# FACIAL EXPRESSION RECOGNITION
# Classifying facial expressions using AlexNet, ResNet, and Vision Transformers

Peter Phan, Peter Nguyen, Duc Nguyen

UMassAmherst
College of Information
& Computer Sciences

## MOTIVATION

Our motivation is to classify facial expressions using different machine learning models. We are using the FER (Facial Expression Recognition) - 2013 dataset by Ian Goodfellow.

## AIM

**Classifying facial expressions using Convolutional Neural Networks and Vision Transformers**

## METHODS

**FERNet**: A CNN model with architecture inspired by AlexNet. Initial hyperparameters were chosen based on previous projects on Kaggle. A random hyperparameter search was deployed to improve the model's test accuracy. The tuned model is trained until convergence.

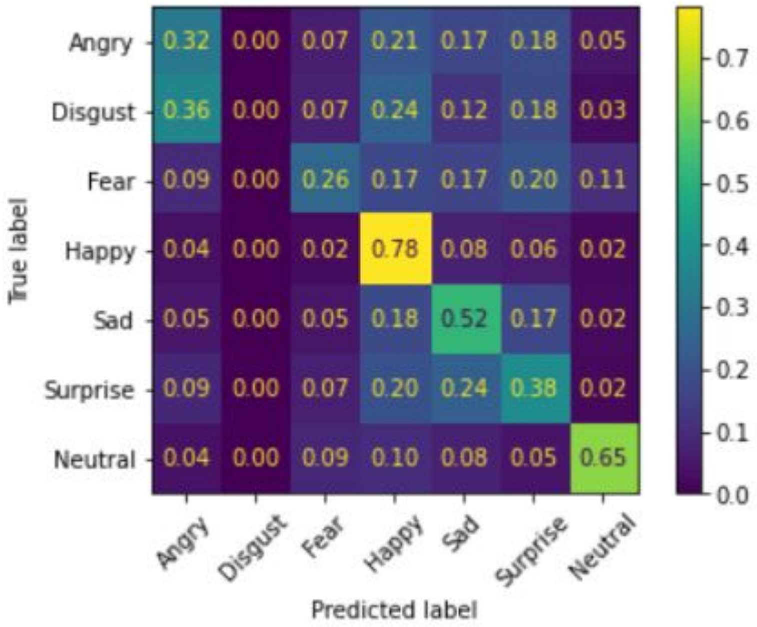
**ResNet:** A Residual Network model is a variant of CNN models. It incorporates the concept of residual learning which solves many performance problems seen in classic CNN models. The main difference in the implementation of a CNN model is the use of residual blocks. Instead of stacking many nonlinear layers on top of each other, we incorporate an underlying identity mapping in the ResNet model between our stacked layers. We essentially give the mapping a reference point which in theory is easier to train from than from an unreferenced mapping. We ran the training on the same training and testing dataset as our AlexNet model.

**ViT Image Classifier:** Our model was developed using the HuggingFace Transformers library. The FER dataset was prepared using the pretrained "vit-base-patch16-224-in21k" feature extractor which is pre-trained on ImageNet-21K at resolution 224x224. The training dataset was scaled down to one-sixth of the original dataset but was made sure to have the same distribution of labels. The same pretrained model was also utilized by our ViT image classification model. The model was trained until convergence.

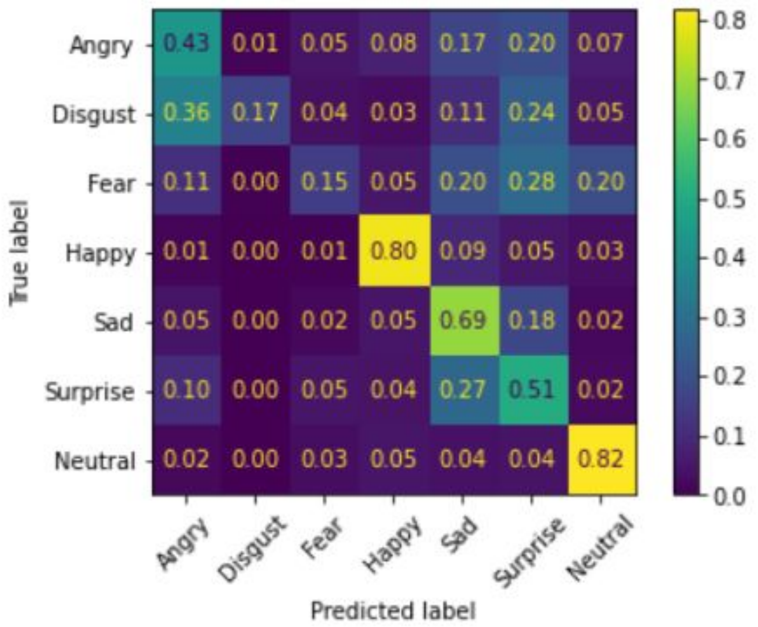
**Model Evaluation:** By keeping the dataset's characteristics consistent, we developed several tests to measure the performances of each model. These tests include the test accuracy, training-to-testing accuracy ratio, confusion matrix, and convergence rate. These scores were assessed to quantify the model's performance as well as identify any overfit.

**Literature Evaluation:** The results of the experiment are compared with results from published literature as well as from previous submissions on Kaggle. The process is carried out to verify the validity of our findings.

