Make the Best of Face Clues in iQIYI Celebrity Video Identification Challenge 2019

TOP 4
MAP: 0.8983

Xi Fang
Shanghai Jiao Tong University
seefun@sjtu.edu.cn

Ying Zou
Shanghai Jiao Tong University
zouying@sjtu.edu.cn

iQIYI-VID-2019 dataset

iQIYI-VID-2019 dataset [1]
· The first video dataset for multi-modal person identification.
· Composed of more than 200k video clips of 10,034 celebrities, divided into three parts, 40% for training,30% for validation and 30% for testing.
· The dataset contains multi-modal features extracted by iQIYI baseline method described in [1].
· The face features are 512-dimensional semi-precision float points vectors, and the face quality scores are also provided with face feature vectors.
· The dataset also provides 512-dimensional head, body and audio feature vectors.

Benchmark

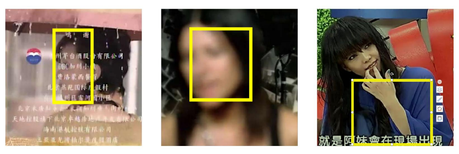
Mean Average Precision (MAP) score in retrieval [2]:

$$MAP(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{j=1}^{n_i}\text{Precision}\left(R_{i,j}\right)$$

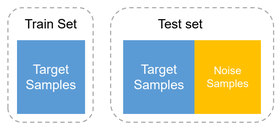
where Q is the set of person IDs to retrieve, m_i is the number of positive examples for the i-th ID, n_i is the number of positive examples within the top k retrieval results for the i-th ID, and $R_{i,j}$ is the set of ranked retrieval results from the top until you get j positive examples. In iQIYI's implementation, only top 100 retrievals are kept for each person ID.

Difficulties

· **Noises of face detection and face quality** in iQIYI-VID-2019 dataset:



· 10034 classes of identities (target samples) and **open-set** retrival:



· **Noises in feature extraction**

· **Huge size of data** (more than 200k video clips)

Our results

In **Validation** Set:

Method	MAP(%)
Baseline face model	81.97
+ qFF	82.93
+ qFF + qFDA	84.47

In **Test** Set:

Method	MAP(%)
iQIYI face baseline	85.19
Ours (training set only)	85.09
+ validation set	88.11
+ 5-fold blending	89.55
+ Ensemble	89.83

Reference

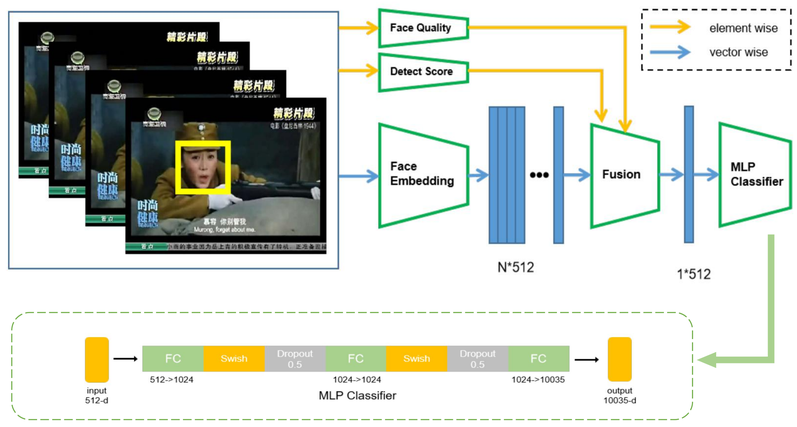
[1] Yuanliu Liu, Peipei Shi, Bo Peng, He Yan, Yong Zhou, Bing Han, Yudi Zheng, Chao Lin, Jianbin Jiang, Yin Fan, Tingwei Gao, Ganwen Wang, Jian Wei Liu, Xiangju Lu, and Danming Xie. 2018. iQIYI-VID: A Large Dataset for Multi-modal Person Identification. ArXiv abs/1811.07548 (2018).

[2] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schutze. [n. d.]. In_x0002_troduction to Information Retrieval? Cambridge University Press 2008. Ch 20 ([n. d.]), 405–416.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4690–4699.

