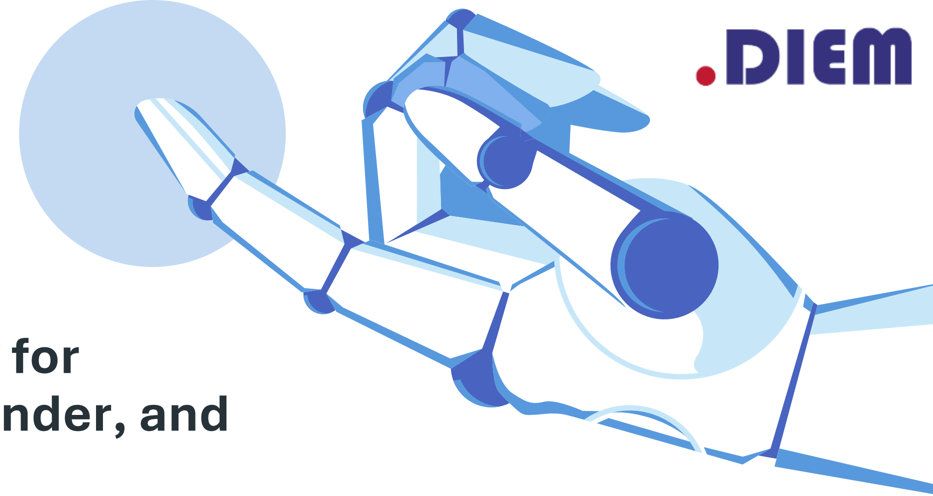


Adapting Vision Language Models via parameter-efficient fine-tuning for Multitask Classification of Age, Gender, and Emotion

Relatori:
Prof. Mario Vento
Prof. Antonio Greco

Canditato: *Antonio Sessa*
Matricola: *0622702311*

UNIVERSITÀ DEGLI STUDI DI SALERNO

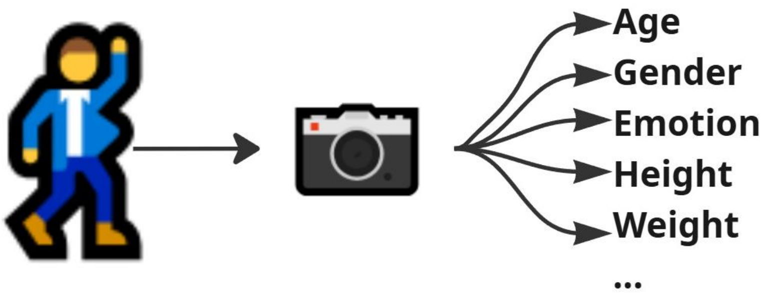
Index

Soft Biometric Recogntion, what and why?

.DIEM

Definition

Soft biometrics are non-unique human attributes that can be indirectly collected from images.



Applications

Social Robotics		*A robot estimates a user is a child and automatically switches to a simpler speech and more playful voice*
Marketing & Commerce		*A digital sign detects a shopper's likely age and gender to show a targeted ad, like for a video game or a new perfume*
Security & Access		*A website uses facial age estimation to automatically block a user who appears underage from accessing mature content*
Healthcare & Wellness		*A wellness app monitors a user's vocal tone or facial expression through their phone to detect signs of stress or fatigue*

Our domain, facial attributes

Facial Emotion Recognition

Labels (7 classes):
Happy, Surprise, Disgust, Angry, Fear, Sad, Neutral

Challanges:

- Class imbalances
- Small datasets
- Low annotator agreements

Age group Classification

Labels (9 classes):
0-2, 3-9, 10-19, 30-39, 40-49, 50-59, 60-69, 70+

Challanges:

- Class imbalances
- High class intra-variance

Gender Recognition

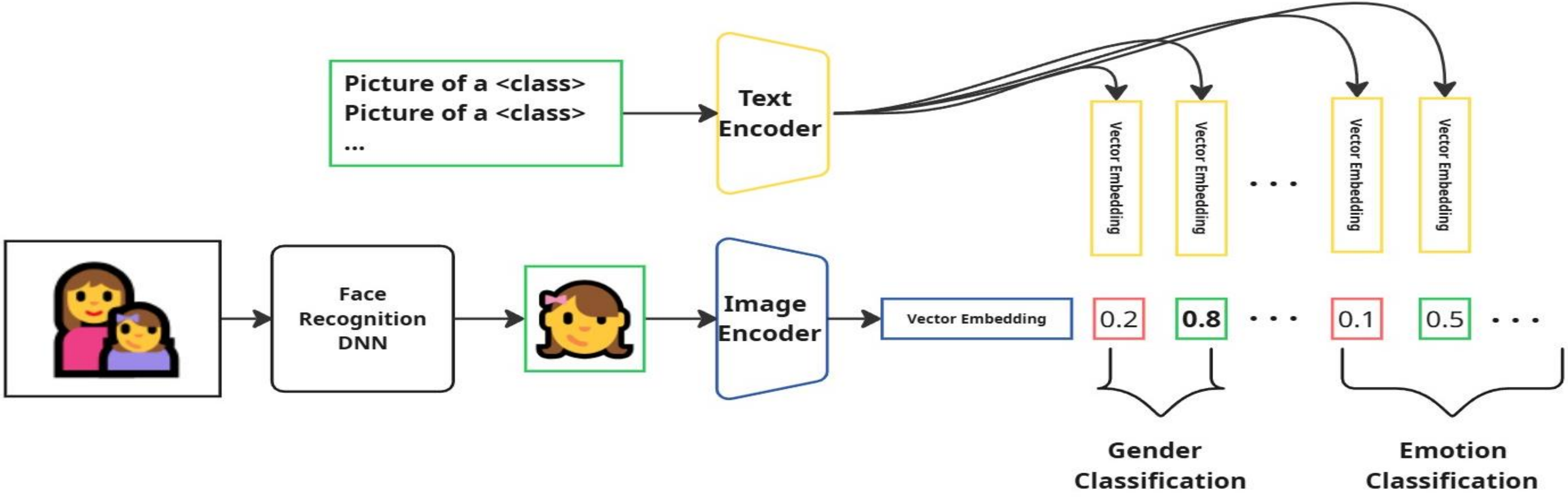
Labels (2 classes):
Male and Female

Challanges:

- Different age groups and ethicities

Our Approach, Vision Language Models

Vision language model (VLM) are large neural networks, trained on billion of image-text pairs, that can be use in a **zero-shot manner**.
We can create a soft-biometric recognition system using a **VLM** (we choose Perception Encoders).



Hard-Prompting, not good enough:

.DIEM

Problems of hard-prompting

- Poor accuracy
- High memory footprint
- High latency

Since **visual understanding stems from the vision encoder**, we can **omit the text encoder** and use the image encoder as a foundation vision model, doing so we **halve the inference time and memory footprint**.