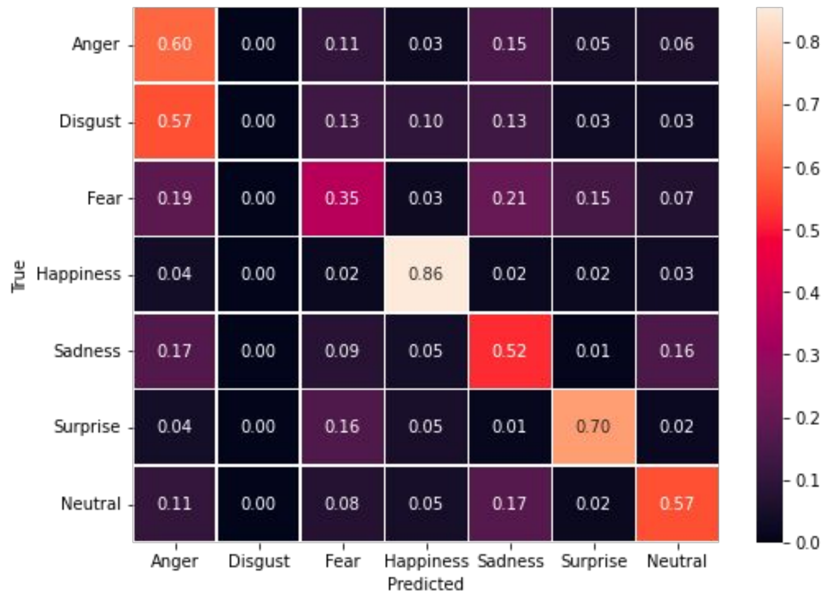
## RESULTS



Normalized confusion matrix for FERNet



Normalized confusion matrix for ResNet



Normalized confusion matrix for ViT

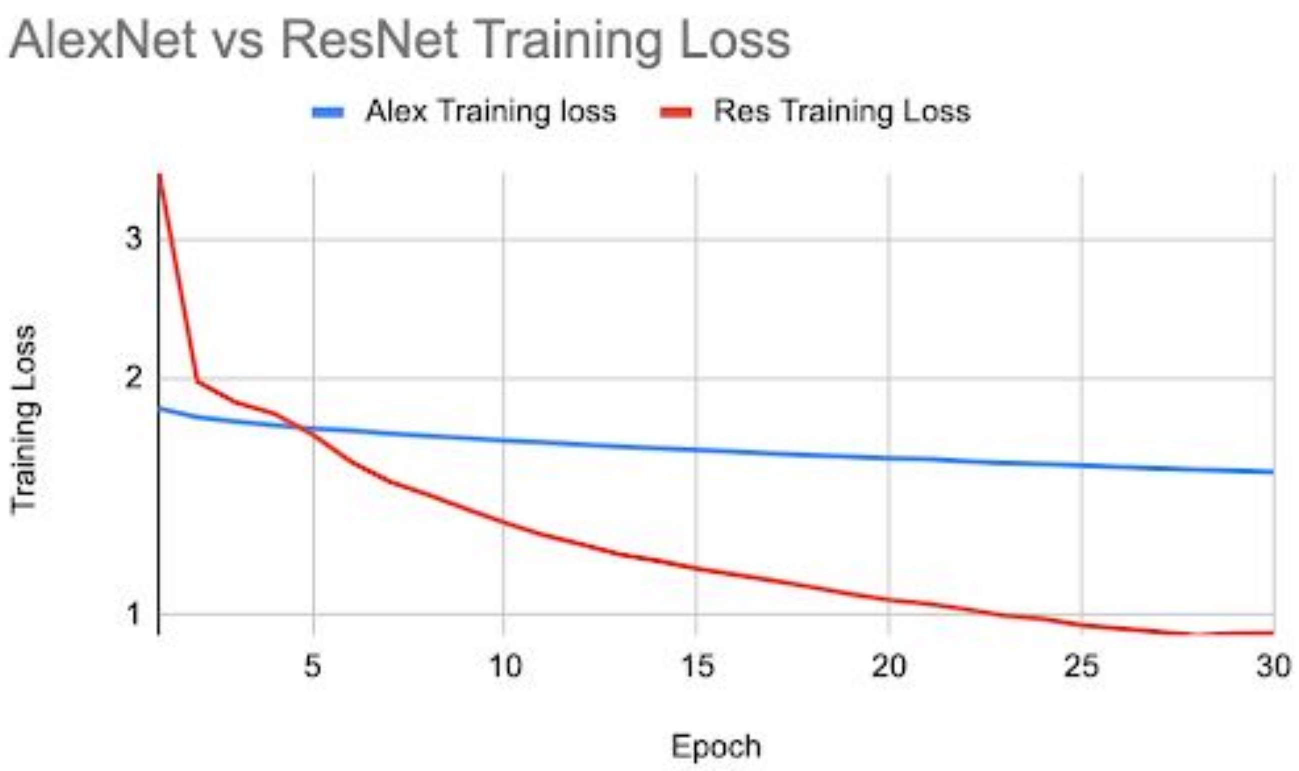
| | AlexNet (FERNet) | ResNet | ViT |
|---|---|---|---|
| Training time | ~13 hours | ~5 hours | 37 minutes (smaller dataset) |
| Number of parameters | 29,069,064 | 223,847 | 86,394,631 |
| Final test accuracy<br>(Human: 65±5%) | 50.4% | 60% | 61% |
| Final training accuracy | 72.4% | 63.8% | N/A |
| Number of epochs | 30 | 30 | 6 |

## CONCLUSION

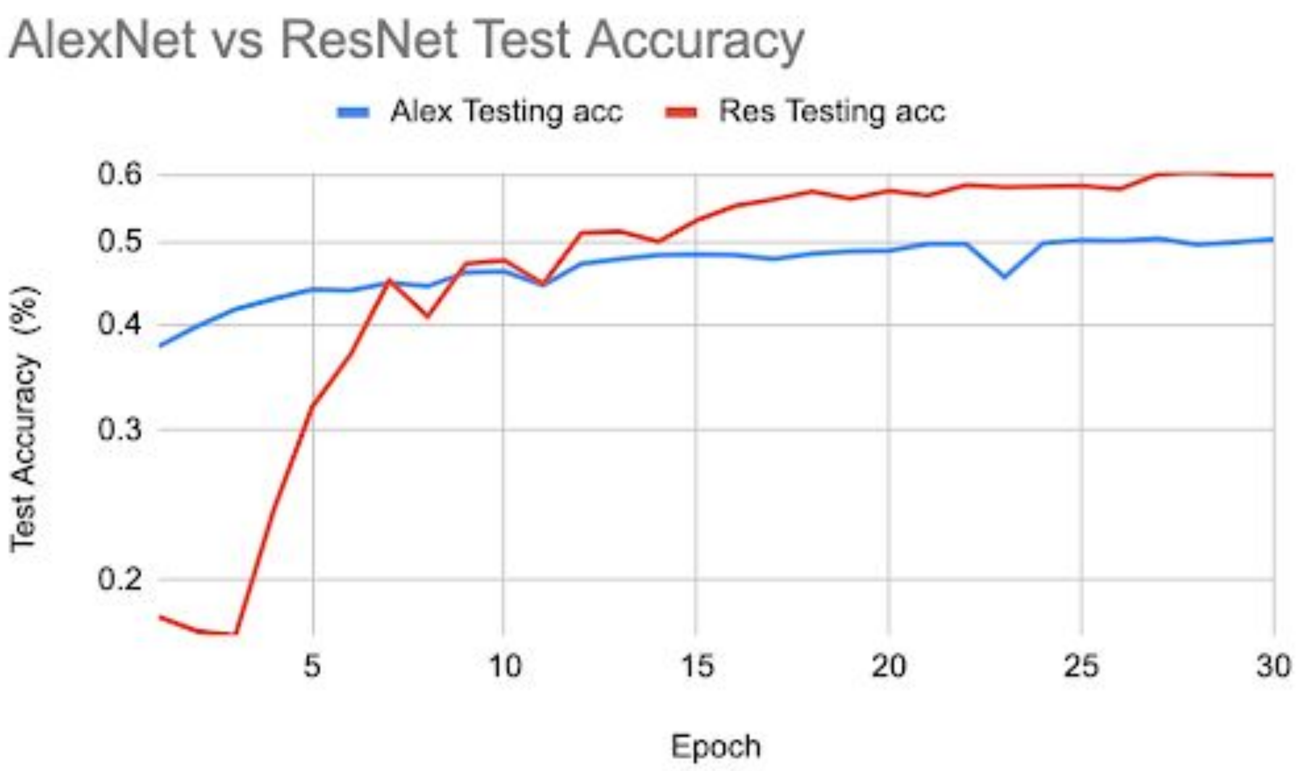
Both ResNet and ViT showed significantly higher test accuracy scores when compared with the older AlexNet model. These models, despite their implementation differences, both performed relatively well with test accuracy scores in the range of the human accuracy of 65.5+-5% on the FER2013 dataset. Notably, ResNet was able to correctly predict the "disgust" label which is impressive because of its substantially smaller sample size compared to other labels. On the other hand, the ViT model was able to achieve the highest test accuracy score despite being trained on a smaller training dataset.

