Latent Class Inference for Calibrated Vision-Language Models in Few-Shot OOD Learning

Assignee Research

June 12, 2026

Abstract

In the realm of few-shot learning, foundation models like CLIP have proven effective but exhibit limitations in cross-domain robustness especially in few-shot settings. Recent works add text as an extra modality to enhance the performance of these models. Most of these approaches treat text as an auxiliary modality without fully exploring its potential to elucidate the underlying class visual features distribution. In this paper, we present a novel approach that leverages text-derived statistics to predict the mean and covariance of the visual feature distribution for each class. This predicti

1 Introduction

This paper examines: Inferring Latent Class Statistics from Text for Robust Visual Few-Shot Learning. Research question: To what extent does inferring latent class distributions from text descriptions improve the calibration and accuracy of vision-language models in out-of-distribution few-shot learning scenarios relative to prompt-based baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

10 papers retrieved. 12 claims extracted; 9 independently verified. Quality review score: 7.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The experiments use two base datasets: ImageNet and iNaturalist.	✓	0.16
iNaturalist is a hierarchical dataset with fine-grained classes.	×	0.11
The covariance matrix plays a crucial role in iNaturalist due to its fine-grained classes.	✓	0.17
The test datasets include Caltech, EuroSAT, Food, Flowers, SUN397, DTD, Pets, Cars, and UCF101.	✓	0.18
Visual and text features are extracted using the pre-trained CLIP ResNet50 trained on LAION400M.	✓	0.16
The method aims to predict the mean and covariance of a class distribution in the feature space from text.	✓	0.18
The learning phase involves predicting the mean and covariance of visual features using textual descriptions.	✓	0.22
The visual backbone (fv) and text encoder (ft) are pre-trained models used to extract image and text features, respectiv	✓	0.20
Text contexts such as 'a photo of a {class}' or GPT3-generated visual descriptions are used for text features.	×	0.14
The target is to estimate the mean (μi) and covariance (Σi) of the visual features for each class ci.	✓	0.19
A diagonal covariance matrix is inferred due to the high-dimensionality of the features.	×	0.14
Two mapping networks, gμ (s, $\theta$$\mu$) and g$\Sigma$ (s, $\theta$$\Sigma$), are employed for predicting the mean and covariance.	✓	0.26

References

- http://arxiv.org/abs/2311.14544v1
- http://arxiv.org/abs/2405.16091v2
- http://arxiv.org/abs/2502.07409v5