Boulder

Computer Vision; Face Recognition and Detection

Maziar Raissi

Assistant Professor

Department of Applied Mathematics

University of Colorado Boulder

maziar.raissi@colorado.edu

DeepFace: Closing the Gap to Human-Level Performance in Face Verification

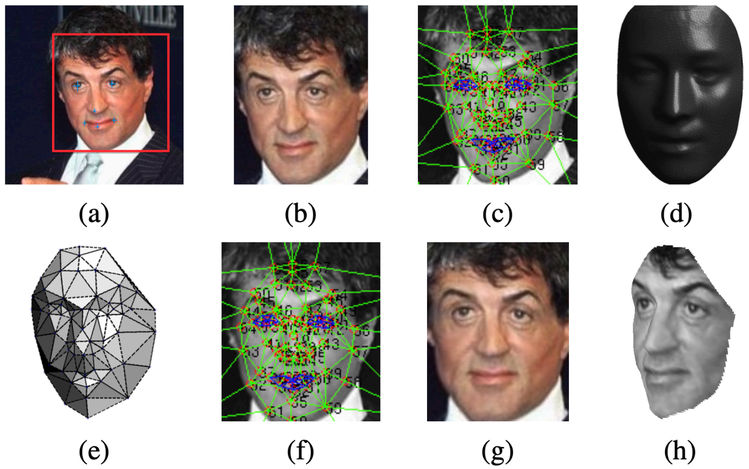
detect ▷ align ▷ represent ▷ classify

Model-based alignment: Explicit 3D face modeling in order to apply a piecewise affine transformation

Face representation from a nine-layer deep neural network

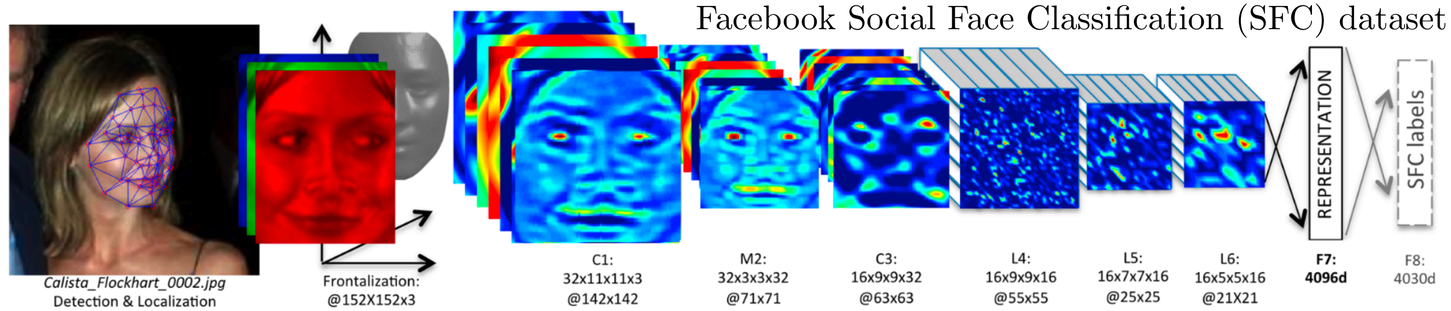
– Labeled Faces in the Wild (LFW) dataset

– YouTube Faces (YTF) dataset



Alignment pipeline. (a) The detected face, with 6 initial fiducial points. (b) The induced 2D-aligned crop. (c) 67 fiducial points on the 2D-aligned crop with their corresponding Delaunay triangulation, we added triangles on the contour to avoid discontinuities. (d) The reference 3D shape transformed to the 2D-aligned crop image-plane. (e) Triangle visibility w.r.t. to the fitted 3D-2D camera; darker triangles are less visible. (f) The 67 fiducial points induced by the 3D model that are used to direct the piece-wise affine warpping. (g) The final frontalized crop. (h) A new view generated by the 3D model (not used in this paper).

Facebook Social Face Classification (SFC) dataset



Normalization | convolutional layer | max-pooling layer

$G(I) \rightarrow$ representation

$$f(I) := \overline{G}(I)/\|\overline{G}(I)\|_2$$

$$\overline{G}(I)_i = \frac{G(I)_i}{\max(G(I)_i, \epsilon)}$$

every location in the feature map learns a different set of filters

Since different regions of an aligned image have different local statistics, the spatial stationarity assumption of convolution cannot hold