[4] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017).

Pipeline

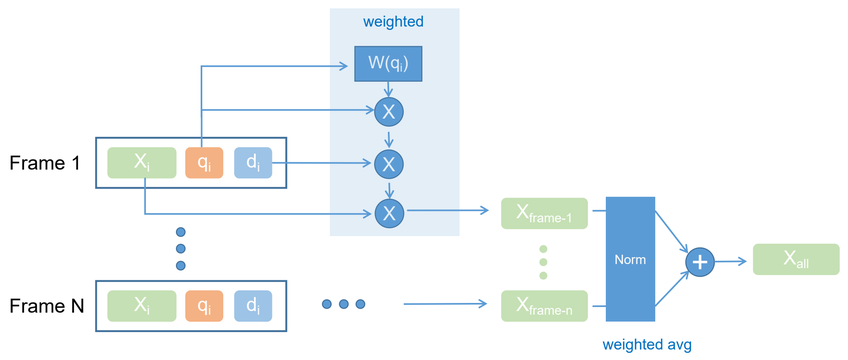


· Using geometric characteristics of face feature vectors extracted by Arcface [3] to fuse face features vectors in single video to **video-wise face feature**.

· Fusion operation based on **face quality score** and **detection confidence**.

· Using **Swish activation [4]** function and **Dropout** in Multi-Layer Perceptron (MLP) classifier.

Feature fusion in testing



qFF (quality-based feature fusion):

$$X_{all} = \frac{1}{\sum_{i=0}^{N} q_i d_i W(q_i)} \sum_{i=0}^{N} q_i d_i W(q_i) X_i$$

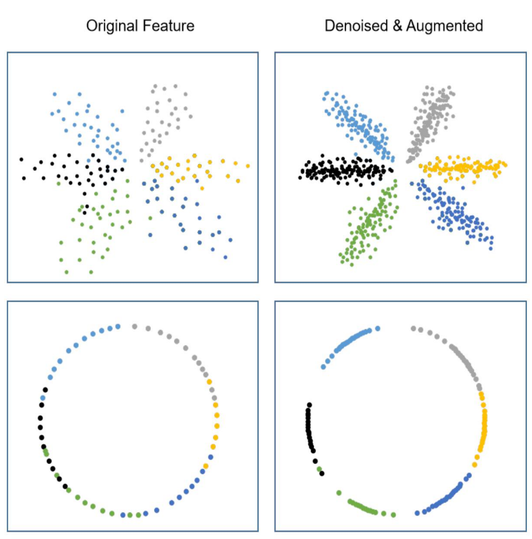
$$W(q_i) = \begin{cases} 0 & q_i \le 0, \\ 0.2 & 0 < q_i \le 20, \\ 0.3 & 20 < q_i \le 30, \\ 0.6 & 30 < q_i \le 60, \\ 1 & q_i > 60. \end{cases}$$

X_i	Face feature vector	X_{frame}	Frame-wise face feature vector
q_i	Face quality score	X_{all}	Video-wise face feature vector
d_i	Face detection confidence	Norm	Normalization
X	multiplication	+	addition

· We fuse face feature vectors in each frame in to **video-wise face feature vector** based on **face quality score** and **face detection confidence.**

· We only classfiy video-wise face feature vectors, which **increase the speed by thousands of times** compared to inference every frame. It only takes about **20 seconds** to inference in test set (60k videos).

Feature denoising and augmentations in training



pre-process feature vectors by threshold screening (quality and det score)

↓

for each training step, random choose 1-5 face feature vector in one video

↓

qFDA: quality-based feature denoising and augmentation

$$X = \frac{1}{\sum_{i=0}^{N} q_i d_i} \sum_{i=0}^{N} q_i d_i X_i$$

↓

train models using online augmentation and denoising method above

Advantages:

1. Preserve and refine class edges in face features' domain
2. Reduce noise from false detection and low-quality face
3. As a regularization method to prevent over-fitting

Training setup:
threshould in detection: 0.8
threshould in quality : 30
Adam + CosineAnnealingLR+ Warmup
Loss: Focal loss + Softmax Loss