STRUCTURE-AWARE GRADIENT REGULATION FOR GENERALIZABLE VISION–LANGUAGE MODELS

Author's: Ashil P[1], Dr. K. Sujith[2]

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in cross-modal understanding, enabling tasks such as image-text retrieval, zero-shot classification, visual question answering, and multimodal reasoning. Despite their strong performance on in-distribution benchmarks, VLMs often struggle to generalize under domain shifts where superficial correlations, spurious visual cues, or dataset biases dominate training signals. Such shortcut learning results in unstable gradient behavior, overfitting to domain-specific patterns, and poor transferability across unseen environments.This project proposes a novel framework titled **Structure-Induced Gradient Regulation (SIGR)** to enhance the generalization capacity of Vision-Language Models. The core idea is to incorporate semantic structure information into the training process and regulate the gradient dynamics across related concepts. By modeling conceptual relationships as a structured graph and applying Laplacian-based smoothness constraints on per-class gradient magnitudes, the framework ensures that learning signals are distributed coherently across semantically related categories. This reduces gradient spikes caused by spurious correlations and promotes stable, structure-aware representation learning.A synthetic multi-domain dataset is constructed to simulate domain shifts with controlled spurious cues. A lightweight CLIP-like architecture consisting of an image encoder, text encoder, and shared embedding space is implemented. The model is trained using contrastive learning with an auxiliary classification objective, and the SIGR regularizer is integrated into the optimization process. Experimental evaluation demonstrates improved robustness in target-domain retrieval and classification tasks compared to baseline models without gradient regulation., The proposed approach offers a principled mechanism for improving cross-domain generalization in multimodal systems and can be extended to real-world applications involving knowledge graphs, ontologies, and large-scale pre-trained models.