Facial Emotional Recognition Research for Videos Sentiment Analysis from Usability Tests

Prepared By: Basma Elhoseny

Contents

Article Approach

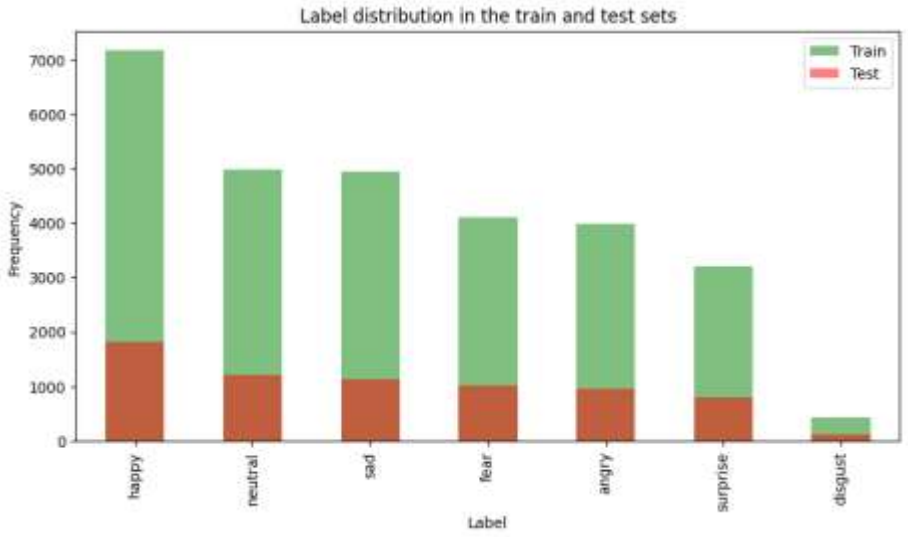
Dataset:

The Data Set used by the article has 35.9K .jpg images with 7 classes (`angry, disgust, fear, happy, neutral, sad, surprise`)

I have run a simple analysis on this data and saw that the disgust label has a very small no. of examples this means the data is biased so we need balancing techniques to tackle this problem. [Colab Notebook]

The techniques to be used [My Suggestions to ideas to be used with the original article approach]:

- ✓ Data augmentation for the examples from this class
  - Flipping
  - Random Rotation
  - Addition of Random Gaussian Noise
- ✓ Weighted loss function to be used to penalize misclassifications of the minority class more than the majority class during training.



The input is image of size 48*48.

Network:

My implementation for the article approach [Colab Notebook]

The article suggests using CNN network with 4 Conv2D layers as features extraction with Max-Pooling in between to reduce feature map and drop out layers to prevent overfitting.

This backbone is followed by 2 fully connected layers with relu activations and a final SoftMax layer with output 7 labels. That represents the probability of the image being classified as the corresponding label.

Optimizer: Adam Optimizer is used

Loss Function: Unweighted Categorical cross entropy. I suggest using weighted one with weights to be calculated form the frequency of the labels obtained in the analysis stop for the data set so that the model doesn't become biased to certain label,

Training:

The article suggests training on the dataset for 100 epoch with batch size 128.It used the test data set as validation. The training took 20.41 minutes on Colab free plan with a 12GB GPU, which is good for such a problem.

Evaluating:

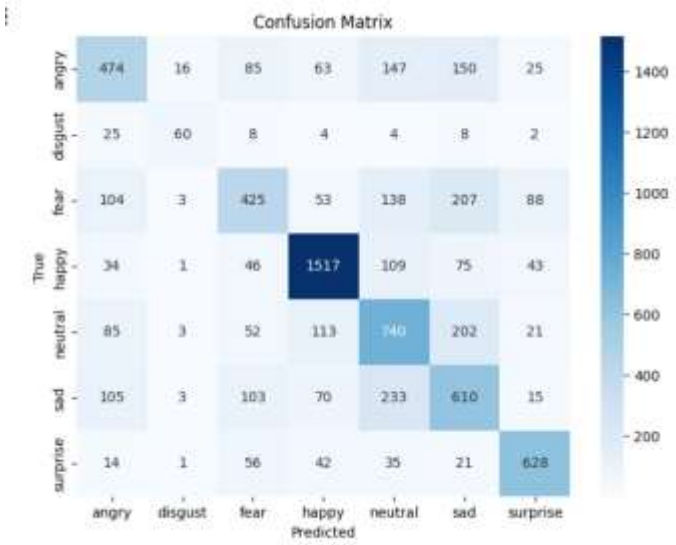


Figure 1 Test Data Set Confusion matrix.

Notes:

- Disgust has lowest Value this is the data analysis above shows that label **Disgust** has smallest no of examples in both training and test.
- This test data set is used for validation, so these results have some bias.

Real Time Detection:

The article used code for real time detection which I think will just be a useless constraint because we just want the analysis of the unmoderated test so no need for real time performance 😀.