## REFERENCES

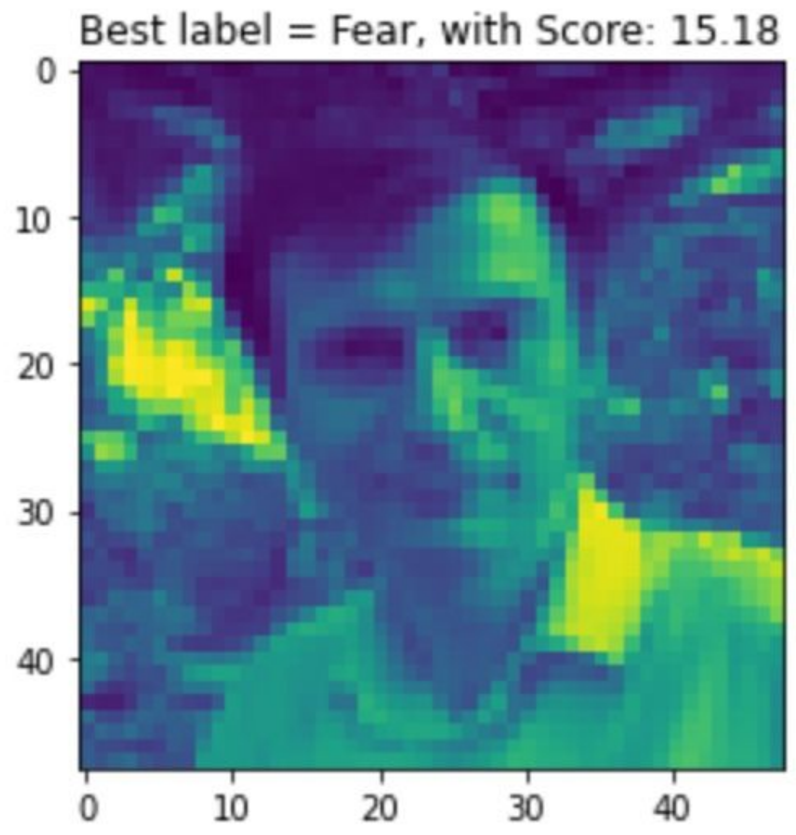
arXiv:1307.0414
arXiv:1512.03385
arXiv:2010.11929



