Final Project Proposal Update

Topic: Improving Fine-Grained Image Classification with Prompt Engineering Using CLIP

1. Introduction

Fine-Grained Image Classification is a task focused on distinguishing between subcategories that are visually alike, for instance, classifying different car models or dog species. Pre-trained models like CLIP have shown good performance in zero-shot classification for general tasks, but their effectiveness diminishes when applied to fine-grained scenarios. This project aims to explore how prompt engineering - structuring domain-specific textual inputs - can improve CLIP's ability to classify fine-grained categories. Different approaches of prompt refinement will be taken until CLIP's alignment with subtle visual distinctions is enhanced.

2. Problem Definition and challenges

In zero-shot classification tasks, CLIP performs well with general categories like comparing cats and dogs, but it struggles with fine-grained categories (e.g. specific species of animals or car models). The limitation could arise from prompts being too generic to capture specific and subtle distinctions. Fine-grained categories like bird species in the CUB-200-2011 dataset, require prompts that highlight subtle differences in visual and contextual features.

This project considers to confront three main challenges. First, crafting prompts that capture domain-specific features without overfitting. Second, Balancing prompt design simplicity with effectiveness. Third, analyzing failure cases to iteratively refine prompts for improved accuracy.

3. Related Works

Fine-Grained Visual Classification (FGVC)

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 attention0based architectures to focus on distinguishing features. These methods often require task-specific datasets and extensive labeled data, limiting their scalability.

CLIP and Vision-Language Alignment

CLIP (Contrastive Language-Image Pretraining)

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.

In order to apply this method for fine-grained tasks, Context-Aware Prompt Tuning (CPT) explores dynamic prompt generation, demonstrating the effectiveness of prompts that takes in context-specific information, such as environmental or object-specific cues. These works highlight the potential of prompt engineering to adapt CLIP for fine-grained classification.

4. Datasets

In this project, two datasets will be used. CUB-200-2011 (Caltech-UCSD Birds) 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. Stanford Cars dataset contains 16,185 images of 196 classes of car types. This dataset provides a domain-specific challenge for recognizing subtle differences in the car design.

5. SOTA methods and baselines