Project Proposal: Mini-CLIP

Contrastive Language-Image Pretraining for Multimodal Learning

ES 667 – Deep Learning

Team Members: Aditya Borate (aditya.borate@iitgn.ac.in), Aryan Solanki (aryan.solanki@iitgn.ac.in),
Nishchay Bhutoria (nishchay.bhutoria@iitgn.ac.in)
Indian Institute of Technology Gandhinagar

1. Project Overview

We propose to implement a miniature version of CLIP (Contrastive Language-Image Pretraining), a multimodal deep learning model that learns joint embeddings of images and text. Our Mini-CLIP will demonstrate the power of contrastive learning by enabling zero-shot image classification and cross-modal retrieval without task-specific training.

2. Motivation

CLIP represents a paradigm shift in computer vision by learning visual concepts from natural language supervision. This project will provide hands-on experience with:

- Contrastive learning and multimodal embeddings
- Transfer learning with vision and language models
- Zero-shot learning capabilities
- Real-world applications of vision-language models

3. Technical Approach

Model Architecture

- **Image Encoder:** Pre-trained ResNet-50 or Vision Transformer (ViT-Small) with projection head
- **Text Encoder:** Pre-trained DistilBERT or BERT-Tiny with projection head
- **Training Objective:** InfoNCE contrastive loss to align image-text embeddings

Dataset

We will use the Flickr8k or MS COCO dataset (subset), containing images paired with descriptive captions, suitable for training within computational constraints.

Training Strategy

Fine-tune pre-trained encoders using contrastive learning on image-caption pairs, optimizing cosine similarity between matched pairs while minimizing similarity with negative samples in each batch.

4. Deliverables

Core Features

1. **Image Retrieval:** Given a text query (e.g., "A child flying a kite on a beach"), retrieve and rank top-K most similar images from the dataset.
2. **Text Retrieval:** Given an input image, retrieve the most relevant text descriptions based on learned embeddings.
3. **Zero-Shot Classification:** Classify images into categories without explicit training on class labels by computing similarity with candidate text prompts.

Evaluation Metrics

Recall@K for retrieval tasks, accuracy for zero-shot classification, and qualitative analysis with visualizations.

5. Expected Outcomes

A functional Mini-CLIP model demonstrating multimodal understanding, with interactive demos for image-text retrieval and zero-shot classification. The project will provide insights into contrastive learning and the effectiveness of vision-language models for downstream tasks.

6. Timeline

Week 1: Data preprocessing, model architecture implementation
Week 2: Training and hyperparameter tuning
Week 3: Feature implementation, evaluation, and visualization
Week 4: Final report and presentation preparation

References

[1] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/pdf/2103.00020

[2] OpenAI, "CLIP: Connecting Text and Images," https://openai.com/index/clip/

[3] OpenAI CLIP GitHub Repository, https://github.com/openai/CLIP

[4] "CLIP: The Most Influential AI Model from OpenAI," Viso.ai, https://viso.ai/deep-learning/clip-machine-learning/