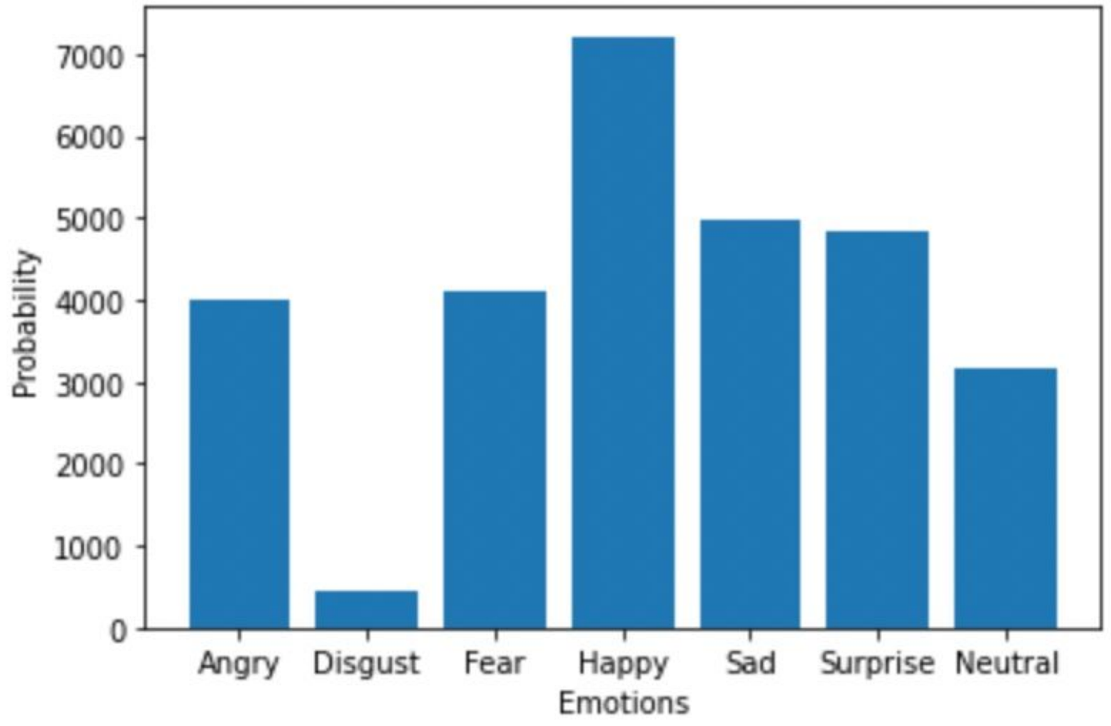
AlexNet vs ResNet Training Loss



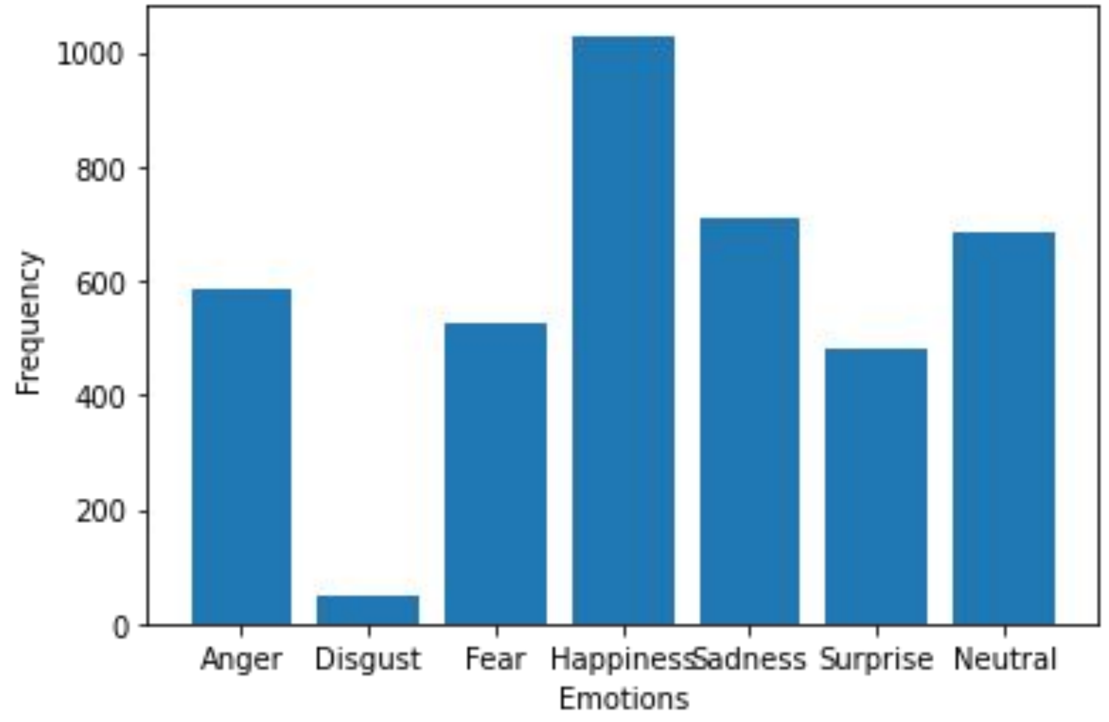
AlexNet vs ResNet Test Accuracy



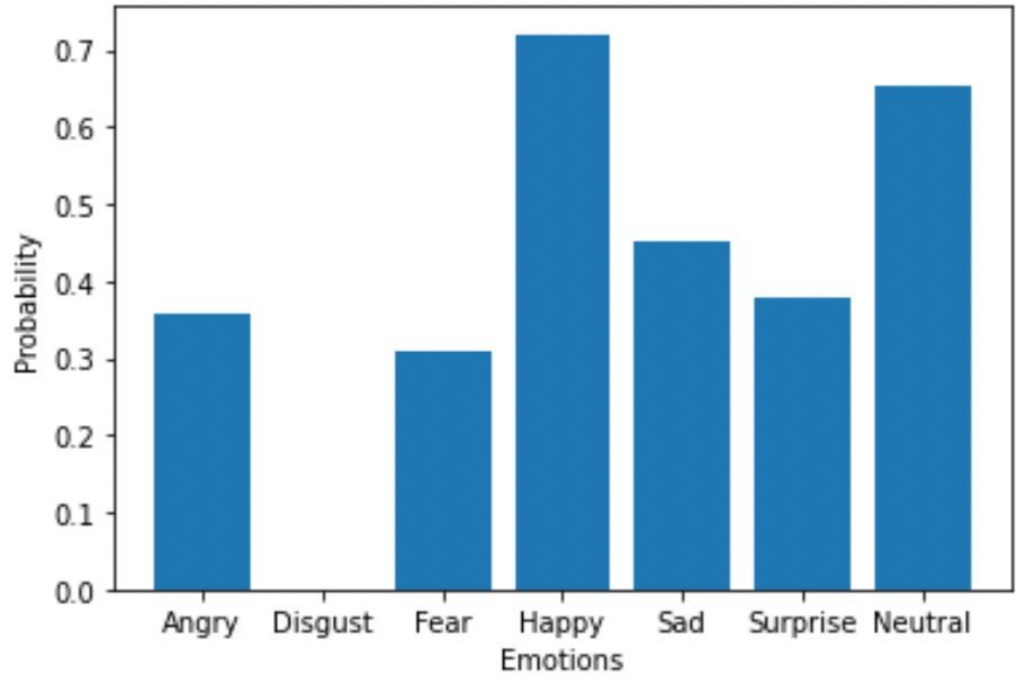