Age*	Gender	Emotion	Global
47.36%	96.67%	66.57%	69.61%

*Average calculation excludes the VggFace2 dataset as its age-labels data are synthetically obtained.

Component	Parameters
Text Encoder	353,986,561
Visual Encoder	318,212,106
Total Parameters	671,137,793
GFLOPs	699.76

Table 4.3: Number of parameters used by the zero-shot baseline during inference

Datasets used to adapt the pre-trained ViT

Training Set

FairFace	~ 97k	Gender & Age
Lagenda	~ 67k	Gender & Age
RAF-DB	~ 17k	Emotion & Gender
CelebaHQ	~ 17k	Gender

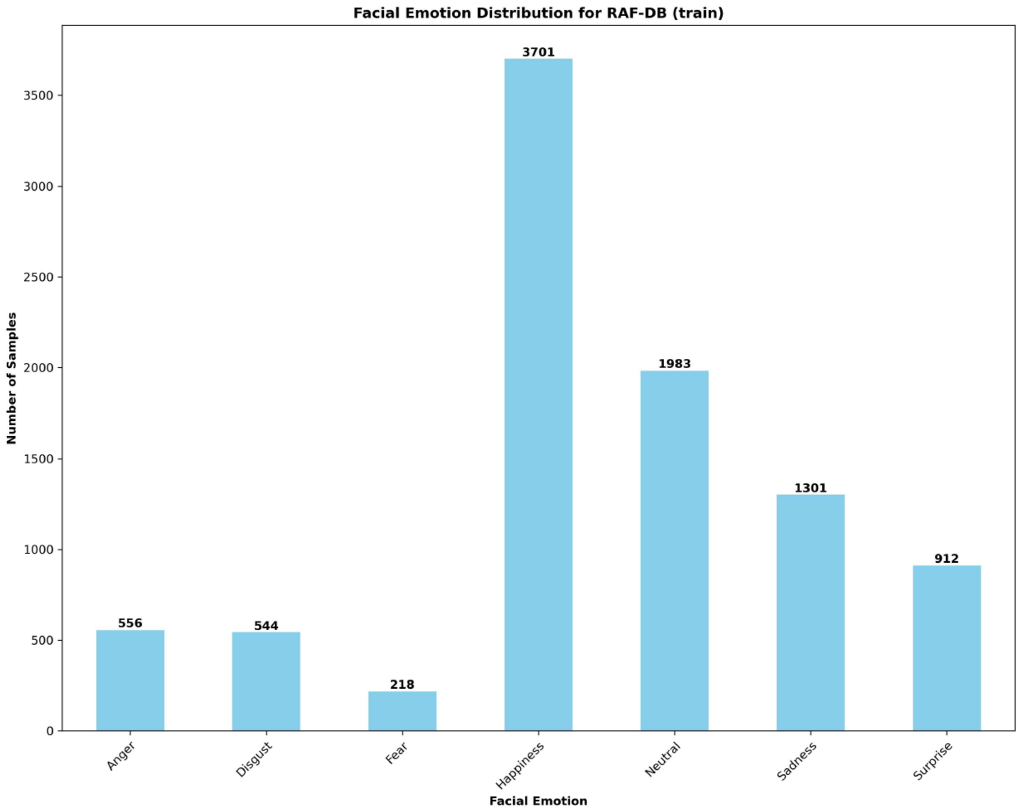
Test Set

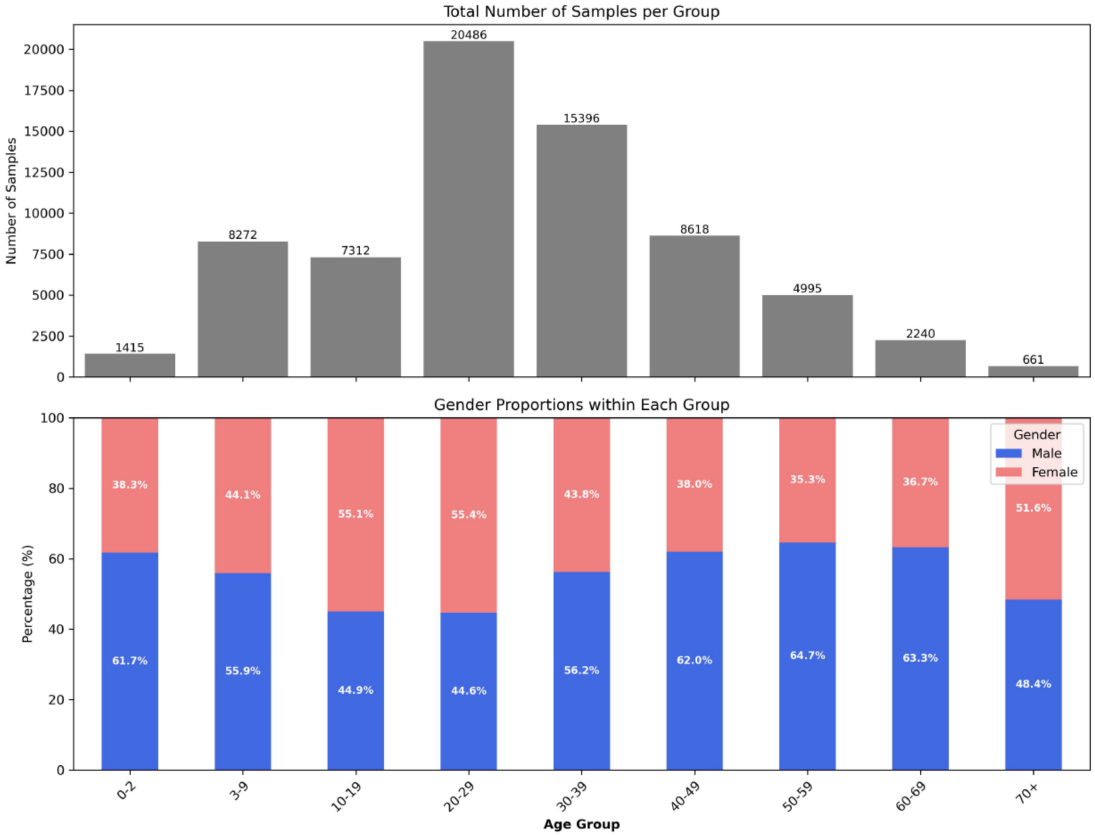
VggFace2	~ 170k	Gender & Age
UTKFace	~ 24k	Gender & Age

FairFace & RAF-DB; Test-splits

. Low amount of celebrity data
. Diverse ethinicities rappresented
. High annotator agreements
. Cross-dataset generalization

. Intra task-class imbalance
. Task imbalance
. Missing labels


Facial Emotion Distribution for RAF-DB (train)
Number of Samples
0
500
1000
1500
2000
2500
3000
3500
556
544
218
3701
1983
1301
912
Anger
Disgust
Fear
Happiness
Neutral
Sadness
Surprise
Facial Emotion



