FaceNet: Facial Expression Recognition

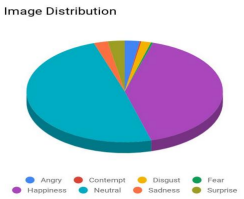
Martin Mbuthia ~ Joseph Wang
maina@stanford.edu ~ wangjoe@stanford.edu

Objective

Our goal is to devise an end-to-end model for classifying images of human faces into eight classes of common facial expressions. We train a deep neural network model based on the Inception V3 architecture. Our highest performance, 80.2% on the test set, after 20 epochs of training. We tackle our biggest challenge, data imbalance, through modifying loss function, data synthesization, etc.

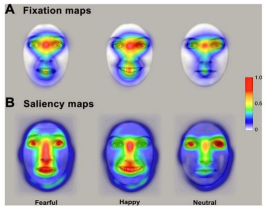
Data

We trained the models on a set of 14,000 RGB images from a Kaggle dataset [1]. Each image also comes with a ground-truth label that will be converted into a one-hot vector. All image data are normalized based on the 256-point scale and have been randomly cropped into 224*224 tensors.



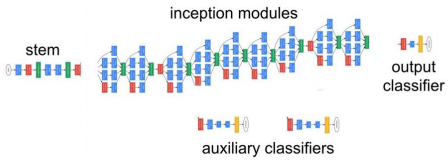
Features

All image data are converted into 2-D tensors of kernel size 3. During training, we also append each image with a class weight that will help the model penalize the loss. The reason we apply weights to the model is largely due to severe data imbalance.



Sample Data

Results



$$L_{cos}((\phi^R, \phi^v), y) = \begin{cases} 1 - cos(\phi^R, \phi^v), & \text{if } y = 1. \\ max(0, cos(\phi^R, \phi^v)) - \alpha, & \text{if } y = -1. \end{cases}$$

$$Loss_{total} = 1/M \sum_{i=0}^{m} \alpha \times \text{cross\_entropy\_loss}(x_m)$$

Model Performances

Architecture	Train Acc.	Test Acc.
Inceptionv3	77.4%	64.0%
MobileNet	80.1%	60.9%
NasNet	73.3%	51.5%

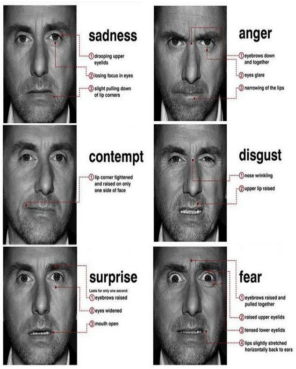
Model

1. Inception V3:
We experimented with various architectural modifications including adding more intermediate loss functions and more fully-connected layers.[2]

2. MobileNet:
We picked this model because it has much fewer number of parameters and thus more efficient in training stage.[3]

3. NasNet:
It has the state-of-art performance, and we use this model as a benchmark of other models' overall performance.[4]

The following table illustrate each model's best performance recorded on the same dataset.



Discussion

After this project, we learned that our model uses very similar approach to understand facial emotions as humans do. We both rely on the most contrasting features to perceive other people's emotions of the face, such as wide-open mouths, frowned eyebrows etc. In terms of results, our model doesn't achieve the state-of-art accuracies, but it shows new, promising characteristics such as learning new features, nuanced semantic representations.

Future

If time permits, we will spend more time gathering new data to enrich our collection and alleviate the data imbalance issue. We also hope to train a GAN network to help our discriminative model learn better.

[1] Dataset: https://github.com/muxspace/facial_expressions
[2] Shin, Hoo-Chang, Roth, Holger R, Gao, Mingchen, Lu, Le, Xu, Ziyue, Nogues, Isabella, Yao, Jianhua, Mollura, Daniel & Summers, Ronald M (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE
[3] Howard, Andrew G, Zhu, Menglong, Chen, Bo, Kalenichenko, Dmitry, Wang, Weijun, Weyand, Tobias, Andreetto, Marco & Adam, Hartwig (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications
[4] Zoph, Barret, Vasudevan, Vijay, Shlens, Jonathon & Le, Quoc V (2017). Learning transferable architectures for scalable image recognition.
[5] Scheller, Elisa, Bi"uchel, Christian & Gamer, Matthias (2012). Diagnostic features of emotional expressions are processed preferentially.