Face verification metric: inner product between the two normalized feature vectors

$$\chi^2(f_1, f_2) = \sum_i w_i (f_1[i] - f_2[i])^2 / (f_1[i] + f_2[i])$$

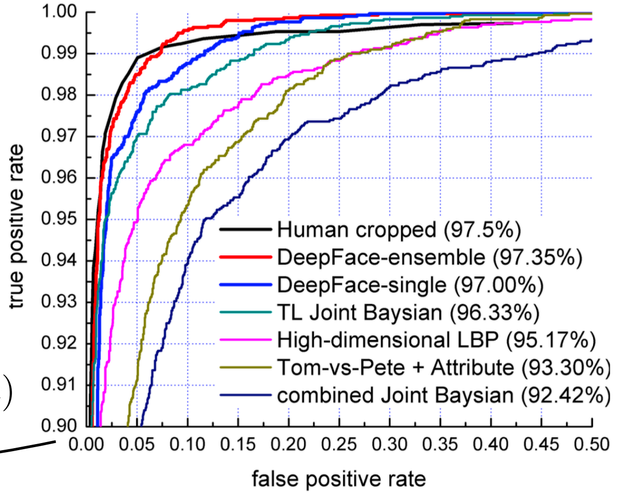
w_i: learned using a linear SVM

χ^2: Weighted $\mathcal{X}^2$ distance

$$d(f_1, f_2) = \sum_i \alpha_i |f_1[i] - f_2[i]|$$

Siamese Network induced distance (trained by cross entropy loss)

The ROC (Receiver Operating Characteristic) curves on the LFW dataset



Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

FaceNet: A Unified Embedding for Face Recognition and Clustering

Boulder

Face verification: Is this the same person?
Face recognition: Who is this person?
Clustering: Find common people among these faces?

A threshold of 1.1 would classify every pair correctly

Triplet Loss

$x \mapsto f(x) \in \mathbb{R}^d$ (image → embedding)

$\|f(x)\|_2 = 1 \rightarrow$ constraint (hypersphere)

1) The squared distance between all faces (independent of imaging conditions) of the same identity should be small
2) The squared distance between a pair of face images from different identities should be large

$x_i^a \rightarrow$ anchor (image of a specific person)
$x_i^p \rightarrow$ positive (other images of the same person)
$x_i^n \rightarrow$ negative (any image of any other person)

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2, \forall (x_i^a, x_i^p, x_i^n) \in \mathcal{T}$$

(α: margin; $\mathcal{T}$: set of all possible triplets)

$$L = \sum_{i=1}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

Triplet Selection

$\arg\max_{x_i^p} \|f(x_i^a) - f(x_i^p)\|_2^2 \rightarrow$ hard positive

$\arg\min_{x_i^n} \|f(x_i^a) - f(x_i^n)\|_2^2 \rightarrow$ hard negative

$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$$

semi-hard negative — avoid model collapse (i.e., $f(x) = 0$)

Face Verification Task

$D(x_i, x_j) \rightarrow L_2$ distance threshold (same/different)

$\mathcal{P}_{\text{same}} \rightarrow$ all pairs (i, j) of the same identity

$\mathcal{P}_{\text{diff}} \rightarrow$ all pairs (i, j) of different identities

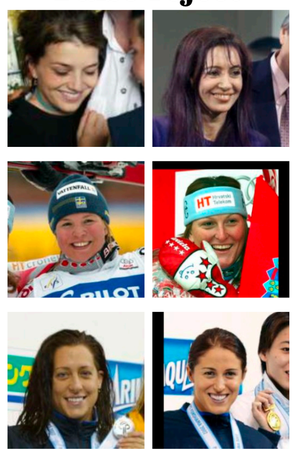
$\text{TA}(d) = \{(i, j) \in \mathcal{P}_{\text{same}}, \text{with } D(x_i, x_j) \leq d\}$ — true accepts

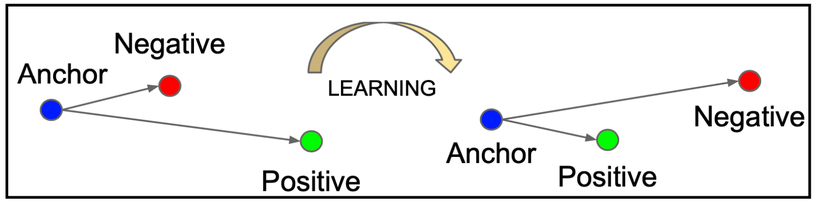
$\text{FA}(d) = \{(i, j) \in \mathcal{P}_{\text{diff}}, \text{with } D(x_i, x_j) \leq d\}$ — false accepts

$$\text{VAL}(d) = \frac{|\text{TA}(d)|}{|\mathcal{P}_{\text{same}}|}, \quad \text{FAR}(d) = \frac{|\text{FA}(d)|}{|\mathcal{P}_{\text{diff}}|}$$

Validation rate and the false accept rate for a given face distance d

False accept **False reject**





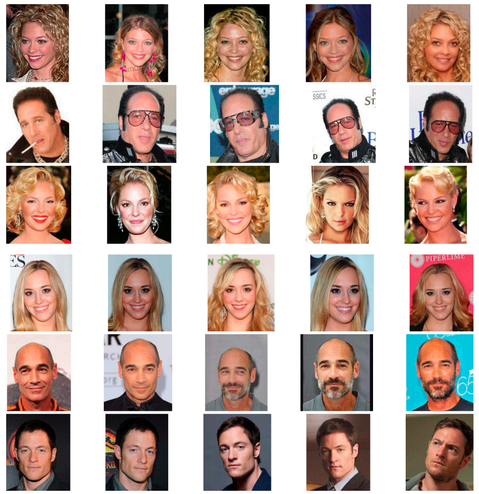
Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

