COSE474-2024F: Final Project
"Improving Fine-Grained Image Classification with Prompt Engineering Using CLIP"

Shin Minseo 2022320338

1. Introduction

The emergence of foundation models like CLIP (Contrastive Language-Image Pre-training) has revolutionized the way images and text are processed in a shared embedding space. It has a good performance in zero-shot classification for general tasks, but its effectiveness diminishes when applied to fine-grained scenarios where the classification is focused on distinguishing between subcategories that are visually alike, for instance, classifying different bird species.

In the context of fine-grained image classification, small inter-class differences like changes in color, size, or shape, can be difficult to capture using generic prompts. Given this context, this project aims to explore the question: How do different prompt engineering strategies affect the performance of CLIP on fine-grained classification? Different approaches of prompt refinement will be taken until CLIP's alignment with subtle visual distinctions is enhanced.

Approach involves simple, descriptive, and context-aware prompts. We will observe the effects of prompt design complexity and evaluating the role of contextual information in improving classification accuracy . As for the contributions, a comprehensive analysis of different prompt design strategies for fine-grained classification using CLIP is presented, along with experimental insights.

2. Related Works

Fine-Grained Visual Classification Fine-Grained Visual Classification (FGVC) aims to categorize distinct subcategories within a bigger category that is visually similar. Traditional approaches relied on domain-specific feature engineering which is a technique that uses knowledge of a specific field or problem to create features. This often involved hand-crafted features or leveraging multi-scale architectures. Fine-Grained Visual Classification Feature-Encoding method emphasizes the local and global features for capturing subtle differences in visual attributes, which required extensive manual design and were not easily generalizable. The introduction of deep learning to fine-grained classification proposed attention-based architectures to focus on distinguishing features. These methods often require task-specific datasets and extensive labeled data, limiting their scalability.

Prompt Engineering for Vision-Language Models Prompt engineering rose as a crucial tool for adapting pre-trained models to specific tasks. In vision-language models (VLM), carefully crafted prompts align the textual input with the visual content, improving task-specific performance. Works like prompt tuning for VLMs introduced methods for optimizing prompts to improve zero-shot classification.

Descriptive Prompts: One approach to improve performance is to make the prompts more descriptive. Proposed by Zhou et al., CoOp(Context Optimization)is where a learnable context is attached to the class name. For example, instead of basic "A photo of a [class]", the CoOp method might generate a prompt like "A large, colorful [class] with sharp claws." Their results show that learned prompts can improve classification for certain datasets.

Contextual Prompts: Another strategy is to introduce contextual information about the object. For example, the environment (e.g. forest) where a type of bird is known to reside in will be added along with the prompt. This approach proposed by Gao et al. were adapted to reflect image-specific context, such as scenes or habitats. This method was shown successful for large-object categories like cars in streets.

3. Methods

3.1 Overview of the Approach The goal of the framework is to improve CLIP's performance in fine-grained image classification tasks by leveraging prompt engineering as a lightweight yet effective adaption method. CLIP's generic prompts, such as "A photo of a [class]" fail to capture subtle distinctions necessary for fine-grained categories like bird species or car models. To address this limitation, an in-depth investigation into prompt engineering techniques is conducted. Intuitively, adding richer descriptions or contextual details to prompts should provide CLIP with additional semantic information, leading to better alignment between text and image embeddings.

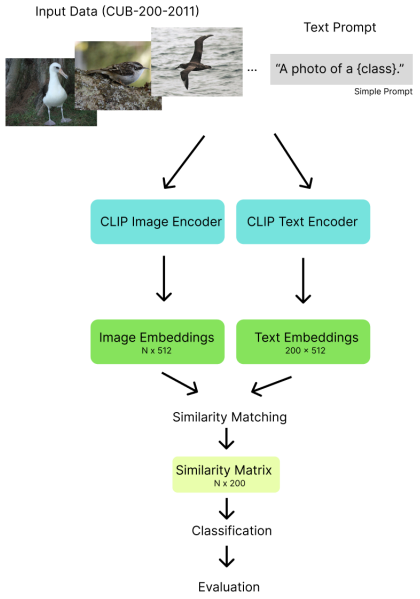

Figure 1. Experimental pipeline for fine-grained classification. The input data is processed by CLIP encoders to generate embeddings, followed by similarity matching, classification, and evaluation.

3.2 Experimental Pipeline

As illustrated in figure 1, the image for input data is from CUB-200-2011(Caltech-UCSD Birds) dataset. This dataset contains 200 categories of bird species and 11,788 images. The different bird species requires fine-grained distinctions, such as variations in plumage patterns or beak shapes. Three set of prompt templates will be used as input to define the text embeddings for classification.

3.2.1 Embedding Generation For each of the 200 bird classes, we generate three types of prompts. Each prompt is passed through the pre-trained CLIP text encoder to generate a 512-dimensional text embedding. Simple Prompts: Minimal text like "A [class] bird." Serve as the baseline. Descriptive Prompts: Attribute-rich prompts like "A large [class] bird with colorful plumage." Captures class-specific features. Context-Aware Prompts: Additional contextual details like "A [class] bird found in forest habitats." aims to introduce environmental context. For the adaptive prompt aggregation strategy, we generate multiple prompts for each class and compute the mean embedding across all prompts. The input images are passed through the CLIP image encoder, which outputs a 512-dimensional embedding for each image.