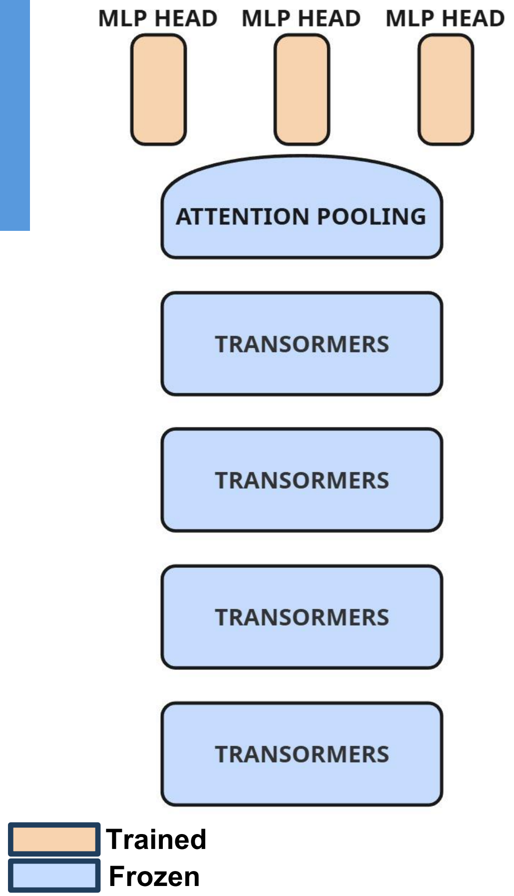
Total Number of Samples per Group
Number of Samples
20000
17500
15000
12500
10000
7500
5000
2500
0
1415
8272
7312
20486
15396
8618
4995
2240
661
Gender Proportions within Each Group
Percentage (%)
100
80
60
40
20
0
Gender
Male
Female
38.3%
44.1%
55.1%
55.4%
43.8%
38.0%
35.3%
36.7%
51.6%
61.7%
55.9%
44.9%
44.6%
56.2%
62.0%
64.7%
63.3%
48.4%
0-2
3-9
10-19
20-29
30-39
40-49
50-59
60-69
70+
Age Group

Multi-task learning, problems and solutions

Multi-Task Problems

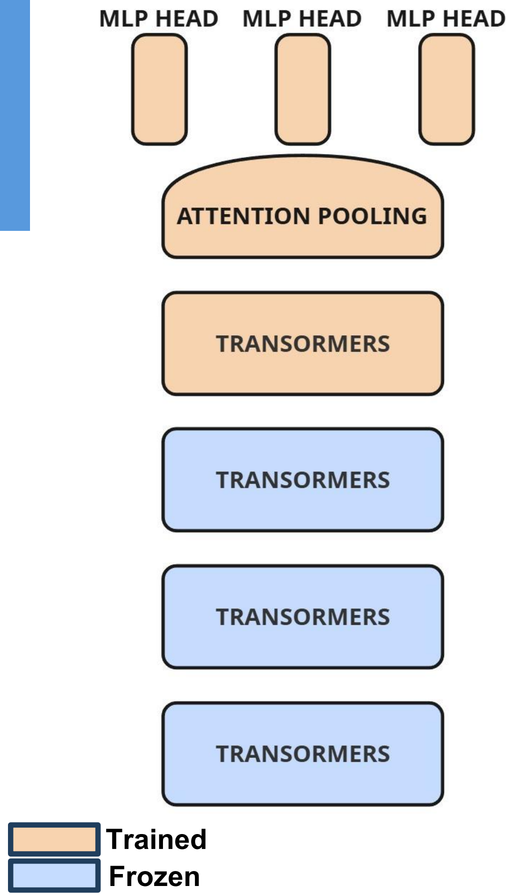
Solutions

Missing Labels	**Masked Loss**	In a batch, average the weight per task only per labelled samples
Class Imbalance	**Weighted Sampling**	Oversampling, utilizing the inverse class frequency as a scoring mechanism.
Task imbalance	**Task Augmentation**	Duplicate emotion samples until they become a third of the datasets
Unbalanced losses	**Uncertainty Weighting**	$\mathcal{L}(W, \sigma_a, \sigma_g, \sigma_e) = \sum_{k \in \{a,g,e\}} \left(\frac{1}{\sigma_k^2} \mathcal{L}_k(W) + \log(\sigma_k^2) \right)$



Linear Probing: attaching a MLP classifier per task to the output of the model, to harness the out-of-the box features of the vision transformers. As there can be no benefit to train the classification head simultaneously, as there is no parameter sharing, they are **trained sequentially** on fully labelled datasets split.

Training: 0.33% of parameters



Partial fine-tune: unfreezing the attention pooling layer and last four transformers block of the ViT, using a differential LR strategy (using 1/10 of the LR for transformers blocks)

Training: 20.13% of parameters

Can we go deeper with full fine-tune?

Not really: Not enough VRAM

?
Use gradient checkpointing and mixed precision? Still not enough.
Use a smaller batch size? Noisy gradients, accentued by the multi-task setting.

Do we want to go deeper? Yes*

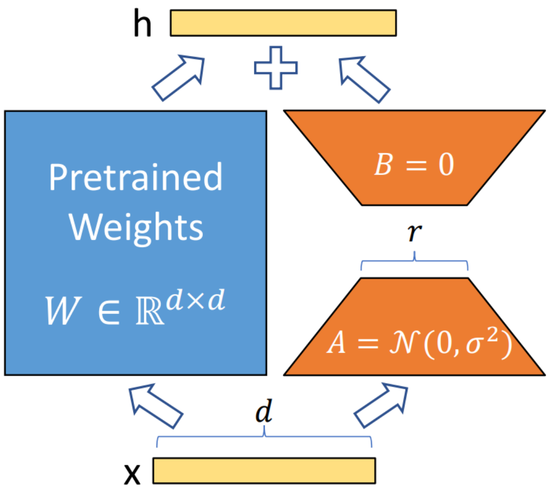
Adjusting the early-layer feature representations within the vision encoder may offer significant benefits.

*
Increase capacity with relatively **small dataset,** possible **overfitting**.

Full fine-tuning could allow task-specific gradients to destroy the network powerful, generalized knowledge, **"Catastrophic forgetting"**

Parameter Efficient Fine-tune with Low Rank Adaptation

Main Idea: the weight updates for large pre-trained model can be effectively represented in a low-dimensional subspace.