Big Boss



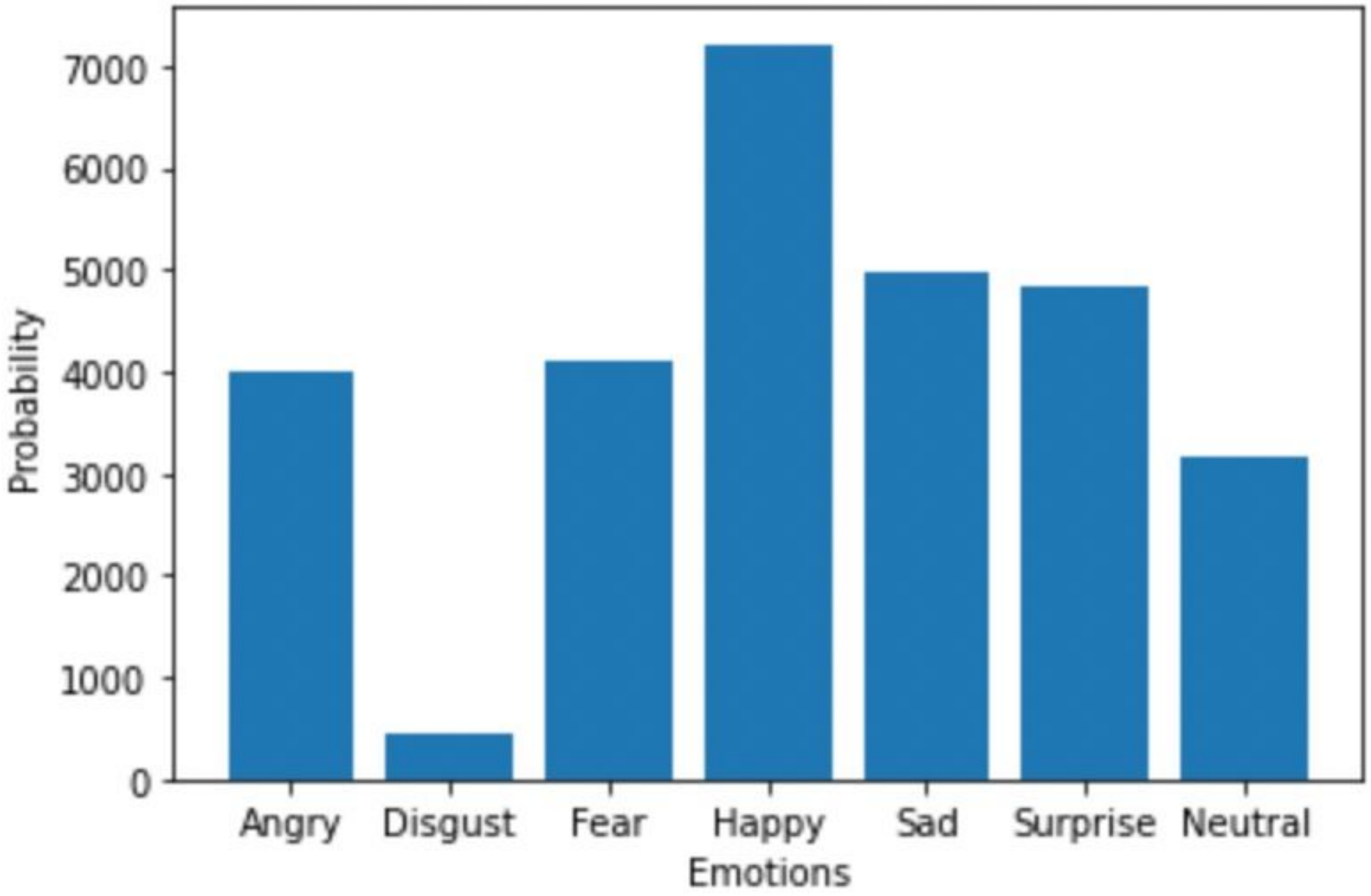
Full FER2013 Dataset



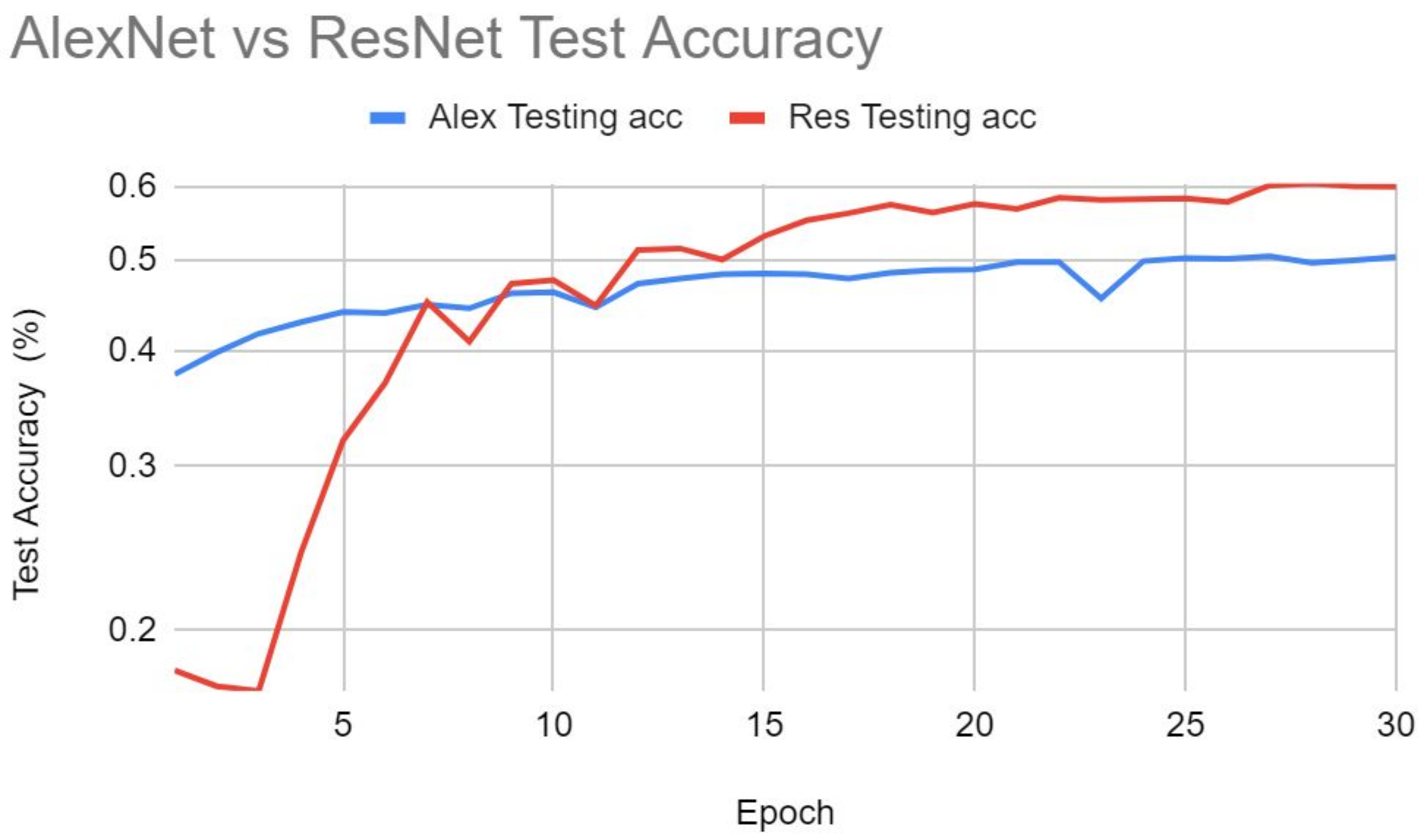
Scaled-down dataset used for ViT



Test accuracy on each label (AlexNet)