Deep Face Recognition

Boulder

Dataset	Identities	Images
LFW	5,749	13,233
WDRef [4]	2,995	99,773
CelebFaces [25]	10,177	202,599

Dataset	Identities	Images
Ours	2,622	2.6M
FaceBook [29]	4,030	4.4M
Google [17]	8M	200M



Stage	Aim	Type	# of persons	# of images per person	total # images	Annotation effort	100%-EER
1	Candidate list generation	A	5,000	200	1,000,000	–	–
2	Image set expansion	M	2,622	2,000	5,244,000	4 days	–
3	Rank image sets	A	2,622	1,000	2,622,000	–	96.90
4	Near dup. removal	A	2,622	623	1,635,159	–	–
5	Final manual filtering	M	2,622	375	982,803	10 days	92.83

Dataset Collection

– IMDB (Internet Movie Data Base)
– Freebase Knowledge Graph
– Google Image Search

(F, stage 3)
(C, stage 5)

No.	Config	Data	Train Align.	Test Align.	Embedding	100% - EER
1	A	C	No	No	No	92.83
2	A	F	No	No	No	95.80
3	A	F	No	Yes	No	96.70
4	B	F	No	Yes	No	97.27
5	B	F	Yes	Yes	No	96.17
6	D	F	No	Yes	No	96.73
7	B	F	No	Yes	Yes	99.13

The top 50 images (based on Google search rank in the downloaded set) for each identity are used as positive training samples, and the top 50 images of all other identities are used as negative training samples. A one-vs-rest linear SVM is trained for each identity using the Fisher Vector Faces descriptor.

VLAD descriptor ▷ Clustering ▷ Near Duplicate Removal

Final manual filtering is aided with AlexNet trained to discriminate between the 2,622 face identities.

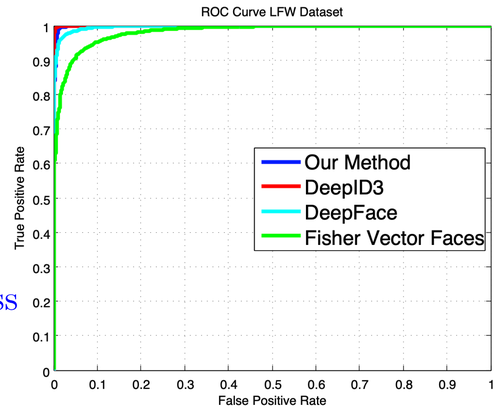
layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

EER: Equal Error Rate
error rate @ the ROC operating point where false positive & false negative rates are equal

Labelled Faces in the Wild

No.	Method	Images	Networks	Acc.
1	Fisher Vector Faces [21]	-	-	93.10
2	DeepFace [29]	4M	3	97.35
3	Fusion [30]	500M	5	98.37
4	DeepID-2,3		200	99.47
5	FaceNet [17]	200M	1	98.87
6	FaceNet [17] + Alignment	200M	1	99.63
7	Ours	2.6M	1	98.95



– Learning a face classifier
– Learning a face embedding using a triplet loss

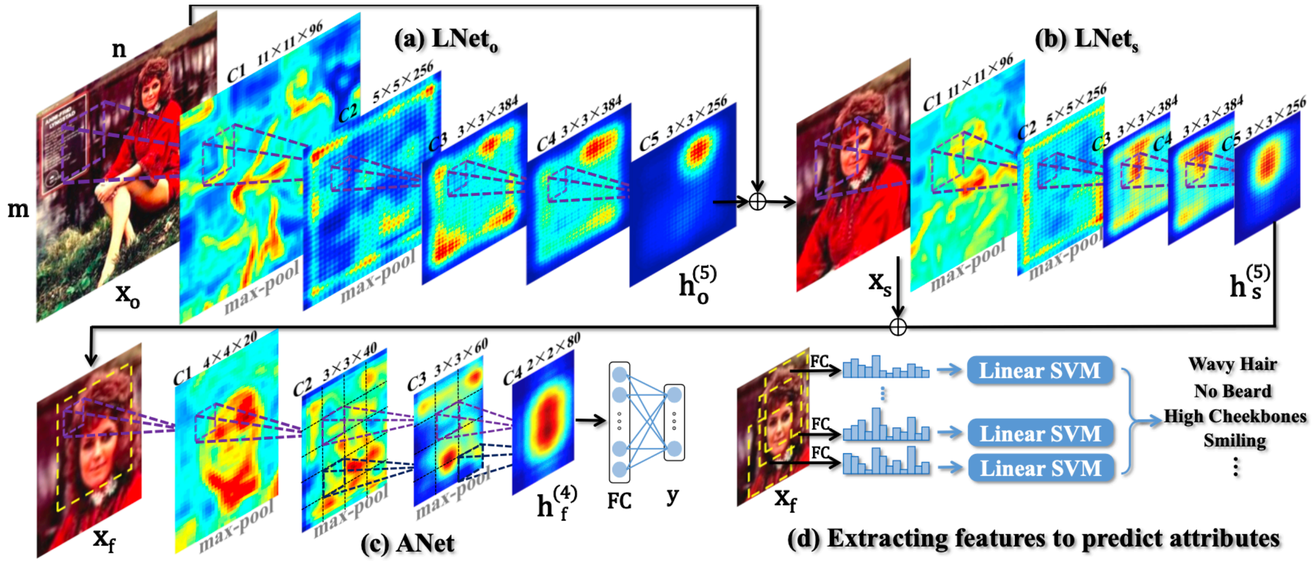
K indicates the number of faces used to represent each video