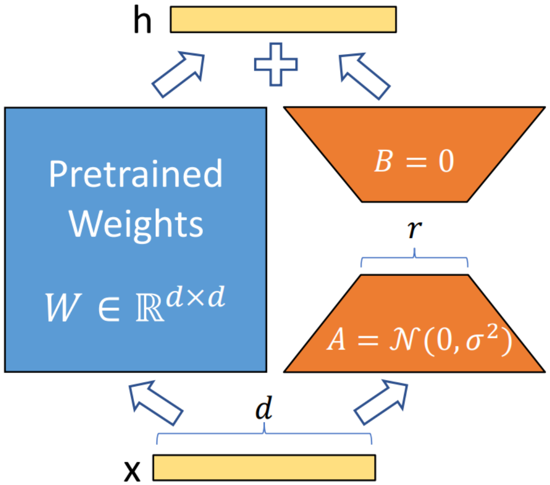
$$h = Wx + BAx$$

Benefits of LoRA:

- Small checkpoints
- **No added inference latency and memory footprint**
- **Lower VRAM consumption** due to parameter-efficient updates
- **Prevents «catastrophic forgetting»** by limiting weight update in a low-rank space

Enhancements for LoRA: **LoRA+** & DoRA

Main Idea of LoRA+: we can achieve a better LoRA fine-tune by employing a differtential learning rate strategy for the LoRA matrices.

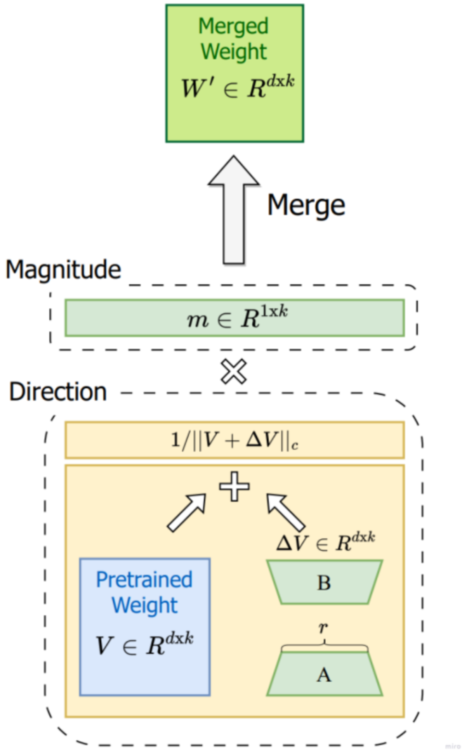


Training update for LoRA+:

$$B^t = B^{t-1} - \lambda \eta G_B$$
$$A^t = A^{t-1} - \eta G_A$$

In our implementation, we set $\boldsymbol{\lambda} = \mathbf{6}$

Enhancements for LoRA: LoRA+ & **DoRA**

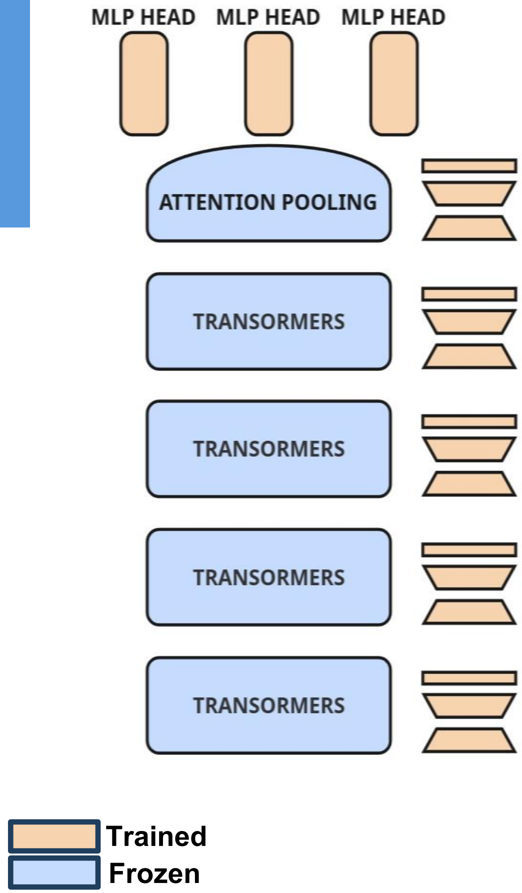


Main Idea: enhance the weight update by seperately training the direction and magnitude of the update.

$$h = x\left(m \cdot \frac{W + BA}{\|W + BA\|_c}\right)$$

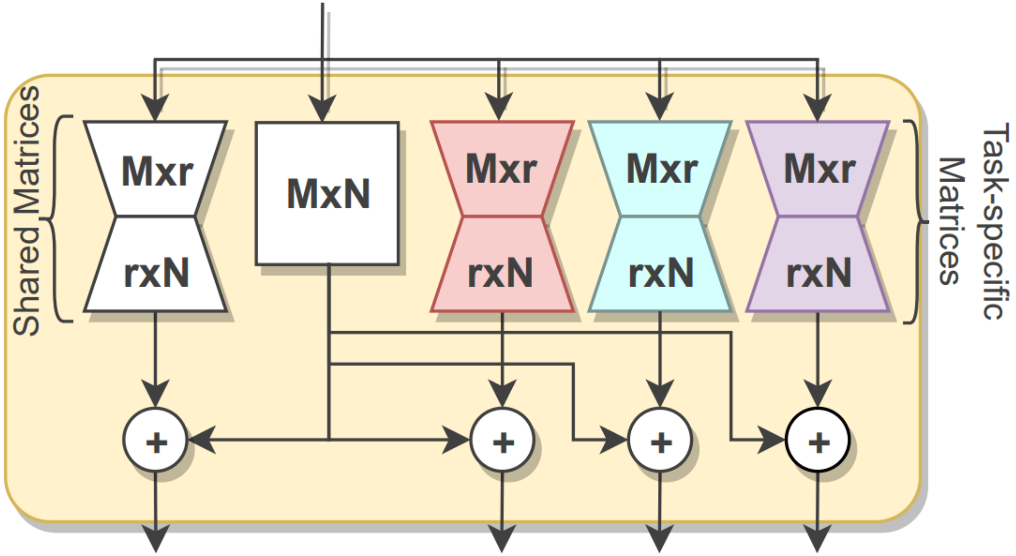
Benefits of DoRA:

- Same benefits of LoRA
- Increased learning capacity



DoRA & LoRA+ : adding DoRA adapters with **rank=64** to **each linear layer** of the Vision Transformer and using a learning rate x6 to train the B matrices of the adapters.
Training: 8.47% of parameters (less than half of FT4)

