UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE ED
ELETTRICA E MATEMATICA APPLICATA

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

ELABORATO FINALE

Adapting Vision Language Models via parameter-efficient fine-tuning for Multitask Classification of Age, Gender, and Emotion

Relatore

Prof. Mario Vento

Prof. Antonio Greco

Candidato

Antonio Sessa

Matr. 0622702305

Abstract

Description of the problem addressed

Recognizing facial attributes such as age, gender, and emotion is an inherently difficult computer vision task due to high intra-class variability and challenging real-world conditions. In this field, large-scale Vision Language Models (VLM) can offer a powerful, generalized visual representations from large scale image-text pre-training, but their direct application to this specialized domain can be inefficient, due to the unnecessary computational overhead of their full architectures, which are not optimized for a pure classification objective. Therefore the central challenge of this work is to develop an efficient and effective adaptation framework to leverage these powerful, pre-trained vision encoders for a unified, multi-task classification objective.

Thesis framework in the contemporary technical scenario

In the current landscape of computer vision, Vision Language Models represent a major shift in the field, demonstrating exceptional zero-shot capabilities through massive-scale pre-training on billions of image-text pairs, in fact, state-of-the-art models like CLIP, SigLIP, and the recent Perception Encoders have shown that joint vision-language training yields powerful, transferable visual representations. Concurrently, the field has witnessed the rise of Parameter-Efficient Fine-Tuning techniques, particularly LoRA and its variants, which enable adaptation of this large pre-trained models with minimal trainable parameters and computational cost. This thesis positions itself at the intersection of these two trends, proposing a framework that leverages the rich visual representations learned by state-of-the-art VLMs while employing PEFT methodologies to enable efficient, specialized adaptation for multi-task facial analysis to address both the performance and efficiency demands of real-world scenarios.

Personal contribution of the candidate to the solution of the problem described

This thesis contributes a comprehensive framework for the efficient multi-task adaptation of a VLM's vision encoder, encompassing its design, implementation, and rigorous evaluation, with a systematic comparison of multiple adaptation techniques such as linear probing, attention probing, partial fine-tuning, and Parameter-Efficient Fine-Tuning (PEFT). The final result of this work is a unified multi-task model that achieves strong accuracy and generalization across all three facial analysis tasks, while also being computationally efficient by discarding the VLM's text encoder to halve inference GFLOPs.

Description of the experimental contents of the work

The experimental work provides a rigorous empirical evaluation of the proposed framework. The PE-Core-L vision encoder is adapted for the three tasks using a composite dataset (FairFace, Lagenda, RAF-DB, CelebA-HQ), with generalization tested on unseen benchmarks (UTKFace, VggFace2). A comprehensive comparison is conducted between multiple adaptation strategies: a zero-shot baseline, linear and attention probing, partial fine-tuning of the final blocks, and PEFT (LoRA+/DoRA). These methods are evaluated in both single-task and multi-task settings, with the latter employing uncertainty-weighting to balance the loss functions, masked labeling to handle partially annotated data and balanced sampling to address the unbalanced datasets. Performance is measured by accuracy, balanced accuracy and also by computational efficiency, using GFLOPs and the number of trainable parameters to quantify the final model's efficiency.