Youtube Faces Dataset

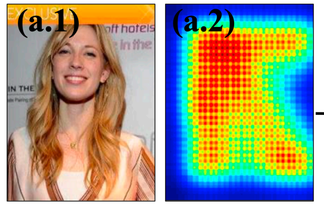
No.	Method	Images	Networks	100%- EER	Acc.
1	Video Fisher Vector Faces [15]	-	-	87.7	83.8
2	DeepFace [29]	4M	1	91.4	91.4
3	DeepID-2,2+,3		200	-	93.2
4	FaceNet [17] + Alignment	200M	1	-	95.1
5	Ours ($K = 100$)	2.6M	1	92.8	91.6
6	Ours ($K = 100$) + Embedding learning	2.6M	1	97.4	97.3

Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." (2015).

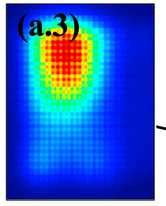
Deep Learning Face Attributes in the Wild



1) Pre-train LNet$_o$ with 1,000 general object categories from the ImageNet dataset



averaged response maps in C5 of LNet$_o$ after pre-training



after fine-tuning

2) Fine-tune LNet$_o$ with attribute tags

$$L = \sum_i y_i \log \underbrace{p(y_i|x)}_{\text{probability of the } i\text{-th attribute given image } x} + (1 - y_i) \log(1 - p(y_i|x))$$

3) For each image, EdgeBox is adopted to propose 500 candidate windows, each of which is measured by a score that sums over its response values normalized by its window size. (Localize a face region using the window with the largest score)

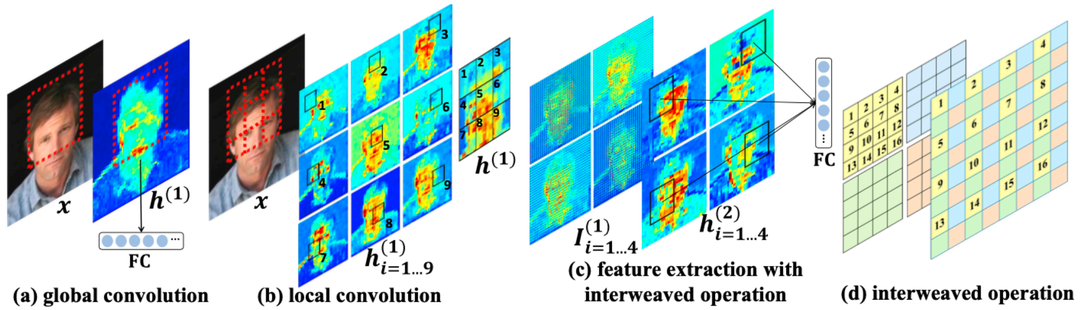
4) Repeat steps 1-3 for LNet$_s$

5) Pre-train ANet with a large number of face identities from the CelebFaces dataset

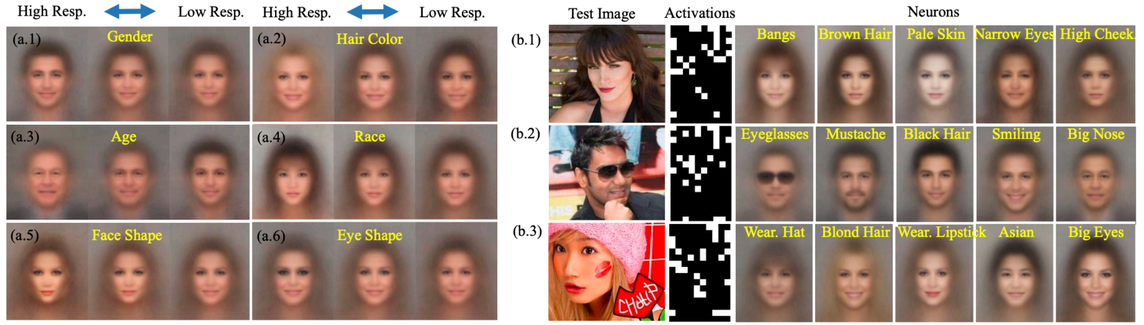
ANet is pre-trained by combining the softmax loss and the similarity loss

$$L = \sum_{i,y_i=y_j} \|\mathrm{FC}_i - \mathrm{FC}_j\|_2^2 \rightarrow \text{improve intra-class invariance}$$

