Aysha Siddika Marua
2104010202229

Self-Contribution Report

Project Overview

This project develops a Visual Question Answering (VQA) system that answers questions based on images using computer vision and NLP. It combines image and question features to predict answers from 582 classes, utilizing pre-trained models like ResNet-50, EfficientNet-B3, ViT-B/16, and BERT. Key steps include dataset preprocessing, feature extraction, model training, comparison of vision backbones, and evaluation using classification metrics.

My Contribution

I led Model Development and EfficientNet-B3 tasks, designing the VQA model architecture, configuring training procedures, and overseeing the implementation and integration of the EfficientNet-B3-based model.

Model Implementation

- **Architecture Design**: I implemented the VQA model using EfficientNet-B3 for feature extraction and BERT for text embedding, followed by a fusion layer with linear transformation, ReLU, and dropout. The output is mapped to 582 answer classes via a linear layer.
- **Freezing Model Components**: I froze the EfficientNet-B3 layers (requires_grad=False) to focus training on the fusion layer and classifier, reducing computational load.
- **Input Handling**: I ensured image inputs were pre processed according to EfficientNet-B3's requirements and textual questions were tokenized and embedded using BERT.

Training and Evaluation Procedure

The dataset was split into 80% training, 10% validation, and 10% testing. Using Cross-Entropy Loss and Adam optimizer (learning rate 1e-4), the model trained for 10 epochs with a batch size of 16. Accuracy and loss were tracked. EfficientNet-B3 achieved a 24% validation accuracy by the final epoch, outperforming ResNet-50 and matching ViT-B/16. Confusion matrices and classification reports highlighted misclassification and underrepresented classes.

Also, in the IEEE format report I authored the methodology, feature extraction, model architecture, results and model evaluation.

Group Contribution

The project was a collaborative effort with shared responsibilities. I led EfficientNet-B3 tasks and supported feature engineering and model comparison. The team analyzed model trade-offs and chose ViT-B/16 for the demo. We also had discussions on optimization strategies and training stability.

Conclusion

This project demonstrates a multimodal deep learning approach to VQA using image and language models. My key contribution was in model development and leading EfficientNet-B3. Through teamwork, we benchmarked models and created a functional VQA pipeline, deepening our understanding of model integration, performance evaluation, and real-world AI deployment challenges.