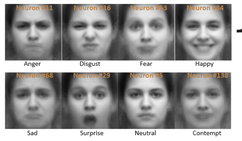


FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition

Hui Ding, Shaohua Kevin Zhou, Rama Chellappa

Problem

Deep convolutional neural network has achieved great success on large-scale image classification task. However, how to effectively train deep network on small dataset is still a challenging problem.

Conventional method is to fine-tune a deep network trained on face recognition dataset to adapt to the facial expression recognition task. This simple strategy has two notable problems:

1. The fine-tuned face net may still contain information useful for subject identification.
2. The network designed for the face recognition domain often has a large capacity, thus the overfitting issue is still severe.

Method

The distribution function of the high-level neurons can be formulated as follows:

$$f(X^l) = C_p \cdot e^{-||X^l||_p^p} \quad (1)$$

To incorporate the knowledge of a face net, we propose to extend (1) to have the following form, i.e., :

$$f(X^l) = C_p \cdot e^{-||X^l - \mu||_p^p} \quad (2)$$

The mean is modeled by the face net. This is motivated by the observation that the fine-tuned face net already achieves competitive performance on the expression dataset, so it should provide a good initialization point for the expression net.

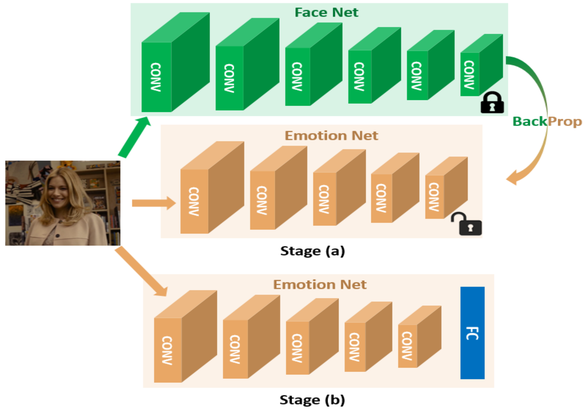
Using the maximum likelihood estimation (MLE) procedure, we can derive the loss function as:

$$\begin{aligned} \max_{\theta_1} L_1 &= \max_{\theta_1} \log f(X^l) \\ &= \max_{\theta_1} \log C_p \cdot e^{-||X^l - \mu||} \\ &= \min_{\theta_1} ||g_{\theta_1}(I) - G(I)||_p^p \end{aligned} \quad (3)$$

- Which layer to transfer?

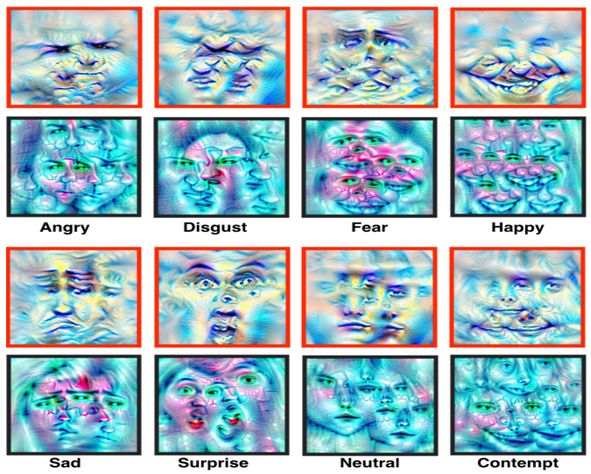
Our experiment results suggest that late middle layer, such as pool5, is a good tradeoff between supervision richness and representation discriminativeness.

Training Strategy



Two-stage Training Algorithm. In stage (a), the face net is frozen and provides supervision for the expression net. The regression loss is back-proped only to the expression net. The convolutional layers are trained in this stage. In stage (b), the randomly initialized fully-connected layers are attached to the trained convolutional blocks. The whole network is trained jointly with cross-entropy loss.

Model Visualization



The red-boxed images are generated by the model trained with our method, while the black-boxed images are from the face network fine-tuned on the expression dataset. We can see the images produced by the face net are dominated with faces, while our model represents the facial expressions better.

Experiments

- Classification Results

CK+

Method	Average Accuracy	#Exp. Classes
CSPL [16]	89.9%	Six Classes
AdaGabor [35]	93.3%	
LBPSVM [36]	95.1%	
3DCNN-DAP [21]	92.4%	
BDBN [19]	96.7%	
STM-ExpLet [20]	94.2%	
DTAGN [22]	97.3%	
Inception [23]	93.2%	
LOMo [37]	95.1%	
PPDN [9]	97.3%	
FN2EN	**98.6%**	
AUDN [18]	92.1%	Eight Classes
Train From Scratch (BN)	88.7%	
VGG Fine-Tune (baseline)	89.9%	
FN2EN	**96.8%**	

OULU-CAS

Method	Average Accuracy
HOG 3D [38]	70.63%
AdaLBP [28]	73.54%
Atlases [39]	75.52%
STM-ExpLet [20]	74.59%
DTAGN [22]	81.46%
LOMo [37]	82.10%
PPDN [9]	84.59%
Train From Scratch (BN)	76.87%
VGG Fine-Tune (baseline)	83.26%
FN2EN	**87.71%**

TFD

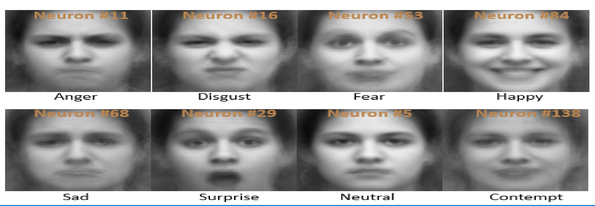
Method	Average Accuracy
Gabor + PCA [40]	80.2%
Deep mPoT [41]	82.4%
CDA+CCA [42]	85.0%
disRBM [43]	85.4%
bootstrap-recon [44]	86.8%
Train From Scratch (BN)	82.5%
VGG Fine-Tune (baseline)	86.7%
FN2EN	**88.9%**

SFEW

Method	Average Accuracy	Extra Train Data
AUDN [18]	26.14%	None
STM-ExpLet [20]	31.73%	
Inception [23]	47.70%	
Mapped LBP [8]	41.92%	
Train From Scratch (BN)	39.55%	
VGG Fine-Tune (baseline)	41.23%	
FN2EN	**48.19%**	
Transfer Learning [25]	48.50%	FER2013
Multiple Deep Network [24]	52.29%	
FN2EN	**55.15%**	

- Top hidden layer neuron visualization

CK+



Oulu-CASIA

Summary

We propose a probabilistic distribution function to model the high level neuron response based on already fine-tuned face net, thereby leading to feature level regularization that exploits the rich face information in the face net. In the second stage, we perform label supervision to boost the final discriminative capability.

As a result, FaceNet2ExpNet improves visual feature representation and outperforms various state-of-the-art methods on four public datasets. In future, we plan to apply this training method to other domains with small datasets.