6) ANet is fine-tuned by attributes to learn high-level feature FC



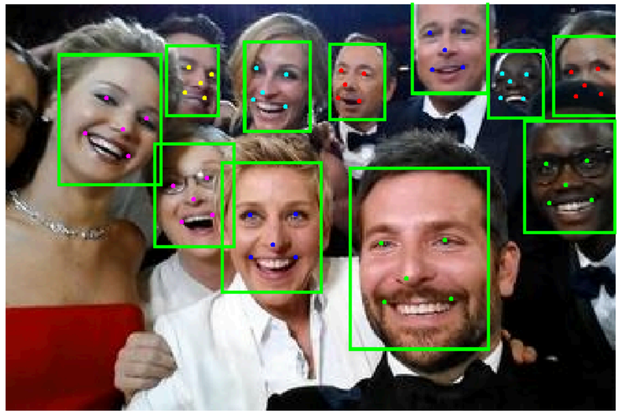
. Detailed pipeline of efficient feature extractions in ANet.

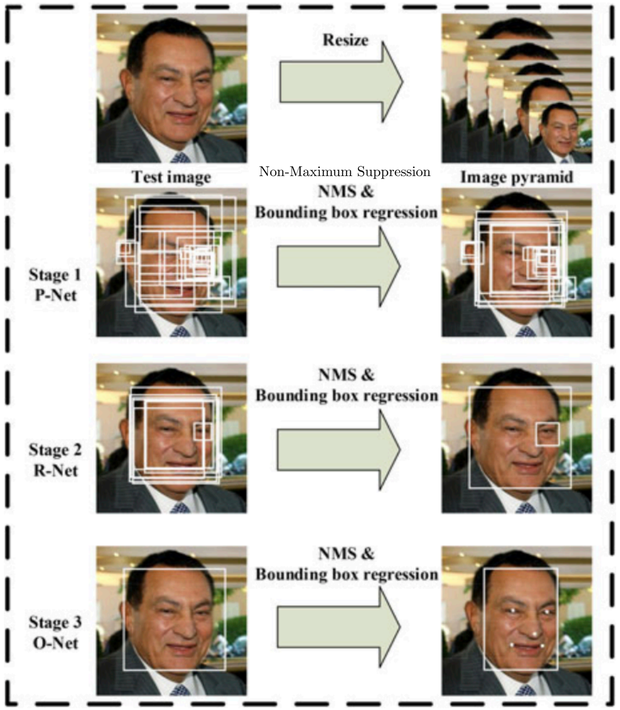


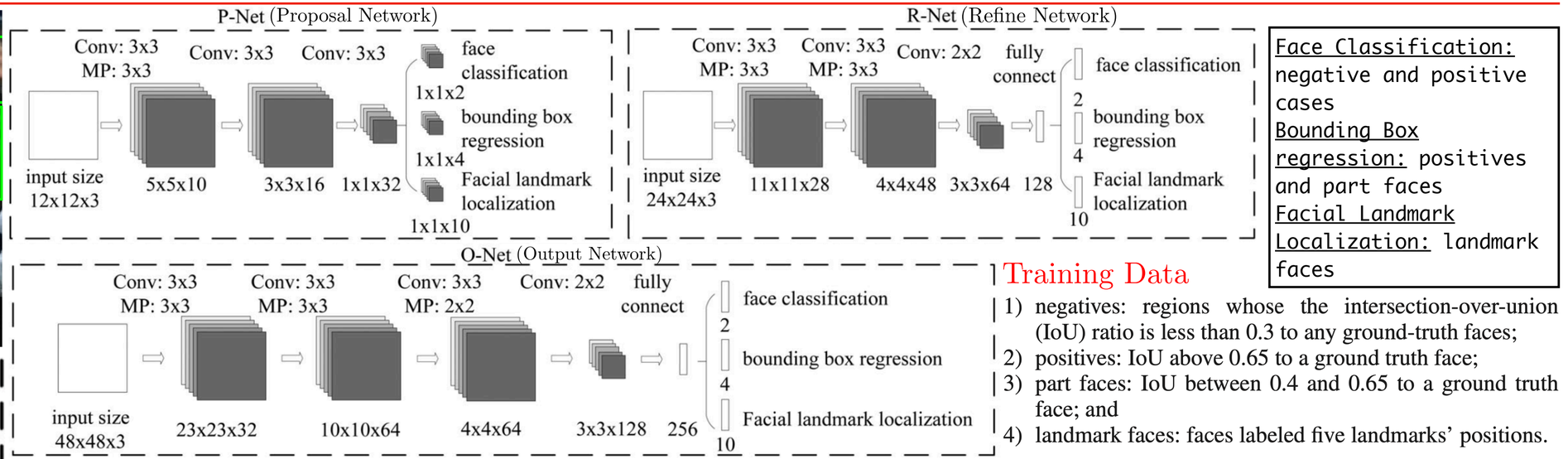
Visualization of neurons in ANet (a) after pre-training (b) after fine-tuning