Coding References for similar architecture:

- ✓ https://github.com/mayurmadnani/fer/blob/master/FER_CNN.ipynb
- ✓ https://www.kaggle.com/code/nakulsingh1289/face-expression-detection-from-scratch

CNNs for FER Problem:

In this section I will present other ideas or enhancements to the approach followed by the article in either literature or public available repositories.

Similar Papers

Paper [1] proposes an implementation of a general CNN building framework for designing real-time CNNs. The Problem of not achieving real time performance is the complexity of the network. The authors proposed 2 models, the first is sequential fully-CNN. Which is composed of 9 Conv Layers with Relu Activations. The FC is removed and replaced by Average Pooling. The model has 600K parameters. It is validated on the same dataset used by the article (FER-2013). Second, uses depth-wise separable convolutions instead of standard convolutions. By this they were able to achieve an arch with about 60K parameters. They were able to achieve inference time 0.22 ± 0.0003 ms including task of emotion classification and gender classification using approach based on

The official implementation is https://github.com/oarriaga/face_classification

Auxiliary Part: Guided back-propagation uncovers the dynamics of the weight changes and evaluates the learned features.
Challenges: Wearing of classes causes model inferencing with the features learned from the image by the model.
Obtain: FER2013 test accuracy of 75.2%
We will use https://github.com/amineHorseman/facial-expression-recognition-using-cnn as coding reference.

Resent18 & VGG19 Implementation [With Model Weights :D]

The authors of https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch has achieved 71.5% occur on the publicly available test and 73.112% on the private test set [in FER 2013 challenge]

Model	Public Test acc	Private Test acc
VGG19	71.496%	73.112%
Resnet18	71.190%	72.973%

In addition, they have shared their Final trained model trained on the EFR2013 https://drive.google.com/open?id=1Oy_9YmpkSKX1Q8jkOhJbz3Mc7qjyISzU

Pretrained Models

Since the dataset for NEF2013 is very popular and used as bench making between different architectures, I found weights for some models trained on this dataset.

Deep Face

- ✓ Facial attribute analysis with deep learning using the DeepFace Library
  - https://viso.ai/computer-vision/deepface/
  - https://pypi.org/project/deepface/0.0.61/
  - https://www.youtube.com/watch?v=n84hBgtzvxo
- ✓ https://www.kaggle.com/code/vumerenko/fer-pre-trained-vggface-fine-tune
- ✓ https://www.robots.ox.ac.uk/~albanie/pytorch-models.html [Emotion Recognition Section]
- ✓ https://www.vlfeat.org/sandbox-matconvnet/pretrained/

FER+ by Microsoft

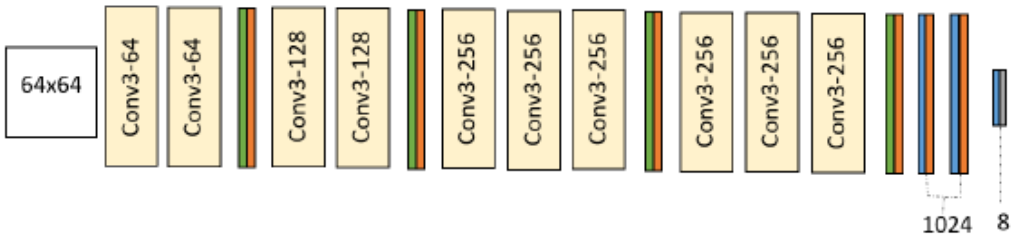
The original FER data set was web crawling face images with emotion related keywords. The images are filtered by human labelers, but the label accuracy is not very high. Microsoft Research suggests in [2] using crowd sourcing to collect ground truth labels (Annotation schema), but the labels resulting from crowd sourcing are noisy, so they propose how to make DCNN (Deep Conv Neural Networks) learn from Noisy Labels.

The new FER + is the same FER but re-tagged by 10 taggers to label the image into one of **8 emotion types** (neutral-happiness-surprise-sadness-anger-disgust-fear-contempt) + **unknown** and **NF** (Not a Face) Label. FER+ is an enhanced annotation and better generalization than the original FER.

The official repo https://github.com/microsoft/FERPlus/blob/master/README.md can be referenced for implementation and coding details as reference.

Approach

Network Proposed is Custom VGG-13 with 10 Conv Layers with Max Pooling and Dropout Layers for feature extraction backbone and for the classification they use 2 Dense layers followed by a SoftMax.



The main idea of the paper is to deal with such noisy labels, the image has votes for its label which reflects the real world since emotions are very subjective, and it is very common that two people have diametrically different opinions on the same face image. So, 4 schemas are proposed to deal with such Noisy labels.