PEFT & Multi-task: **MTLoRA**



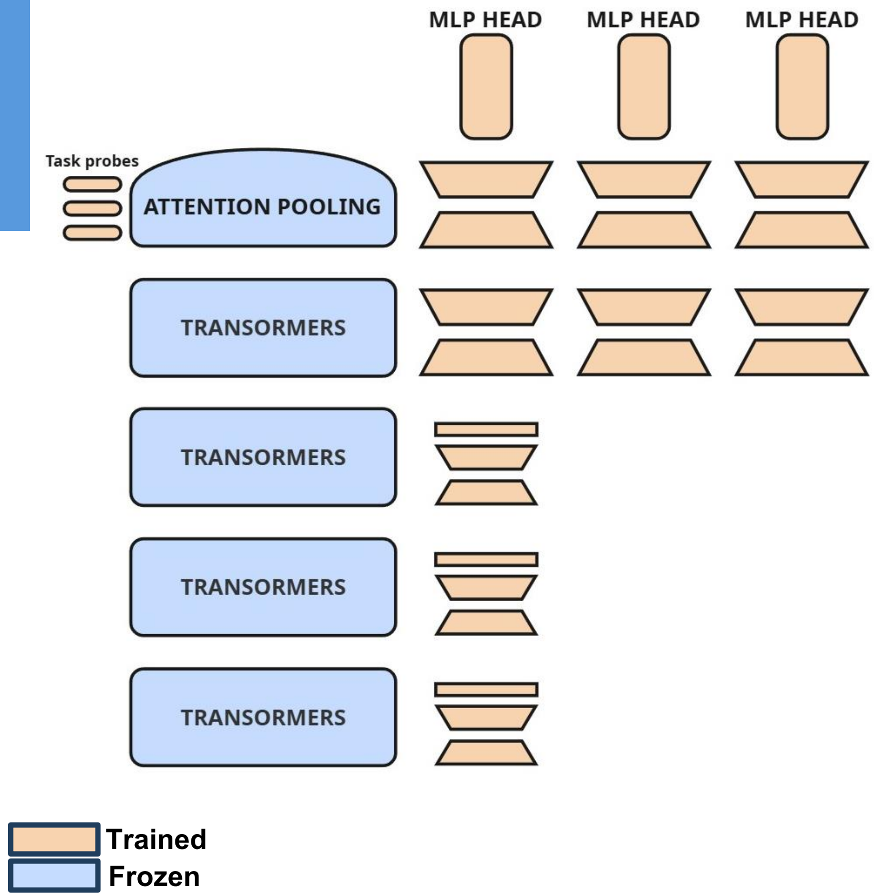
Main Idea: disentangle the parameter space trough **Task-Specific LoRA matrices**

Benefit:

- Train specialized parameters for a task, and use them to create **task-specifics feature-maps**

Deficit:

- The task specific LoRA matrices cannot be merged, small increase in memory footprint and inference latency



MTLoRA: TS-LoRA (r=64) applied to last transformer block and attention pooling layers. TA-DoRA using the same setup describe earlier.

Training: 10.38% of parameters

01 Introduction

02 Methodology

03 Results

Multi-task Vs. Single-task

Model	Avg. Age	Avg. Gender	Emotion	Global Avg
Baseline	47.36%	96.67%	66.57%	69.59%
LP	60.10%	97.51%	84.82%	80.81%
FT_4 (ST)	62.80%	97.57%	88.78%	83.05%
FT_4 (MTL)	63.00%	97.41%	88.43%	82.94%
LoRA (ST)	63.72%	97.56%	90.83%	84.04%
LoRA (MTL)	63.53%	97.49%	91.21%	**84.07%**
MTLoRA	64.03%	97.51%	90.06%	83.86%

-0.11%

+0.03%

-0.18%

Multi-task approach was **successful in preventing negative-transfer**, all of our multi-task model achieve **perfomance parity compared to their single-task equivalent**

Efficiency Comparison

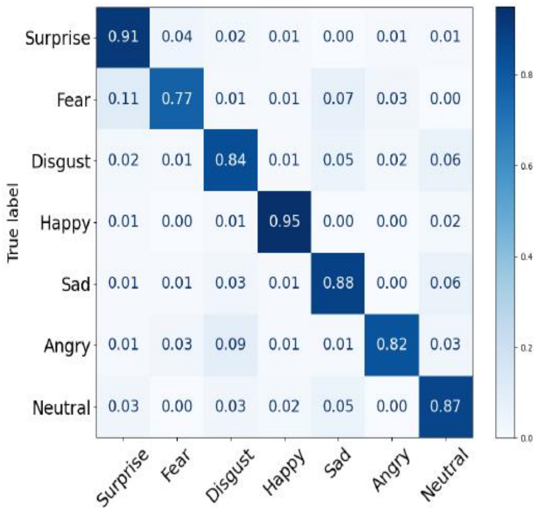
Model	GLOPS	PARAMS	AVG. ACC
Baseline	700	671 M	69.61%
LoRA	352 (-50%)	320 M (-52%)	84.07% (+20%)
MTLoRA	368 (-47%)	329 M (-50%)	83.89% (+20%)

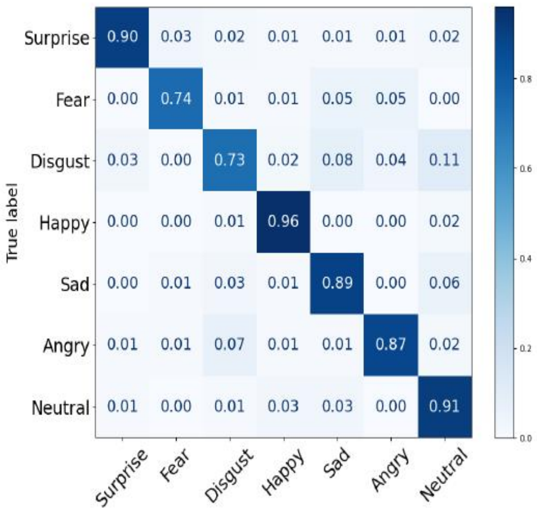
Multi-task model detailed results

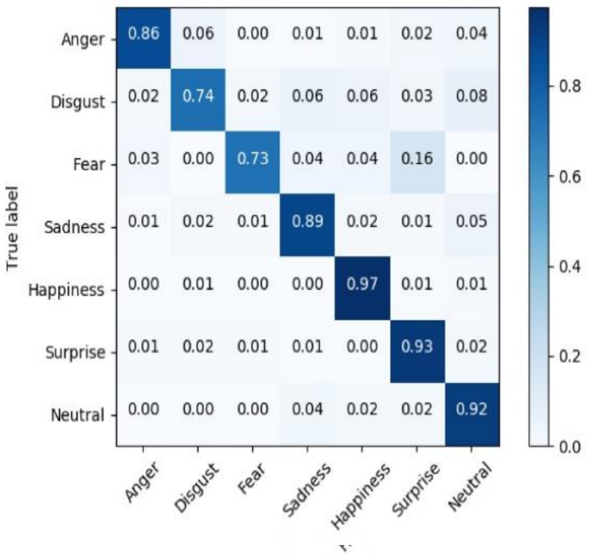
Model	FairFace (Age)	FairFace (Gender)	RAF-DB (Emotion)	UTKFace (Age)	UTKFace (Gender)	VggFace2 (Age)	VggFace2 (Gender)
Baseline	46.11%	97.60%	66.57%	48.43%	96.63%	42.01%	95.78%
LP	61.00%	97.70%	84.82%	61.56%	**97.00%**	57.75%	97.82%
FT_4 (MTL)	63.45%	**97.71%**	88.42%	62.54%	96.68%	61.68%	97.81%
LoRA (MTL)	63.73%	97.57%	**91.21%**	63.34%	96.90%	61.69%	**98.00%**
MTLoRA	**64.11%**	97.62%	90.06%	**63.96%**	96.93%	**62.74%**	**98.00%**