Liu, Ziwei, et al. "Deep learning face attributes in the wild." Proceedings of the IEEE international conference on computer vision. 2015.

Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks







Face Classification: negative and positive cases
Bounding Box regression: positives and part faces
Facial Landmark Localization: landmark faces

Training Data

1) negatives: regions whose the intersection-over-union (IoU) ratio is less than 0.3 to any ground-truth faces;
2) positives: IoU above 0.65 to a ground truth face;
3) part faces: IoU between 0.4 and 0.65 to a ground truth face; and
4) landmark faces: faces labeled five landmarks' positions.

Face Classification

$x_i \rightarrow$ sample

$$L_i^{\text{det}} = -\left(y_i^{\text{det}} \log\left(p_i\right) + \left(1 - y_i^{\text{det}}\right)\left(1 - \log(p_i)\right)\right)$$

$p_i \rightarrow$ prob. of x_i being a face (produce by the network)

$y_i^{\text{det}} \in \{0, 1\} \rightarrow$ ground-truth label

Bounding Box Regression

$$L_i^{\text{box}} = \left\| \hat{y}_i^{\text{box}} - y_i^{\text{box}} \right\|_2^2$$

$y_i^{\text{box}} \rightarrow$ ground-truth coordinate

$\hat{y}_i^{\text{box}} \rightarrow$ regression target obtained from the network

left, top, height, and width $\implies y_i^{\text{box}} \in \mathbb{R}^4$

For each candidate window, predict the offset between it and the nearest ground truth

Facial Landmark Localization

$$L_i^{\text{landmark}} = \left\| \hat{y}_i^{\text{landmark}} - y_i^{\text{landmark}} \right\|_2^2$$

Five facial landmarks (nose, left & right eyes & mouth corners)

$y_i^{\text{landmark}} \in \mathbb{R}^{10}$

Multisource Training

$\alpha_i \rightarrow$ target importance

$$\min \sum_{i=1}^{N} \sum_{j \in \{\text{det}, \text{box}, \text{landmark}\}} \alpha_j \beta_i^j L_i^j$$

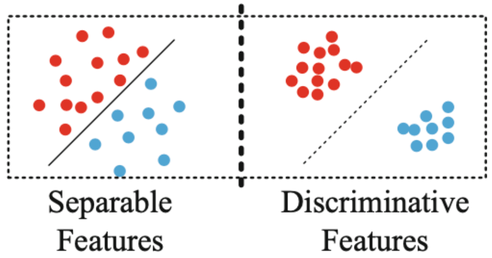
$\beta_i^j \rightarrow$ sample type indicator

Online Hard Sample Mining

In each minibatch, sort the losses computed in the forward propagation from all samples and select the top 70% of them as hard samples.

Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." *IEEE Signal Processing Letters* 23.10 (2016): 1499-1503.

A Discriminative Feature Learning Approach for Deep Face Recognition



– Labeled Faces in the Wild (LFW)
– YouTube Faces (YTF)
– MegaFace Challenge

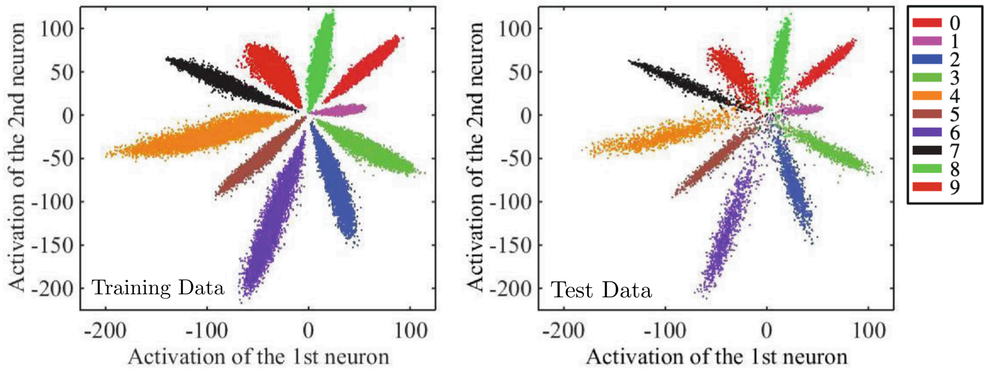
– Face Recognition
– Face Verification

– inter-class dispersion
– intra-class compactness

$$\mathcal{L}_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}} \rightarrow \text{softmax loss}$$