1. Majority Voting
    - The label to be learned is the one with max votes by the taggers.
2. Multi-Label Learning
    - It is Fine for the learning algorithm to match with any of the emotions that has enough taggers labeling them.
    - The believed that this approach is to be the best, but it didn't show the best results [Results are shown in results section below]
3. Probabilistic Label Drawing
    - A random emotion tag is drawn from the example's label distribution.
4. Cross-entropy loss
    - The Label distribution is the target function to be learned by the model.
    - Over the multiple epochs during training, we will approach the true label distribution on average.

Results:

Note: Most of the emotions are well classified except disgust and contempt. This is because we have very few examples in the FER+ training set that are labeled with these two emotions.

	Neutral	Happiness	Surprise	Sadness	Anger	Disgust	Fear	Contempt
Neutral	90.27%	1.91%	1.48%	4.95%	1.13%	0.00%	0.26%	0.00%
Happiness	2.32%	94.47%	1.22%	1.22%	0.77%	0.00%	0.00%	0.00%
Surprise	6.64%	3.08%	86.97%	0.71%	1.18%	0.00%	1.42%	0.00%
Sadness	23.21%	1.67%	0.72%	67.94%	3.59%	0.48%	2.39%	0.00%
Anger	10.16%	3.28%	0.66%	2.30%	82.30%	0.66%	0.66%	0.00%
Disgust	10.53%	0.00%	5.26%	0.00%	57.89%	26.32%	0.00%	0.00%
Fear	4.35%	0.00%	29.35%	8.70%	5.43%	0.00%	52.17%	0.00%
Contempt	54.17%	0.00%	0.00%	12.50%	20.83%	4.17%	4.17%	4.17%

Figure 2 Confusion Matrix for PLD (Probabilistic Label Drawing) Schema

Note: As shown in table 1 PLD (Probabilistic Label Drawing) and CEL (Cross-entropy loss) proved to be the best approaches to deal with the noisy labels.

Scheme	Trials					Accuracy
	1	2	3	4	5	
MV	83.60 %	84.89 %	83.15 %	83.39 %	84.23 %	83.852 ± 0.631 %
ML	83.69 %	83.63 %	83.81 %	84.62 %	84.08 %	83.966 ± 0.362 %
PLD	85.43 %	84.65 %	85.34 %	85.01 %	84.50 %	**84.986 ± 0.366** %
CEL	85.01 %	84.59 %	84.32 %	84.80 %	84.86 %	84.716 ± 0.239 %

Table 1 Testing accuracy from training VGG13 using four different schemes: majority voting (MV), multi-label learning (ML), probabilistic label drawing (PLD) and cross-entropy loss (CEL)

Multi Model Approach:

I had an idea for a fancy approach :D. I see we have multiple data forms audio, video, textual input that we can analyze for sentiment classification. So, building a multi-model we can make use of these different input forms. We can use this model for analyzing both written answers for the test, video answers being developed by Juca, and furthermore the voice recorded answers. I have done an initial search on this idea, if this idea is interesting for the mentors more detailed research will be done on the available approaches and the pretrained models.

Paper [3] proposes use of speech from video and its transcript to predict the sentiment analysis.

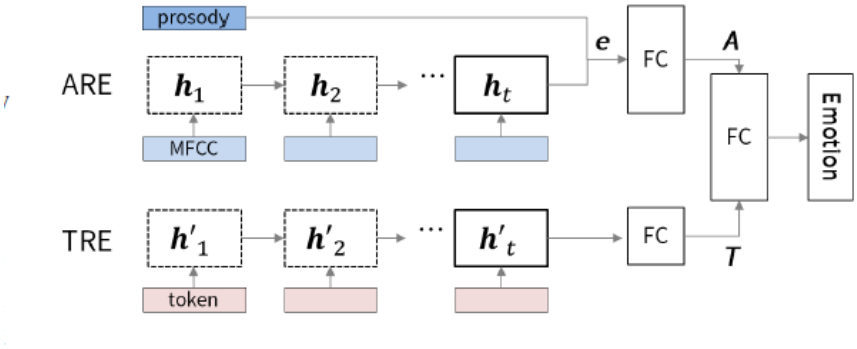


Figure 3 Multimodal dual recurrent encoder. The upper part

This paper used Dataset IEMOCAP https://sail.usc.edu/iemocap/index.html, which is a multimodal and multi-speaker database, recently collected at SAIL lab at USC. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions. This data set is available only for research purposes. **But I have contacted the owners of the dataset, and they granted me the data, I got from them the approval about using the data for open-source project on condition that we share the final weights.**

Terms and Conditions: https://sail.usc.edu/iemocap/Data_Release_Form_IEMOCAP.pdf

Note: The Dataset has the transcript for the videos. For Inference we can used google API Speech to Text

References

[1] Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.

[2] Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016, October). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 279-283).

[3] Yoon, S., Byun, S., & Jung, K. (2018, December). Multimodal speech emotion recognition using audio and text. In 2018 IEEE spoken language technology workshop (SLT) (pp. 112-118). IEEE.