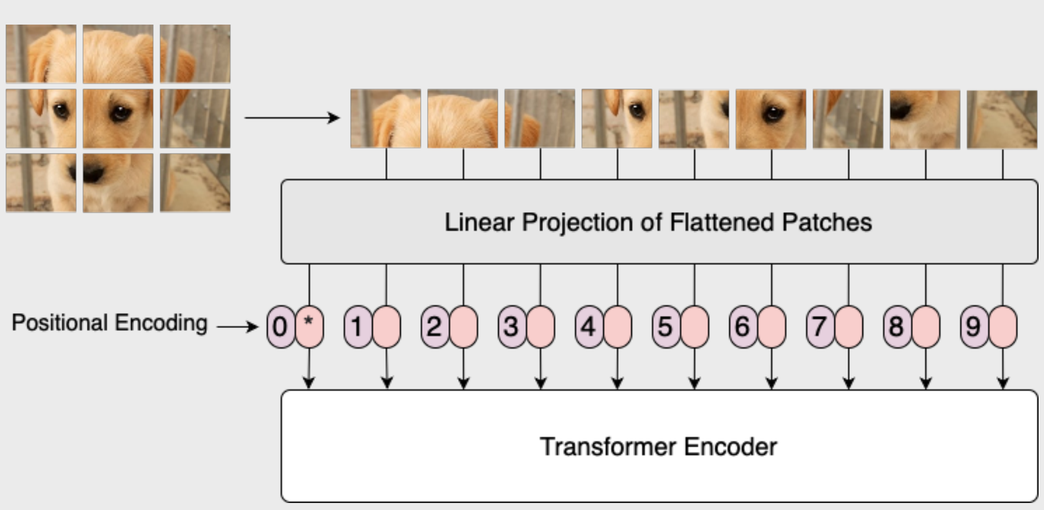
VISION TRANSFORMERS FOR FACIAL RECOGNITION