Comparison with the state of the art

Model	FairFace (Age)	FairFace (Gender)	RAF-DB (Emotion)
MIVOLO	62.28%	97.50%	-
CLIP-ViT-L/14	63.45%	97.10%	-
ResEmoteNet	-	-	**94.76%**
ApViT	-	-	92.21%
Baseline	46.11%	97.60%	66.57%
LoRA	63.73%	97.57%	91.21%
MTLoRA	**64.11%**	**97.62%**	90.06%







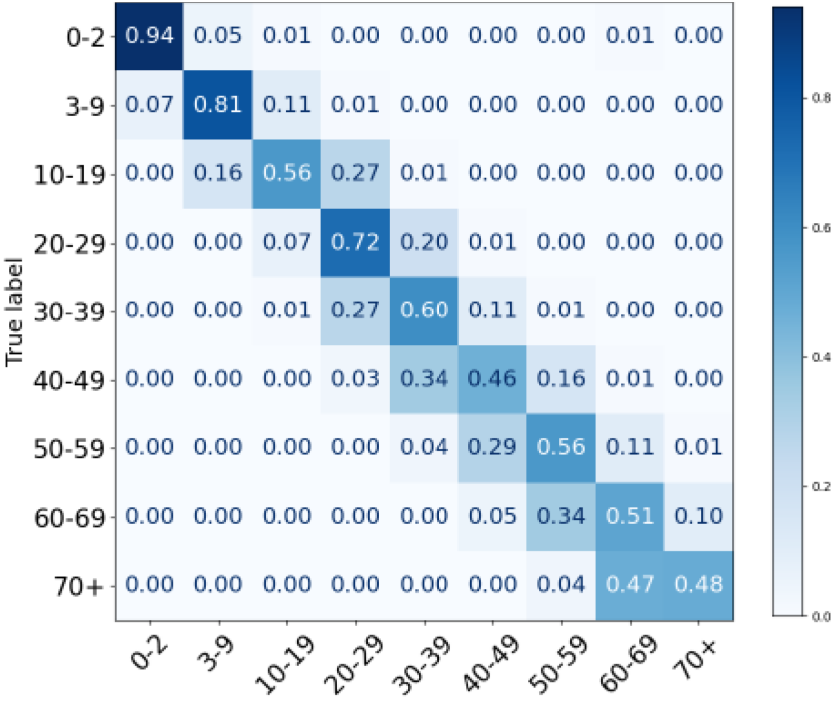
(a) MTLoRA confusion matrix on RAF-DB, balanced accuracy of 86.17 (Acc. 90.06)

-0.19%

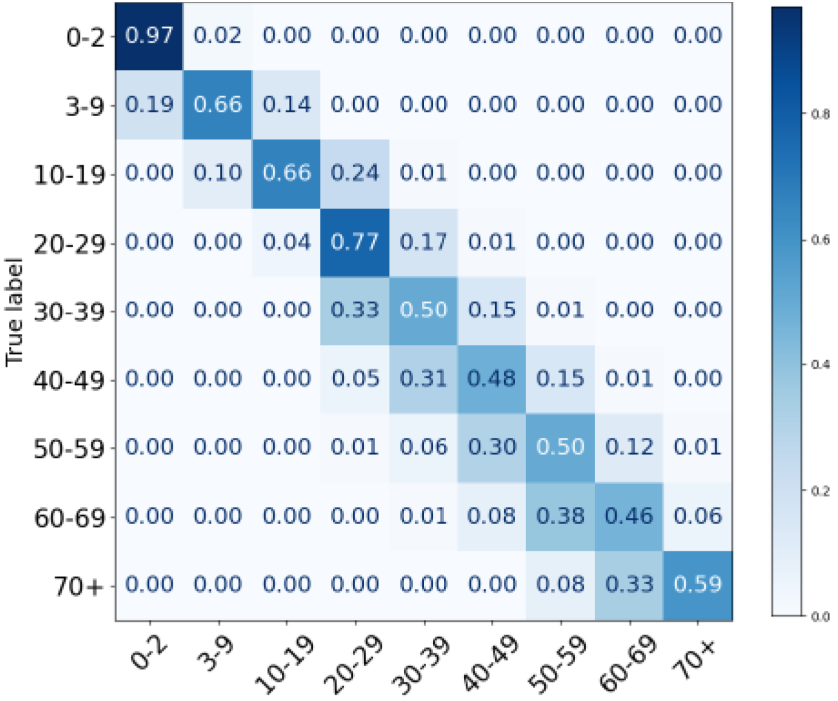
(b) LoRA confusion matrix on RAF-DB, balanced accuracy of 85.90 (Acc. 91.21)

-0.46%

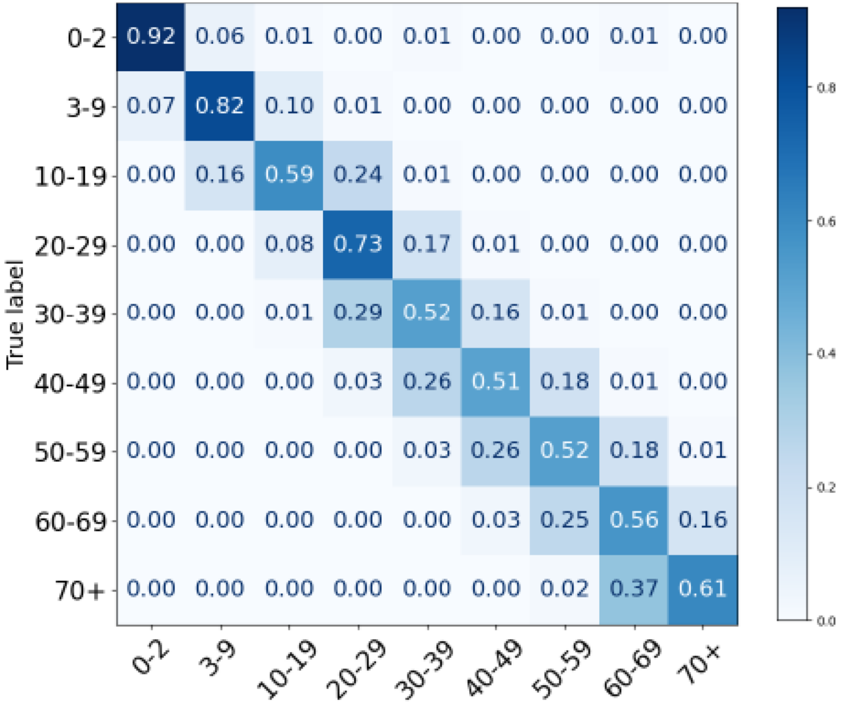
(c) APViT confusion matrix on RAF-DB, balanced accuracy of 86.36 (Acc. 92.21)



