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

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 require explanations in their home language.

Data

- **Dataset:** Flickr8k corpus, originally annotated in English and translated into isiZulu with human annotators to ensure linguistic accuracy.
- **Risks:** Small dataset size and potential translation errors may affect fluency and semantic quality.
- **Mitigation:** Data augmentation, transfer learning from multilingual datasets, and preprocessing to improve robustness.

Baseline & Model Plan

- **Baseline:** CNN (ResNet50/InceptionV3) + LSTM decoder without attention.
- **Proposed Model:** Integrate dual attention (spatial + semantic) to align visual regions with linguistic output; explore transformer-based architectures inspired by low-resource language studies.
- **Implementation:** TensorFlow/Keras or PyTorch frameworks.

Metrics

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

Key Papers (Anchor)

1. **Image Captioning in Bengali** – Introduces a low-resource dataset and CNN+RNN with attention; reference for isiZulu captioning.
2. **Image Caption Generation Using a Dual Attention Mechanism** – Dual attention (spatial + semantic) improves context capture and caption relevance.
3. **Attention-Based Transformer Models for Image Captioning Across Languages** – Multilingual transformers generalize well, highlighting the role of attention in bridging linguistic gaps.