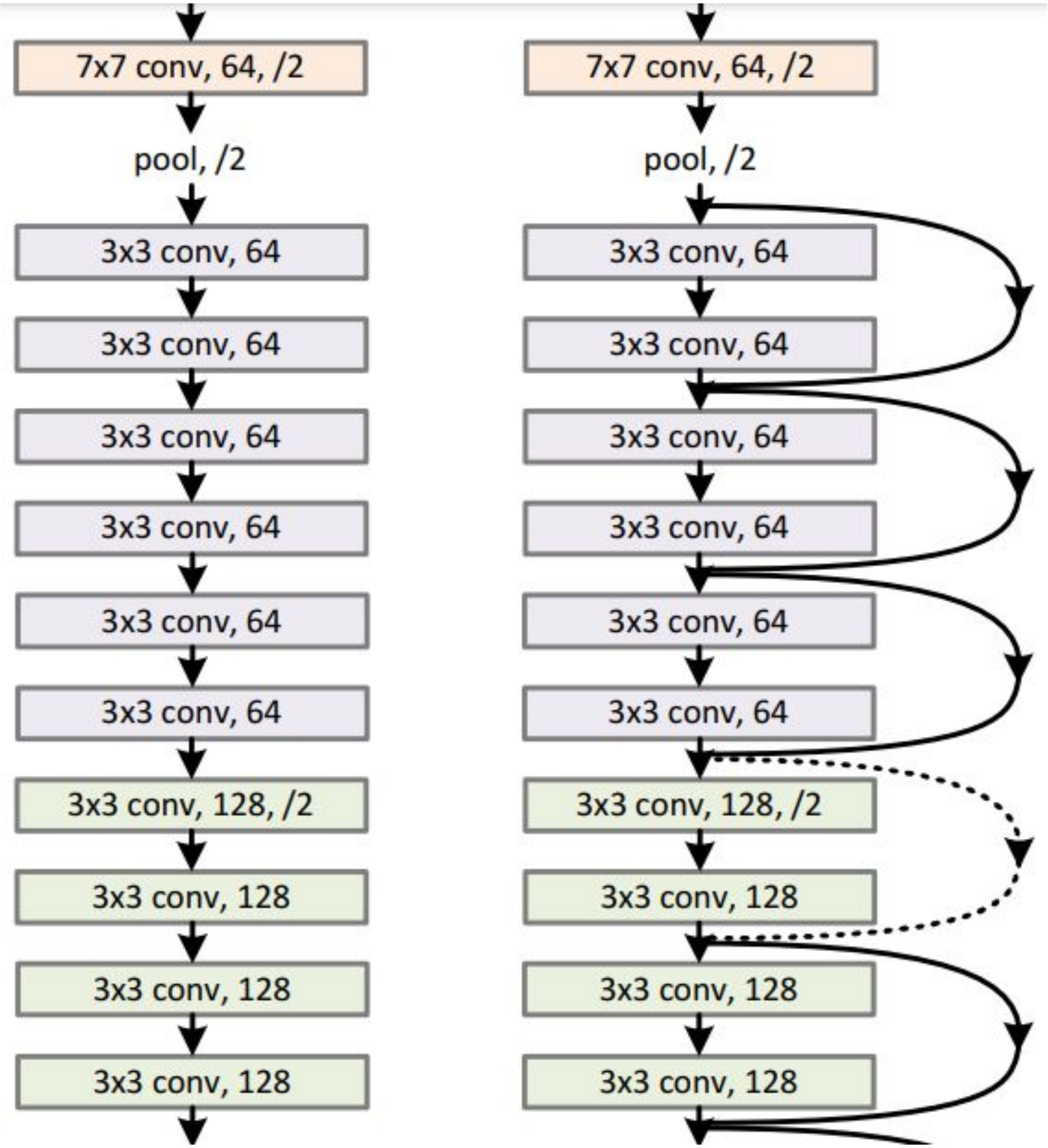


Full FER2013 Dataset

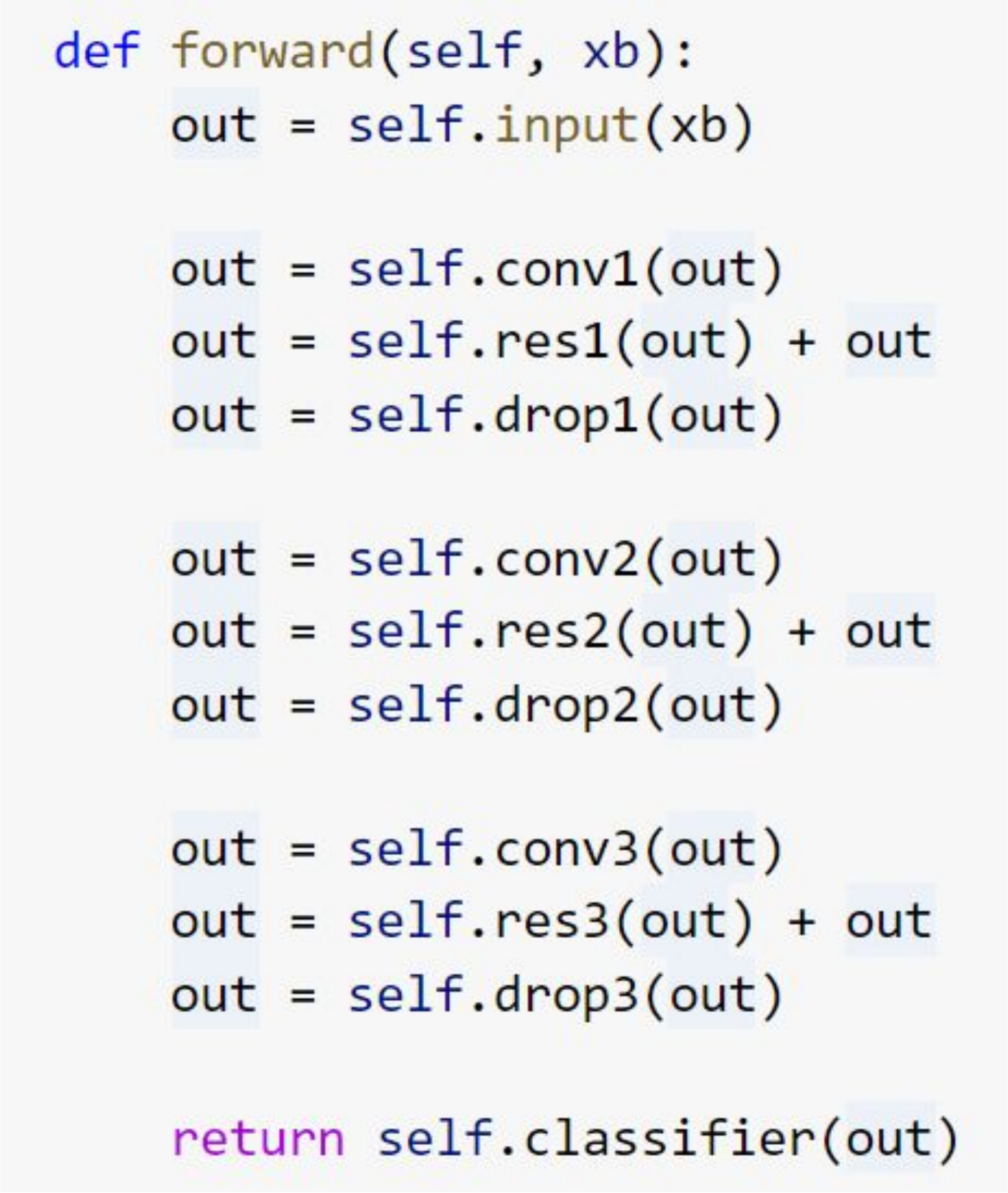

AlexNet vs ResNet Test Accuracy
Alex Testing acc
Res Testing acc
Test Accuracy (%)
0.6
0.5
0.4
0.3
0.2
5
10
15
20
25
30
Epoch