Šimon Strýček
Supervisor: Ing. Jakub Špaňhel

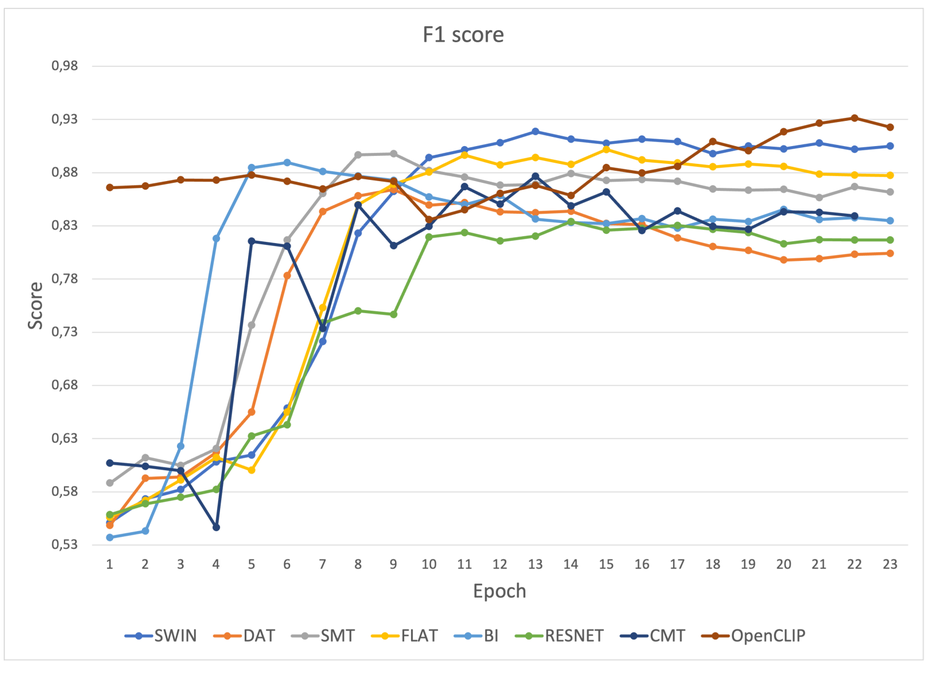
Vision Transformers



Experiments

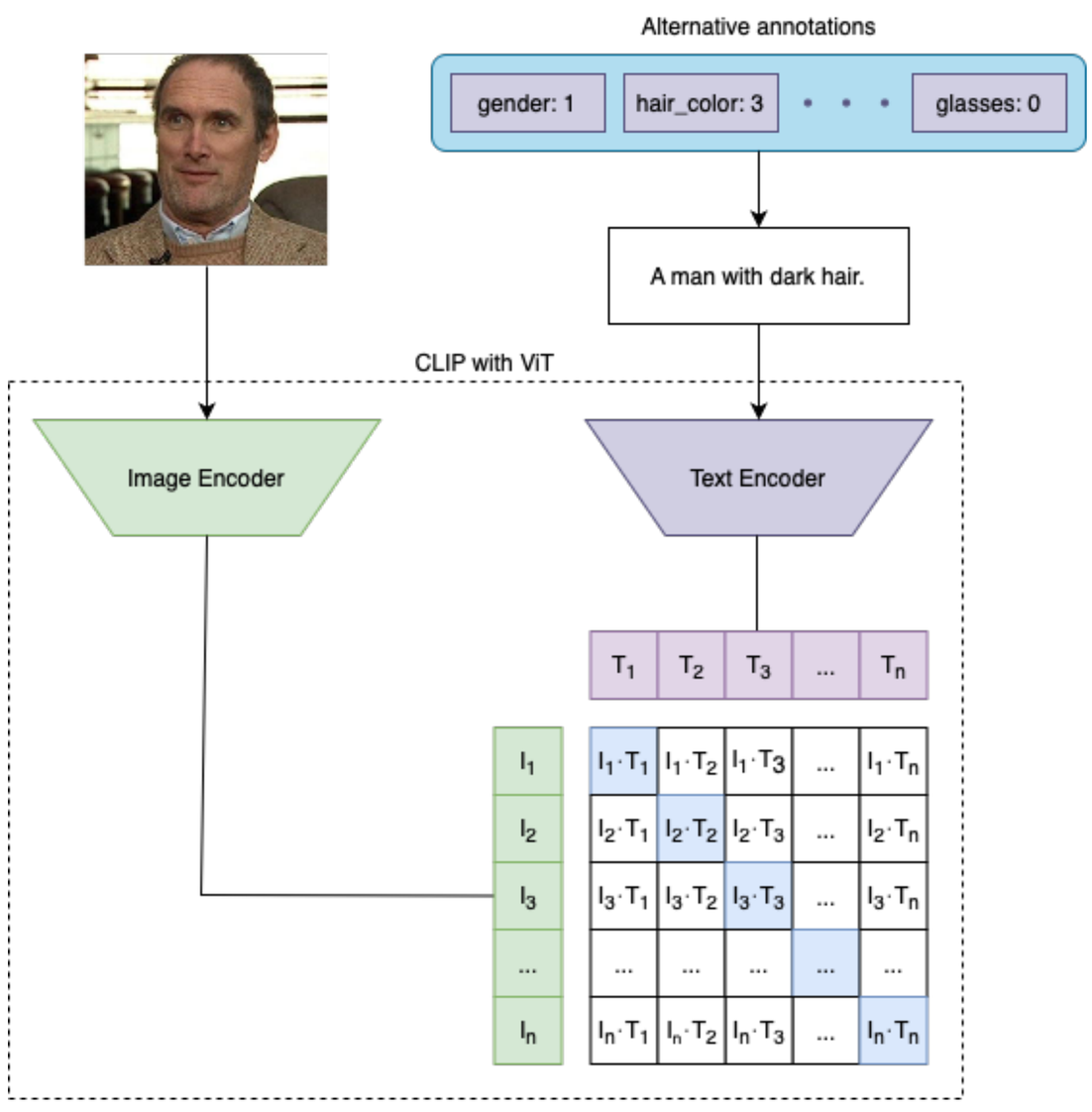
- Testing modern ViT architectures
- Combining different types of training data
- Testing multiple public datasets
- Finding ideal training condititions

Modern ViTs + facial recognition



Combining different types of training data

- Generated textual face description
- CosFace + contrastive CLIP loss



Achieved results

Model	Method	Training dataset	LFW Accuracy [%]
R100	AdaFace	WebFace4M	99.80
CLIP (visual only)	CosFace	MS1Mv3	98.98
SWIN	CosFace	MS1Mv3	98.70
CLIP (multitask)	CosFace + contrast	VGGFace2	94.30

*Only the last three models are subjects of this paper.

Experimenting with loss function

ArcFace

- EER threshold too high
- Plateau in suboptimal state
- Unable to reach >0.95 Acc

CosFace

- EER threshold low
- Slower convergence
- Able to reach >0.98 Acc