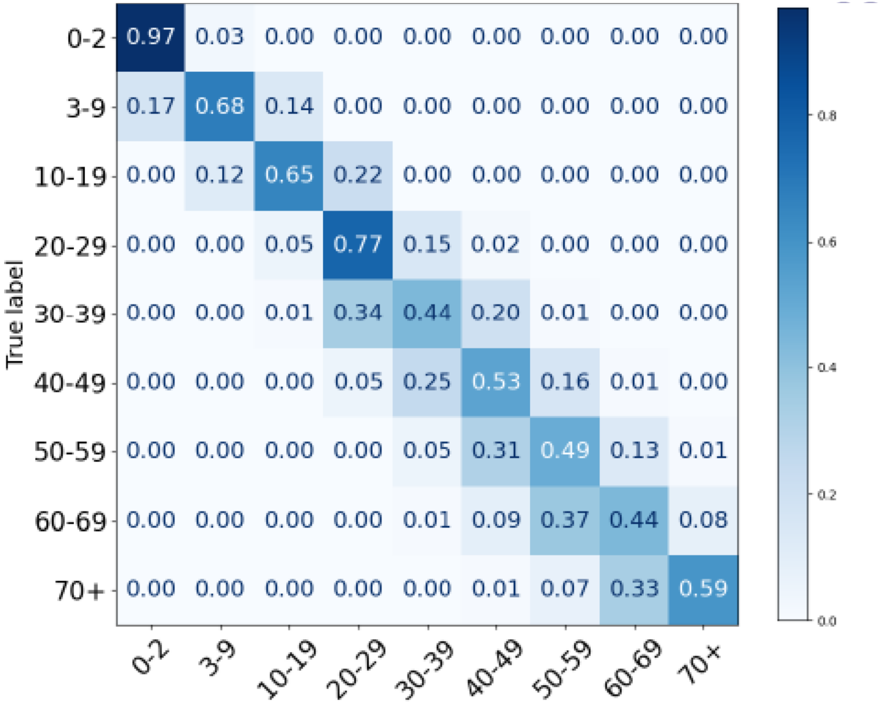
(a) MTLoRA confusion matrix on FairFace, balanced accuracy of 62.78 (Acc. 64.11)



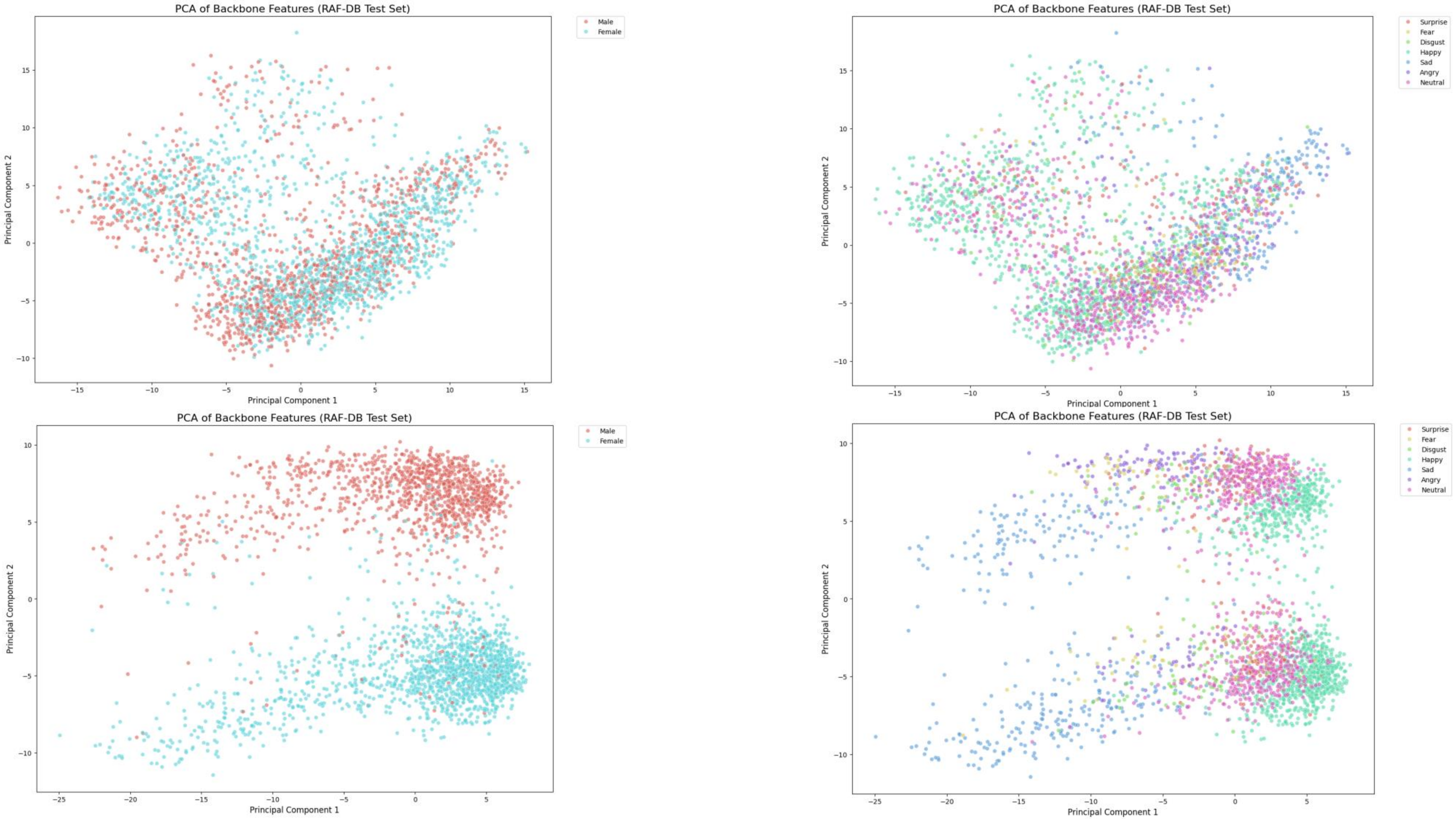
(b) MTLoRA confusion matrix on UTKFace, balanced accuracy of 61.98 (Acc. 63.96)