ImageNet 21K

FER2013

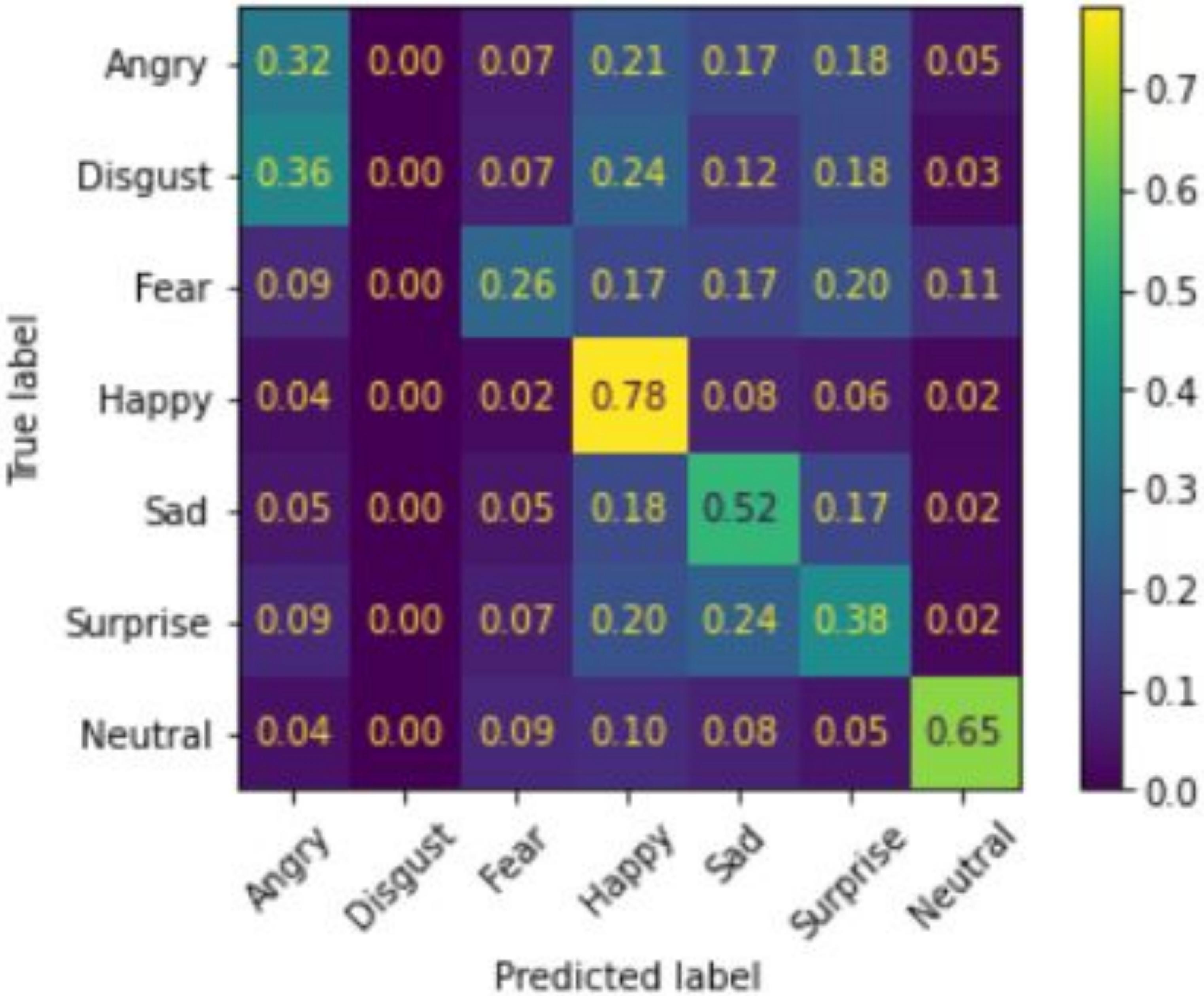


VGG-19 vs ResNet Comparison



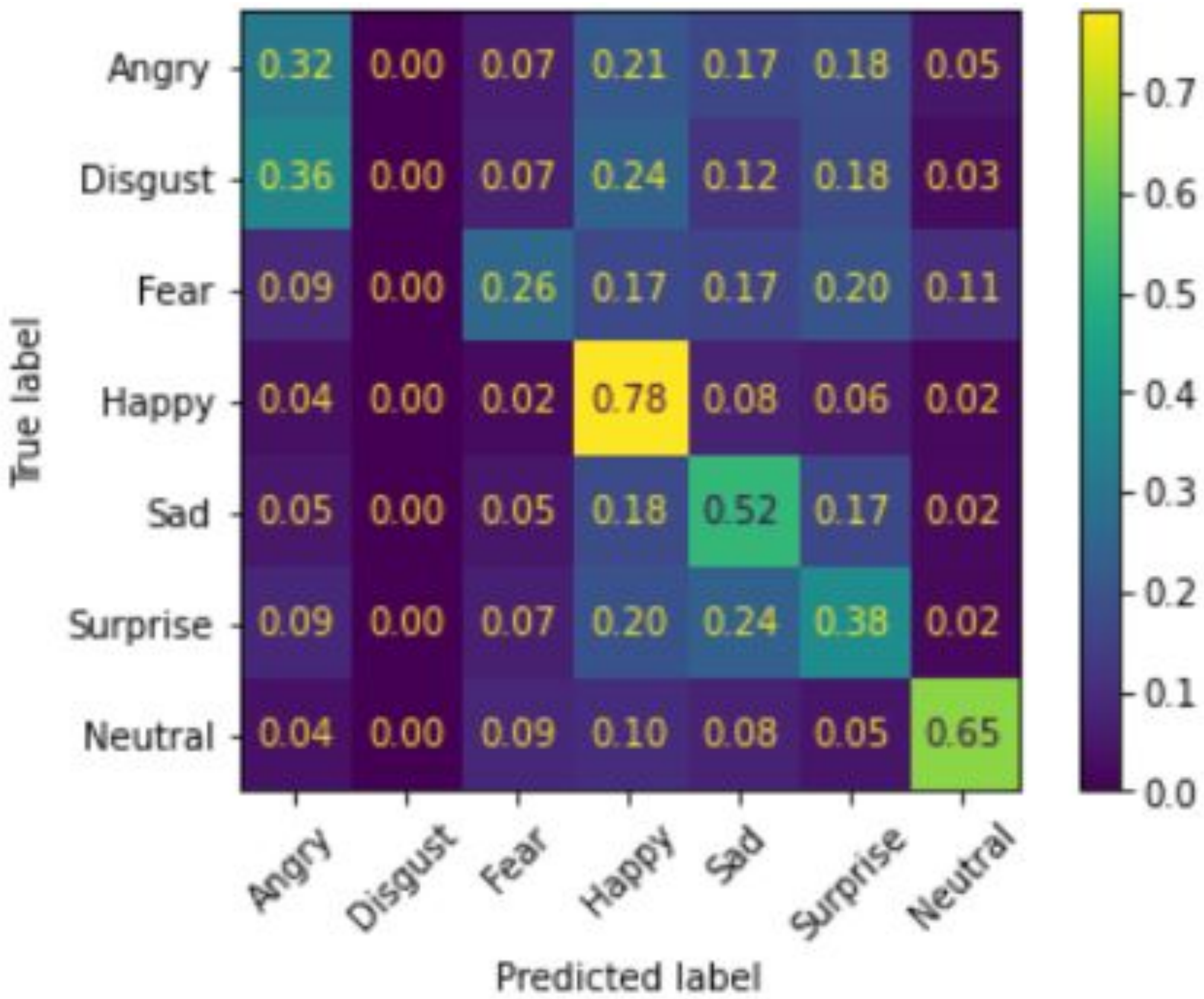
```python
def forward(self, xb):
    out = self.input(xb)

    out = self.conv1(out)
    out = self.res1(out) + out
    out = self.drop1(out)

    out = self.conv2(out)
    out = self.res2(out) + out
    out = self.drop2(out)

    out = self.conv3(out)
    out = self.res3(out) + out
    out = self.drop3(out)

    return self.classifier(out)
```