$x_i \in \mathbb{R}^d \rightarrow$ deep features of i-th example belonging to class y_i
$W_j \in \mathbb{R}^d \rightarrow j$-th column of the weights $W \in \mathbb{R}^{d \times n}$
$b \in \mathbb{R}^n \rightarrow$ bias
$m \rightarrow$ size of mini-batch
$n \rightarrow$ number of classes

MNIST Example (CNN : $I_i \mapsto x_i \in \mathbb{R}^2$)



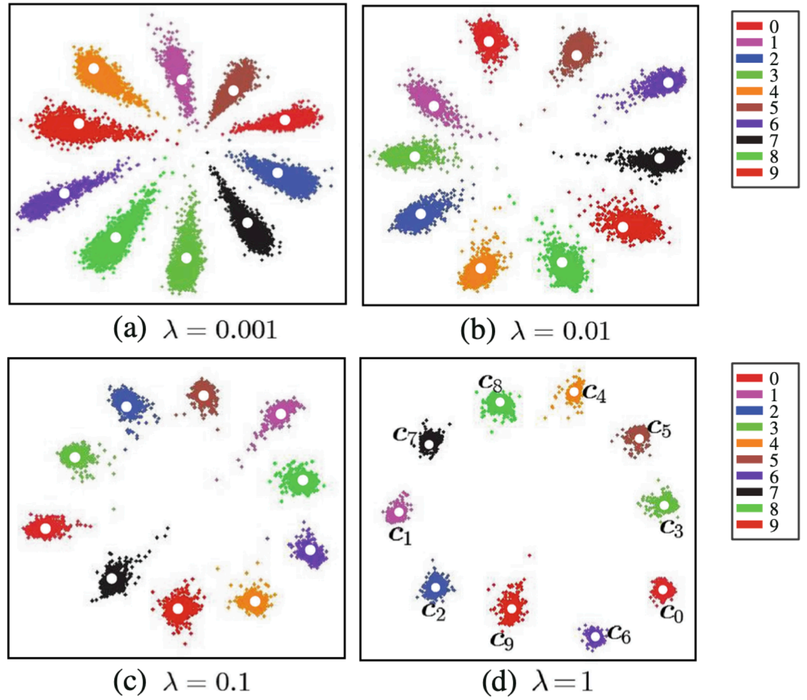
(a) (b)

The Center Loss

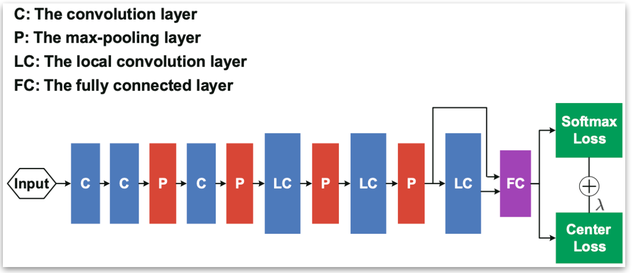
$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

$c_{y_i} \in \mathbb{R}^d \rightarrow y_i$th class center of deep features

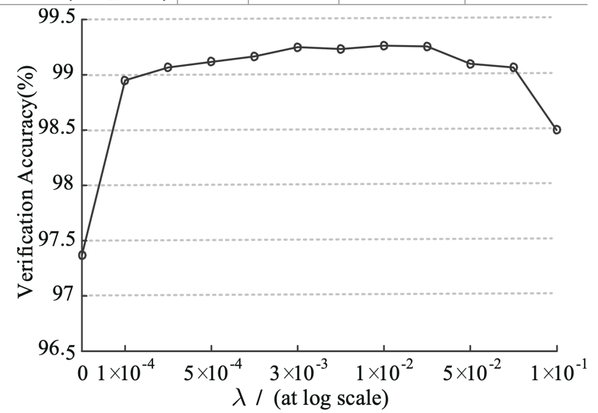
$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C$$



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$
(c) $\lambda = 0.1$ (d) $\lambda = 1$



Method	Images	Networks	Acc. on LFW	Acc. on YTF
DeepFace [34]	4M	3	97.35 %	91.4 %
DeepID-2+ [32]	-	1	98.70 %	-
DeepID-2+ [32]	-	25	99.47 %	93.2 %
FaceNet [27]	200M	1	99.63 %	95.1 %
Deep FR [25]	2.6M	1	98.95 %	97.3 %
Baidu [21]	1.3M	1	99.13 %	-
model A	0.7M	**1**	97.37 %	91.1 %
model B	0.7M	**1**	99.10 %	93.8 %
model C (Proposed)	**0.7M**	**1**	**99.28 %**	**94.9 %**



(a) Face images in LFW

Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition." *European conference on computer vision*. Springer, Cham, 2016.