3.2.2 Similarity Calculation To classify a bird image, the similarity between its embedding and the 200 class text embeddings is calculated using cosine similarity.

Let $\mathbf{v}$ be the image embedding and $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{200}$ be the 200 text embeddings. The similarity between the image embedding and a text embedding is calculated as:

$$\text{similarity}(\mathbf{v}, \mathbf{t}) = \frac{\mathbf{v} \cdot \mathbf{t}}{\|\mathbf{v}\| \, \|\mathbf{t}\|} \tag{1}$$

The predicted label y^* for the image is given by:

$$y^* = \arg \max_{i \in \{1,\ldots,200\}} \text{similarity}(\mathbf{v}, \mathbf{t}_i) \tag{2}$$

3.2.3 Evaluation Metrics The following metrics are used to evaluate the effectiveness of each strategy: Accuracy: proportion of correctly classified images Confusion Matrix: Visual representation of misclassifications, showing which classes are commonly confused Qualitative Analysis: Misclassified images to identify why some strategies fail

In this paper, we hypothesize that both descriptive and context-aware prompts will outperform simple baselines as they provide more detailed semantic information about the classes.

4. Experiments

A series of experiments and analyses to answer the following key research questions:

Q1. Do the proposed prompt engineering strategies improve classification accuracy compared to the baseline prompt? Q2. Which prompt strategy yields the best performance, and how doe sit generalize to unseen test data? Q3. What are the key success factors and failure points of each prompt strategy, and how do they affect misclassification patters?

4.1 Dataset In this section, we evaluate the effectiveness of proposed prompt engineering strategies for fine-grained image classification using the CUB-200-2011(Caltech-UCSD Birds) dataset. This dataset contains 200 categories of bird species and 11,788 images. The different bird species requires fine-grained distinctions, such as variations in plumage patterns or beak shapes. The dataset is first split into train(20%) and test(80%) sets to ensure fair evaluation. Images are resized to 224x224 pixels to match the requirements of CLIP CiT-B/32 image encoder.

4.2 Computing Resources The experiments utilized pre-trained CLIP ViT-B/32 model, and it was conducted on Google Colab Pro's GPU T4 environment. PyTorch 1.13.1 and Transformers 4.27.1 frameworks were used.

4.3 Quantitative Results

Table 1 summarizes the classification accuracy for each prompt engineering strategy. We can observe that all three approaches result in similar accuracy scores, and the performance is low overall. Opposite from what was hypothesized,

Table 1. Zero-shot classification accuracy for different prompt strategies

Prompt Strategy	Accuracy (%)
Simple Prompts	35.89
Descriptive Prompts	35.47
Context-Aware Prompts	35.42

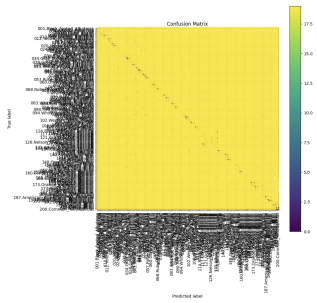

Figure 2. Confusion Matrix

simple prompts outperform the other two strategies. Descriptive and context-aware prompts perform slightly worse.

4.4 Qualitative Results

We visualized a confusion matrix to identify patterns in misclassifications (figure 2). It summarizes the predictions of the model across 200 bird classes. It reveals a strong diagonal dominance, indicating that the model correctly classifies a significant proportion of samples for most classes. There is a high confusion between visually similar classes, which is further evaluated by figure 3 examples of misclassified images.

We can see that misclassifications often occur between species with similar plumage patterns or physical features. For example, Great Grey Shrike (true) and Loggerhead Strike (predicted) shows similar black and white plumage and hooked beaks. In addition, background with similar contexts may contribute to error. For the example of Floria Jay (true) and Western Wood Pewee (predicted), they both have backgrounds of forest or open fields.

4.4 Interpretation

Challenge comes from the point that this particular dataset (CUB-200-2011) requires distinguishing birds of 200 species, many of which exhibit subtle visual differences, with just prompt engineering. Result shows that CLIP struggles to differentiate between fine-grained classes without additional fine-tuning. CLIP's embedding focuses on global features like overall shape and color, rather than localized details such as beak shapes. Also, prompts like "in forest habitats" doesn't work well because CLIP relies on visual alignment and habitats context is not visible in the images.

Simple prompts achieve the highest accuracy, suggesting that concise and direct prompts create better alignment between text and image embeddings in zero-shot classification. Also, adding background context is likely to mislead the task because many images focus solely on the bird not the environment. Without fine-tuning on the dataset, CLIP fails to capture fine-grained distinctions of the birds.

Figure 3. Misclassified Images

5. Further Directions

Multi-prompt ensemble could help address the limitations of single-prompt strategies by averaging embeddings generated from diverse, yet complementary textual descriptions for each class. This can create more robust alignment between text and image embeddings.

Approaches other than prompt engineering can be taken to greatly improve fine-grained classification task of this dataset. Incorporating local attention mechanisms like Grad-CAM could enhance the model by focusing on key region of the image like the bird's features rather than relying on global features. After then, we can apply descriptive approach of prompt engineering.

6. References

A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML), 2021.

K. Zhou, J. Yang, C. Wang, et al. Learning to prompt for vision-language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

L. Gao, T. Zhu, and W. Zhang. Prompt learning for vision-language models. arXiv preprint arXiv:2109.01134, 2021.