Research Project Report

Pedram Abdzadeh Ziabari

This report outlines the implementation and outcomes of a Python application leveraging the CLIP model to classify images based on textual descriptions. The project involves both zero-shot classification and linear probing on CIFAR-10. A total of 12 zero-shot runs and 4 linear probe runs were conducted, covering all prompt variations and pretrained CLIP ViT-B/32 models. All runs and logs are available in src/main.ipynb and can be executed on Google Colab: Open in Colab. The full implementation is hosted in the GitHub repository: https://github.com/pabdzadeh/monash-admission-task, with a thorough guide in the README. Experiments were performed on Google Colab using a T4 GPU with batch size 1200.

Implementation

The project is implemented in Python with a modular pipeline supporting both zero-shot and linear-probe evaluation. A CLI interface in src/main.py allows specification of batch size, training epochs, linear probe type (Simple or Exact), prompt variations, checkpoint paths, and model selection.

For zero-shot classification, images are embedded using CLIP ViT-B/32's vision encoder, while textual prompts are encoded with CLIP's text encoder. Six variations of CIFAR-10 class prompts were considered: default labels, capitalized forms, descriptive templates (e.g., "a photo of an airplane"), contextual prompts (e.g., "a flying airplane"), and synonyms (e.g., "an aircraft"). The pipeline also allows global addition of prefixes or postfixes to all prompts, providing flexibility for prompt exploration. Cosine similarity between image and text embeddings determines the predicted class, and classification accuracy is computed batch-wise (eval function in src/engine.py).

Linear probing freezes CLIP ViT-B/32's visual stream and applies two strategies. The "Exact" method uses regularized logistic regression with L-BFGS optimization and hyperparameter sweep for λ on the penultimate layer (CLIPWithLinearProbeExact in src/linear_probe_model.py). The "Simple" method trains a single linear layer with Adam optimizer and 50% dropout on top of the penultimate layer of the frozen ViT-B/32 visual embeddings for 10 epochs. Both methods rely on proper preprocessing to match pretrained CLIP expectations. For the "Simple" model, checkpoints can be saved and resumed for reproducibility (linear_probe_train and eval_linear_probe functions in src/engine.py).

Results

Zero-Shot Classification

Zero-shot performance is strongly influenced by the choice of textual prompts. Using two pretrained CLIP models, laion2b_s34b_b79k and datacomp_xl_s13b_b90k, descriptive prompts outperformed raw labels especially in the case of pretraining on laion2b_s34b_b79k, with accuracies increasing by over 20% in some cases. Also, overly specific or contextual prompts sometimes decreased performance, likely due to misalignment with the model's pretraining distribution.

Table 1- Accuracy of Zero-Shot CLIP models pretrained on laion2b_s34b_b79k, and datacomp_xl_s13b_b90k for different prompt types
Prompt design clearly influences classification accuracy, with template-based descriptions outperforming raw class names.

Class Name Prompts	CLIP ViT-B/32 (laion2b)	CLIP ViT-B/32 (datacomp_xl_s13b_b90k)
Default Class Names	71.7%	95.2%
Type 1 (an airplane, an automobile, …)	92.0%	95.3%
Type 2 (e.g. a photo of an airplane, …)	93.8%	95.3%
Type 3 (e.g. a flying airplane)	91.0%	91.7%
Type 4 (e.g. a passenger airplane)	89.1%	90.9%
Type 5 (synonym e.g. an aircraft)	90.2%	92.8%

Linear Probe

Linear probes achieved high accuracy for both Exact and Simple strategies, confirming the separability of CLIP embeddings for CIFAR-10. Pretrained model choice had minimal impact.

Table 2- Accuracy of Linear Probe Train CLIP models pretrained on laion2b_s34b_b79k, and datacomp_xl_s13b_b90k for Simple and Exact implenetation of Linear Probe

Probe Type	CLIP ViT-B/32 (laion2b)	CLIP ViT-B/32 (datacomp_xl_s13b_b90k)
Exact	96.8%	97.8%
Simple	96.9%	97.8%

Challenges

The implementation of the Exact linear probe required careful attention to hyperparameter tuning and optimization stability. Preprocessing CIFAR-10 images to match CLIP's expected input format involved precise normalization and resizing. Zero-shot classification revealed that minor changes in textual prompts can noticeably affect accuracy. Managing multiple prompt variations and repeated runs for reproducibility required careful organization and computational resources.

Potential Improvements

Prompt engineering could be further refined by exploring additional variations of textual prompts to optimize zero-shot transfer. Further experiments with other pretrained CLIP ViT-B/32 models could provide insights into differences in performance and prompt sensitivity across model variants. While the primary evaluation metric remains accuracy, future work could optionally consider complementary metrics such as precision, recall, or F1-score to provide additional perspectives on performance.

Conclusion

This project demonstrates that CLIP provides effective visual embeddings for CIFAR-10. Zero-shot performance is sensitive to prompt design, whereas linear probes achieve high accuracy regardless of strategy. The results highlight CLIP's potential for rapid adaptation to new classification tasks and underscore the importance of prompt design in multimodal applications.