ArcFace: Additive Angular Margin Loss for Deep Face Recognition

– sotfmax classifier
– triplet loss

sotfmax drawbacks

– size of the linear transformation matrix $W \in \mathbb{R}^{d \times n}$ increases linearly with the identities number n
– good for closed-set classification problems but not discriminative enough for the open-set face recognition problem

triplet loss drawbacks

– combinatorial explosion in the number of face triplets
– semi-hard-example-mining is a quite difficult problem for efficient model training

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}$$

softmax loss

$x_i \in \mathbb{R}^d \rightarrow$ deep feature of the i-th sample
$y_i \rightarrow$ class of the i-th sample
$W_j \in \mathbb{R}^d \rightarrow j$-th column of $W \in \mathbb{R}^{d \times n}$
$b_j \in \mathbb{R}^n \rightarrow$ bias
N & $n \rightarrow$ batch size & class number
set $b_j = 0$
$W_j^T x_i = \|W_j\|\|x_i\|\cos\theta_j$

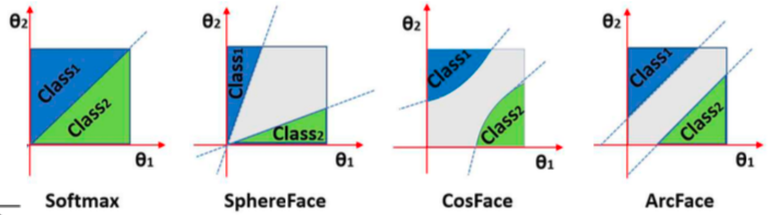
$\|W_j\| = 1$ by l_2 normalization
$\|x_i\| = s$ by l_2 normalization
hyper-sphere with radius s

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}$$

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}$$

$m \rightarrow$ additive angular margin penalty between x_i and W_{y_i} to simultaneously enhance the intra-class compactness and inter-class discrepancy

$$L_4 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(m_1\theta_{y_i}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{y_i}+m_2)-m_3)} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}$$

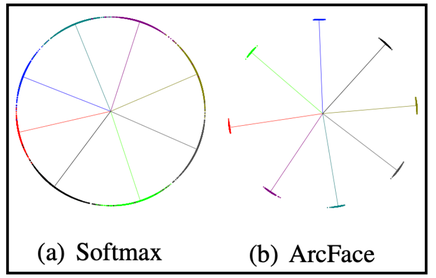


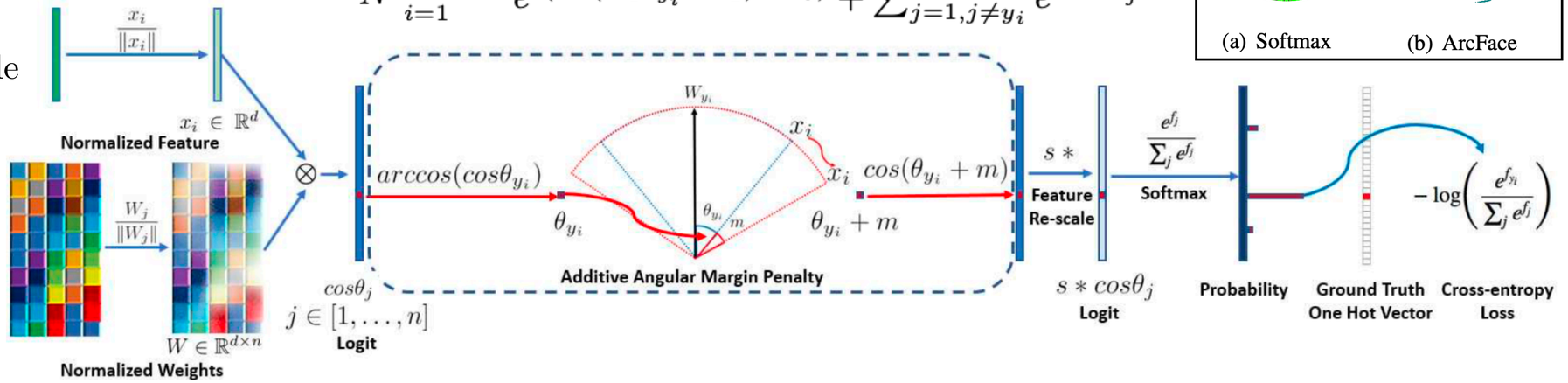
triplet-loss $\leftarrow \arccos(x_i^{pos} x_i) + m \leq \arccos(x_i^{neg} x_i)$

Intra-Loss $\leftarrow$ $L_5 = L_2 + \frac{1}{\pi N}\sum_{i=1}^{N}\theta_{y_i}$

$$L_6 = L_2 - \frac{1}{\pi N(n-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq y_i}^{n}\arccos(W_{y_i}^T W_j)$$

Inter-Loss





Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Boulder

Questions?