COS 801 Project-Group 8

Bridging the Visual-Linguistic Divide: An Automated Image Captioning System for isiZulu using Deep Visual Attention Models

Muphulusi Dzivhani (u18069682)
Ndaedzo Makgatho (u25739906)
Simamkele Mtsengu (u17042845)

1 Problem

Current image captioning systems perform well in English but underperform in low-resource languages such as isiZulu. Captions often lack fluency, semantic richness, and cultural relevance. There is no existing automated system for isiZulu captions, highlighting a research gap in African languages. This project aims to develop an attention-based deep learning model to generate accurate and culturally appropriate isiZulu image captions, enhancing accessibility for people who cannot read or who require explanations in their home language.

2 Data

- **Dataset:** Flickr8k corpus, originally annotated in English and translated into isiZulu with human annotators.
- **Risks:** Small dataset size and translation errors may affect quality.
- **Mitigation:** Data augmentation, transfer learning, and preprocessing.

3 Baseline & Model Plan

- **Baseline:** CNN (ResNet50/InceptionV3) + LSTM decoder without attention.
- **Proposed Model:** Dual attention (spatial + semantic); explore transformer-based architectures for low-resource languages.
- **Implementation:** TensorFlow/Keras or PyTorch.

4 Metrics

- **Automated:** BLEU, METEOR, ROUGE, CIDEr.
- **Human evaluation:** Fluency, semantic accuracy, and cultural relevance.

5 Risks, Compute Plan, and Milestones

Risks include dataset size limitations, cultural/linguistic translation challenges, and compute demands of training deep models. Training will be managed using CPU/GPU-enabled environments with optimizations like batch normalization and early stopping.

5.1 Risks

- Dataset limitations may hinder fluency and cultural relevance.
- Compute constraints for training deep models with attention.

5.2 Compute Plan

- Use GPU-enabled environments (e.g., Google Colab Pro or university HPC clusters).
- Optimize training with batch normalization and early stopping.

5.3 Milestones

- Translate and annotate Flickr8k captions into isiZulu.
- Implement baseline Bengali model adaptation.
- Train and validate attention-based model.
- Conduct SHAP and ablation studies.
- Finalize evaluation and publish findings.

6 Key Papers (Anchor)

1. **Image Captioning in Bengali**: Demonstrates low-resource language captioning using CNN+RNN with attention.
   item **English to Zulu (Marivate, Sefara, Khoboko)**: Demonstrates that carefully designed prompts and parameter-efficient fine-tuning (PEFT) can significantly improve English-to-Zulu translation performance with large language models, enabling effective NLP for low-resource African languages.
2. **Attention-Based Transformer Models for Image Captioning Across Languages**: Discusses multilingual transformer approaches and attention mechanisms.

GitHub Repository: https://github.com/18069682/isiZulu-image-Captioning