Emotion Detection with Vision Transformers and Image Features

Stephan Sharkov
stpshrkv@stanford.edu | CS231N Spring 2022

Background

- **Emotion Detection** is an important CV problem, widely used in healthcare and HCI.
- Models are given an image and try to **predict the emotion** person is having in the picture, or **emotion sentiment(pos. vs. neg.)**
- **Transformers** having success in NLP, have been widely explored in CV
- **State-of-the-art models** attempted to solve the problem with CNN's, Finetuning on networks, as well as Finetuning with Transformers
- I attempted to improve using **Transformer and HOG**

Problems

- How do we achieve the best accuracy in **detecting someone's emotion, or at least its sentiment**?

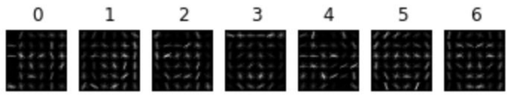
Dataset

- FER2013[1] Dataset uses 7 emotions: **Anger(0), Disgust(1), Fear(2), Happy(3), Sad(4), Surprise(5), Neutral(6).**

Examples of original images for each emotion

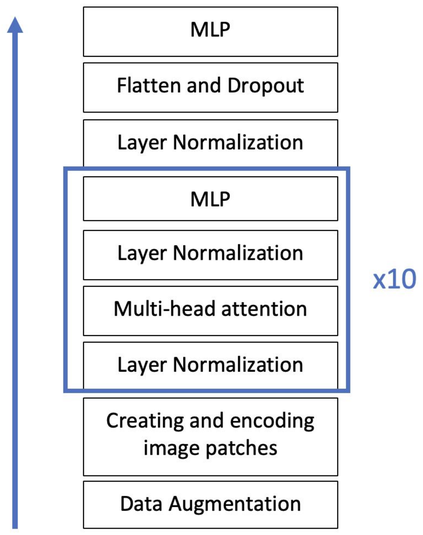
- Training data: 22968 images
  Validation data: 5741 images
  Testing data: 3589 images
- Sentiment analysis: **0-Negative**(Anger, Disgust, Fear, Sad), **1-Neutral-Positive**(Happy, Surprise, Neutral)
- HOG - Histogram of Oriented Gradients, represents direction and intensity of gradients at every pixel of the image, indicating shape

Examples of HOG features for each emotion



Methods

- **Transformer**:
  - Augments Data with Resizing, Random Rotation, Random Zoom
  - Splits images into 144 patches, encodes positional embeddings
  - MLP includes 2 layers of Dense layer with ReLU and Dropout



- **Finetuning**
  - Performed on top of ResNet101 pre-trained on ImageNet data
  - MLP with 4 layers of Dense layer with ReLU and Dropout

- Baseline Models: **KNN and CNN**
  - CNN with Batch-Norm, Conv-32, ReLU, Conv-16, ReLU, MaxPool

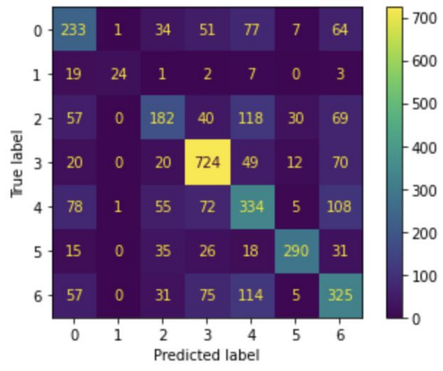
Discussion

- **Finetune overfits**, even with high regularisation (Dropout) and low epochs
- **HOG features are not working with Transformers**, since pictures become too similar
- **HOG helps Finetune on sentiment analysis**, indicating that it can improve existing Finetune models, but performs bad in emotion classification
- **Finetune with ResNet101** in the base performed **better on emotion classification** than other pre-trained models in the base
- **Finetune with ResNet101** in the base performed **worse on sentiment analysis** than other pre-trained models in the base

Experiments

Model	EC ACC	SA ACC
KNN(k=60)	16	49
KNN-HOG(k=50)	35.5	60.5
CNN	47.15	70.27
CNN-HOG	46.93	69.53
Finetune[2]	41.34	73.19
Finetune[3]	45.16	**73.67**
Transformer	**58.51**	71.08
Finetune	48.26	68.86
Finetune-HOG	38.03	71.45

Correctly Classified Image: Disgust



Incorrectly Classified Image: Fear instead of Anger

Future Work

- Investigating and decreasing **overfitting of Finetune method**
- Exploring **other image features** besides HOG
  - Which ones could work with Transformers
- Using **Transformers as Finetune** method on top of pre-trained models

References

1. Ian Goodfellow et al. FER2013 Dataset: "Challenges in Representation Learning: A report on three machine learning contests." On arXiv, 2013.
2. Vasavi Gajarla and Aditi Gupta. "Emotion Detection and Sentiment Analysis of Images." In Georgia Institute of Technology. 2015.
3. Fuyan Ma, Bin Sun, and Shutao Li. "Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion." In Affective Comput. 2021.