision-language model on remote sensing datasets?

B. Objective and Significance

I aim to enhance the PaliGemma model's ability to generate diverse and semantically rich captions by increasing variation in training captions using automated augmentation techniques. This task is significant because it offers improvements without modifying model architecture or training procedure complexity, making it accessible and reproducible.

C. Dataset and Preliminary Analysis

The RISC dataset consists of 44,521 satellite images, each paired with 5 human-written captions, totaling over 222,000 entries. Captions are short and often exhibit high overlap. For instance, many captions use repeated phrases such as "A building with a red roof." This redundancy limits model expressiveness. My initial analysis confirms an average caption length of 10 words and significant repetition across captions for a single image.

D. Methodology

I propose the following augmentation strategy:

- **Paraphrasing:** Using a pre-trained T5-base model to generate 1-2 paraphrased versions of each original caption.

- **Back-Translation:** Translate captions to another language (e.g., English $\rightarrow$ French $\rightarrow$ English) using a translation API or HuggingFace pipeline.
- Combine original and augmented captions to increase dataset diversity.
- Fine-tune the PaliGemma model on the augmented dataset.

E. Performance Metrics

I will evaluate the generated captions using standard metrics:

- **BLEU** (Bilingual Evaluation Understudy): Measures n-gram precision between candidate and reference captions.
- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering): Considers both precision and recall, and includes stemming and synonym matching.
- **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation): Evaluates the longest common subsequence between the predicted and reference captions, capturing fluency and word order.
- **CIDEr** (Consensus-based Image Description Evaluation): Specifically designed for image captioning tasks, it calculates the similarity of a generated caption to a set of references using TF-IDF weighted n-grams.

I will compare baseline performance with results from paraphrased and back-translated datasets.

IV. Conclusion

This project proposes an improvement to transformer-based image captioning models. By leveraging well-established text augmentation methods, I aim to enhance caption diversity and quality in the RISC dataset for better model performance. The outcomes may offer a practical alternative to more complex fine-tuning methods and provide insights into improving VLM performance on domain-specific data.

References

[1] R. Sennrich et al., "Improving Neural Machine Translation Models with Monolingual Data," ACL, 2016.

[2] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, 2020.

[3] J. Zhang et al., "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," ICML, 2020.

[4] B. Zhao et al., "A Systematic Survey of Remote Sensing Image Captioning," IEEE Access, 2021.

[5] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR, 2022.