PART-BASED FACE RECOGNITION WITH VISION TRANSFORMERS

Zhonglin Sun, Georgios Tzimiropoulos
Queen Mary university of London

Contribution

- Investigate the training strategy for ViT in the task of face recognition.
- Parts-based pipeline for deep face recognition with discriminatively patches learning
- Part-based model is able to learn stable landmarks to competitive degree

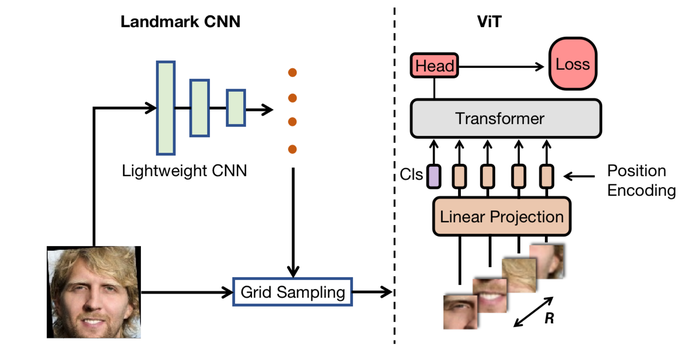

Fig. 1: The pipeline of our part-based fViT

fViT: ViT for Face Recognition

Structure and Flops

- 12 layers, 11 attention heads, d = 768, MLP dim: 2048

Model	Hidden size	Parameters	FLOPS
part fViT-B	768	66M	12.64G
fViT-B	768	63M	12.58G
Resnet-100	-	65M	12.10G
part fViT-S	512	46M	8.96G
fViT-S	512	43M	8.90G
Resnet-50	-	43.59M	6.33G

Fig. 2: The strucutre of our part-based fViT

Hyperparameters for training fViT, inspired by [3]

- fViT is prone to overfitting [4]
- stochastic depth regularization probability: 0.1
- resize & crop: $[0.9, 1.0]$
- RandAugment magnitude: 0.2
- Mixup with alpha and probability:0.5, 0.2
- Cutout probability: 0.1
- weight decay: 1e-1 for fViT and 5e-2 for landmark CNN
- optimazation: AdamW; 34 epochs; train from scratch

part-based fViT

The ViT operates on a sequence of visual token which do not need to be computed on uniform grid. Inspired by work on part-based FR [1], herein we describe how to apply ViT on patches representing facial parts, which can be found in Fig 1.
(1)we use a lightweight weight CNN to predict a set of $R = P \times P$ landmarks. The landmark CNN in our setting is MobilenetV3:

$$\mathbf{r} = \mathrm{CNN}(\mathbf{X}),\ r_i = [x_i, y_i]^T,\ i = 1, \ldots, P^2, \tag{1}$$

(2)Use Differentiable grid sampling method of STN [2] for extracting each patch centered at each landmark coordinate r_i

(3)Send the sampled patches to fViT.

Results

We achieve SOTA results on MS1MV3 dataset and the forward error for evaluating the stability of landmarks, can be found in Fig3 and Fig 4

Method	LFW	CFP-FP	AgeDB	IJB-B	IJB-C	MegaFace/id	MegaFace/ver
CosFace[49]	99.81	98.12	98.11	94.80	96.37	97.91	97.91
ArcFace[9]	99.83	92.27	92.28	94.25	96.03	98.35	98.48
GroupFace[27]	98.85	98.63	96.20	94.93	96.26	98.74	98.79
CircleLoss[45]	99.73	96.02	-	-	93.95	98.50	98.73
DUL[4]	99.83	98.78	-	-	94.61	98.60	-
CurricularFace[21]	99.80	98.37	98.32	94.8	96.1	98.71	98.64
Sub-center ArcFace[11]	99.80	98.80	98.31	94.94	96.28	98.16	98.36
FAN-Face[58]	99.85	98.63	98.38	94.97	96.38	98.70	98.95
BroadFace[28]	99.85	98.63	98.38	94.97	96.38	98.70	98.95
ArcFace-challenge[12]	99.85	99.06	98.48	-	96.81	-	-
VPL[13]	99.83	99.11	**98.60**	95.56	96.76	98.80	**98.97**
ALN[64]	-	96.53	97.25	93.13	95.27	-	-
VirFace[31]	99.56	97.15	-	88.90	90.54	-	-
MagFace[38]	99.83	98.46	96.15	94.51	95.97	-	-
SCL[32]	99.80	98.59	98.26	94.74	96.09	81.40	97.15
Face Transformer [65]	99.83	96.19	97.82	-	95.96	-	-
fViT-B, ours	**99.85**	99.01	98.13	95.97	97.21	98.69	98.91
Part fViT-B, ours	99.83	**99.21**	98.29	**96.11**	**97.29**	**98.96**	98.78

Fig. 3: The results on MS1M dataset

	Method	MAFL	AFLW
Supervised	TCDCN [19]	7.95	7.65
	MTCNN [18]	5.39	6.90
Unsupervised	Thewlis [15]	7.15	-
	Jakab [5]	3.19	6.86
	Zhang [17]	3.46	7.01
	Shu [13]	5.45	-
	Sahasrabudhe [11]	6.07	-
	Sanchez [12]	3.99	6.69
	Mallis [9]	4.12	7.37
	Li [7]	**3.08**	**6.20**
Ours	Landmark CNN	4.87	10.22
	Landmark CNN ($R = 49$)	3.37	7.16
	Landmark CNN ($R = 16$)	3.88	7.69

Fig. 4: The forward error on the MAFL & AFLW dataset

Visualization

Visualization of learned landmarks in R=16, and R=49, can be found in Fig 5 and Fig 6.

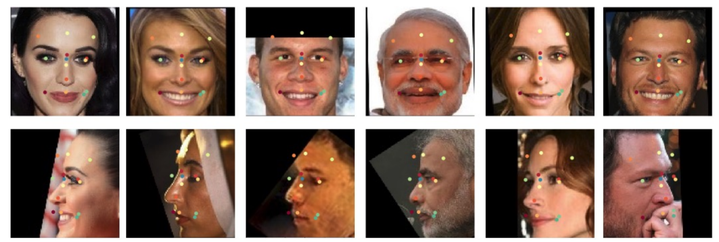

Fig. 5: The learned landmarks when R=16

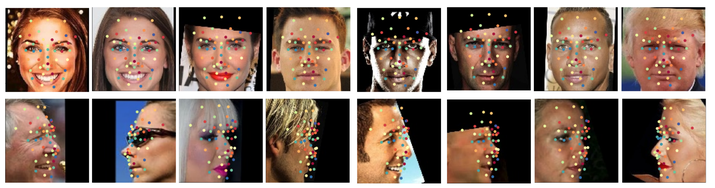

Fig. 6: The learned landmarks when R=49

Acknowledgements

Zhonglin Sun is funded by QMUL-CSC project.

References

[1] Dong Chen et al. “Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3025–3032.

[2] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. “Spatial transformer networks”. In: *Advances in neural information processing systems* 28 (2015), pp. 2017–2025.

[3] Andreas Steiner et al. “How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers”. In: *arXiv preprint arXiv:2106.10270* (2021).

[4] Yaoyao Zhong and Weihong Deng. “Face Transformer for Recognition”. In: *arXiv preprint arXiv:2103.14803* (2021).