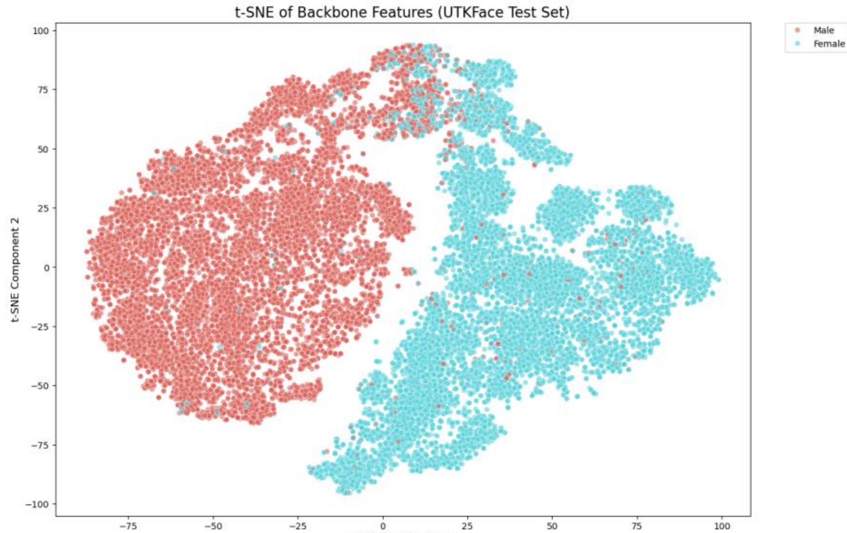
(c) LoRA confusion matrix on FairFace, balanced accuracy of 64.26 (Acc. 63.73)

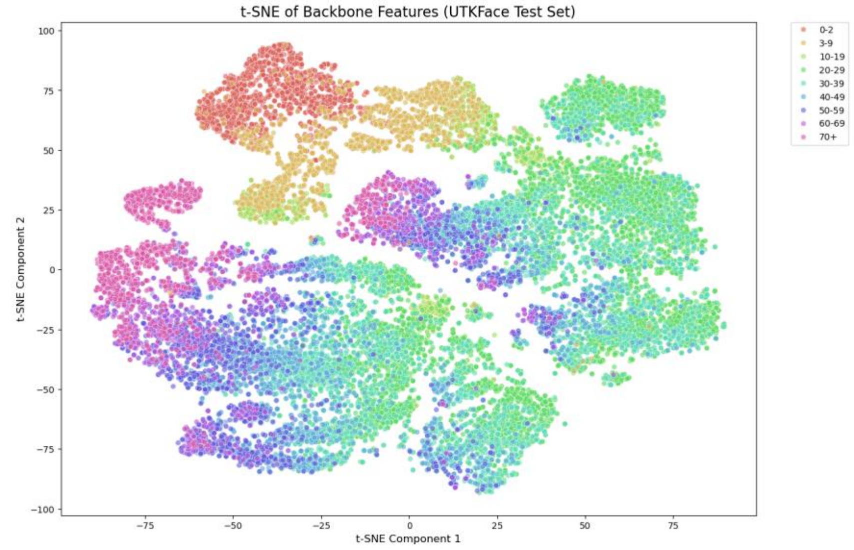


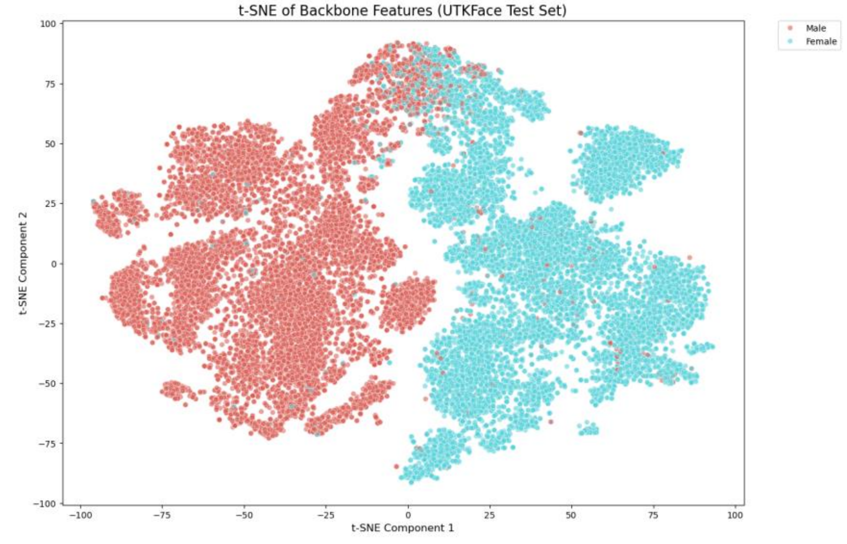
(d) LoRA confusion matrix on UTKFace, balanced accuracy of 61.91 (Acc. 63.34)


PCA of Backbone Features (RAF-DB Test Set)
Male
Female
Principal Component 1
Principal Component 2
PCA of Backbone Features (RAF-DB Test Set)
Surprise
Fear
Disgust
Happy
Sad
Angry
Neutral
Principal Component 1
Principal Component 2
PCA of Backbone Features (RAF-DB Test Set)
Male
Female
Principal Component 1
Principal Component 2
PCA of Backbone Features (RAF-DB Test Set)
Surprise
Fear
Disgust
Happy
Sad
Angry
Neutral
Principal Component 1
Principal Component 2


t-SNE of Backbone Features (UTKFace Test Set)
t-SNE Component 2
0-2
3-9
10-19
20-29
30-39
40-49
50-59
60-69
70+
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75
100


t-SNE of Backbone Features (UTKFace Test Set)
t-SNE Component 2
Male
Female
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75
100


t-SNE of Backbone Features (UTKFace Test Set)
t-SNE Component 2
t-SNE Component 1
0-2
3-9
10-19
20-29
30-39
40-49
50-59
60-69
70+
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75


t-SNE of Backbone Features (UTKFace Test Set)
t-SNE Component 2
t-SNE Component 1
Male
Female
100
75
50
25
0
−25
−50
−75
−100
−100
−75
−50
−25
0
25
50
75
100

Dimostratore

https://huggingface.co/spaces/Antuke/FaR-FT-PE
https://youtu.be/V6-9QTf1xaQ

Face Classification System

Select Model Checkpoint

mtlora.pt

Model Status

Successfully loaded: mtlora.pt

Features

- **Age Classification**: 9 categories (0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+) + Age estimation with weighted centroid average
- **Gender Classification**: M/F
- **Emotion Recognition**: 7 categories (Surprise, Fear, Disgust, Happy, Sad, Angry, Neutral)
- **Automatic Face Detection**: Detects and analyzes multiple faces
- **Detailed Probability Distributions**: View confidence for all classes

Instructions

1. (Optional) Select a model checkpoint from the dropdown.
2. Upload an image or capture from webcam (or select an example below)
3. Click "Classify Image"
4. View detected faces with age, gender, and emotion predictions below

Upload Image

Drop Image Here
- or -
Click to Upload

Annotated Image

Classify Image

📷 Try with example images

Examples