Code Structure



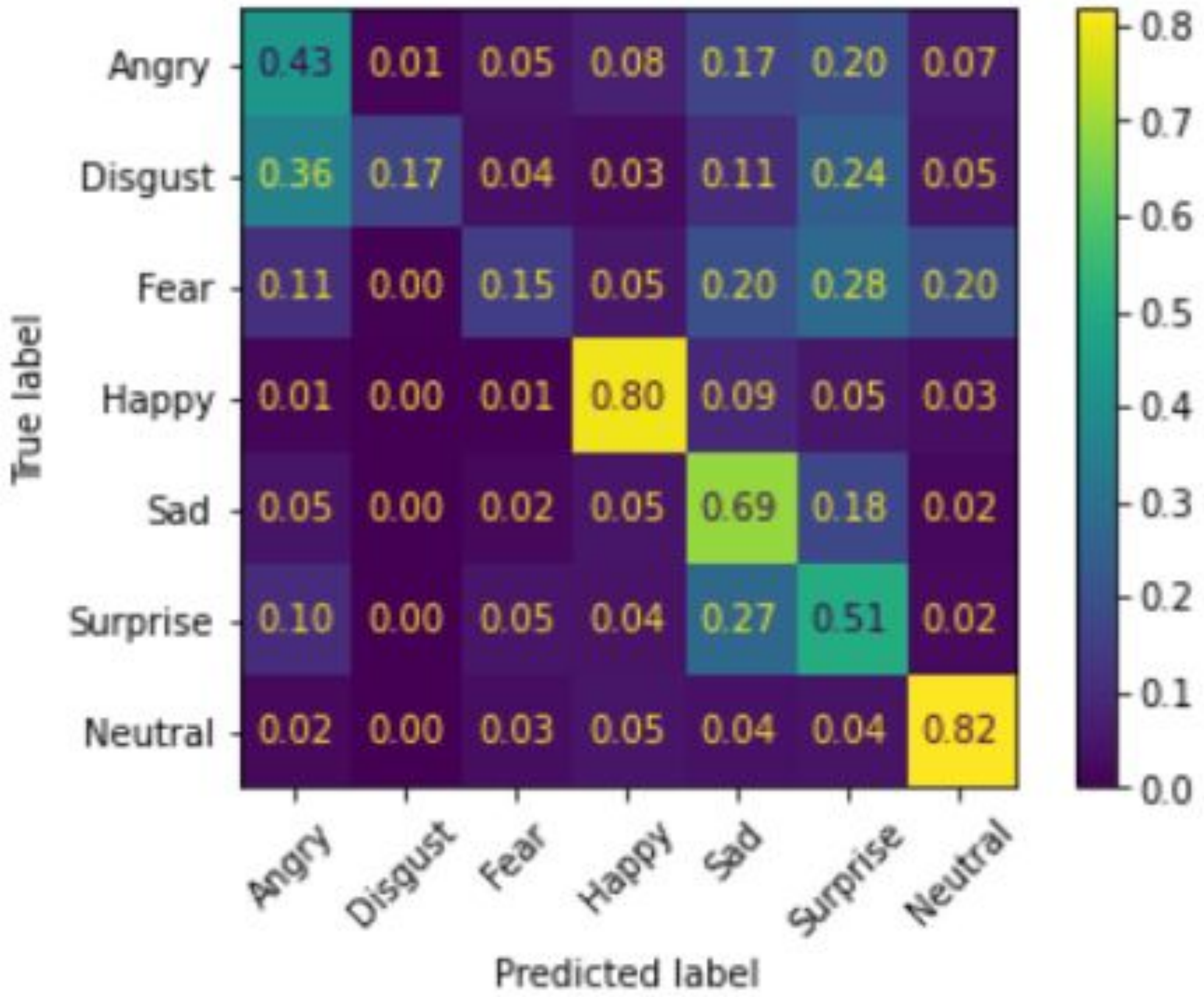
| True label \ Predicted label | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Angry | 0.32 | 0.00 | 0.07 | 0.21 | 0.17 | 0.18 | 0.05 |
| Disgust | 0.36 | 0.00 | 0.07 | 0.24 | 0.12 | 0.18 | 0.03 |
| Fear | 0.09 | 0.00 | 0.26 | 0.17 | 0.17 | 0.20 | 0.11 |
| Happy | 0.04 | 0.00 | 0.02 | 0.78 | 0.08 | 0.06 | 0.02 |
| Sad | 0.05 | 0.00 | 0.05 | 0.18 | 0.52 | 0.17 | 0.02 |
| Surprise | 0.09 | 0.00 | 0.07 | 0.20 | 0.24 | 0.38 | 0.02 |
| Neutral | 0.04 | 0.00 | 0.09 | 0.10 | 0.08 | 0.05 | 0.65 |

Normalized true-label confusion matrix for AlexNet



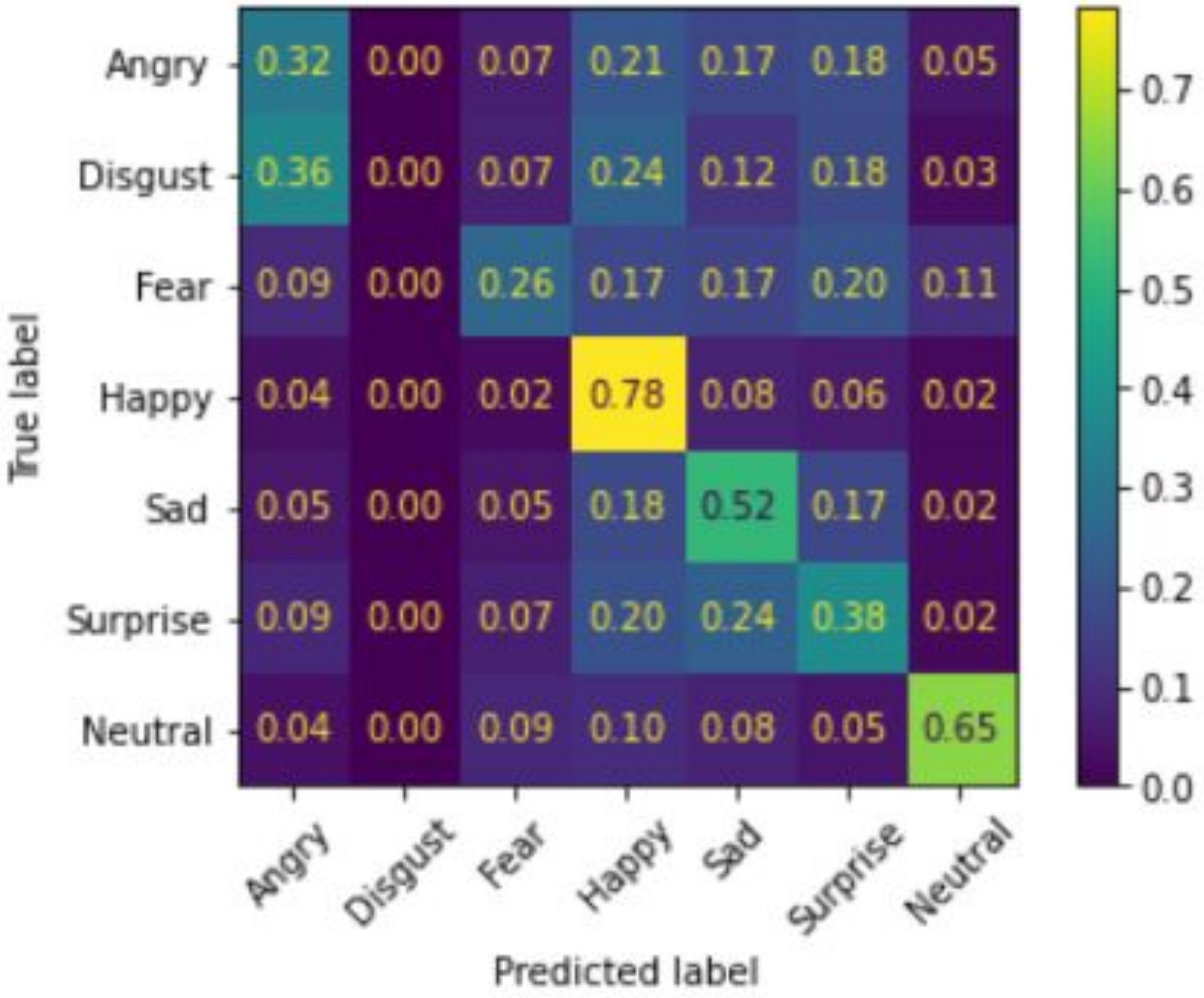
| | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Angry | 0.32 | 0.00 | 0.07 | 0.21 | 0.17 | 0.18 | 0.05 |
| Disgust | 0.36 | 0.00 | 0.07 | 0.24 | 0.12 | 0.18 | 0.03 |
| Fear | 0.09 | 0.00 | 0.26 | 0.17 | 0.17 | 0.20 | 0.11 |
| Happy | 0.04 | 0.00 | 0.02 | 0.78 | 0.08 | 0.06 | 0.02 |
| Sad | 0.05 | 0.00 | 0.05 | 0.18 | 0.52 | 0.17 | 0.02 |
| Surprise | 0.09 | 0.00 | 0.07 | 0.20 | 0.24 | 0.38 | 0.02 |
| Neutral | 0.04 | 0.00 | 0.09 | 0.10 | 0.08 | 0.05 | 0.65 |

Normalized confusion matrix for AlexNet



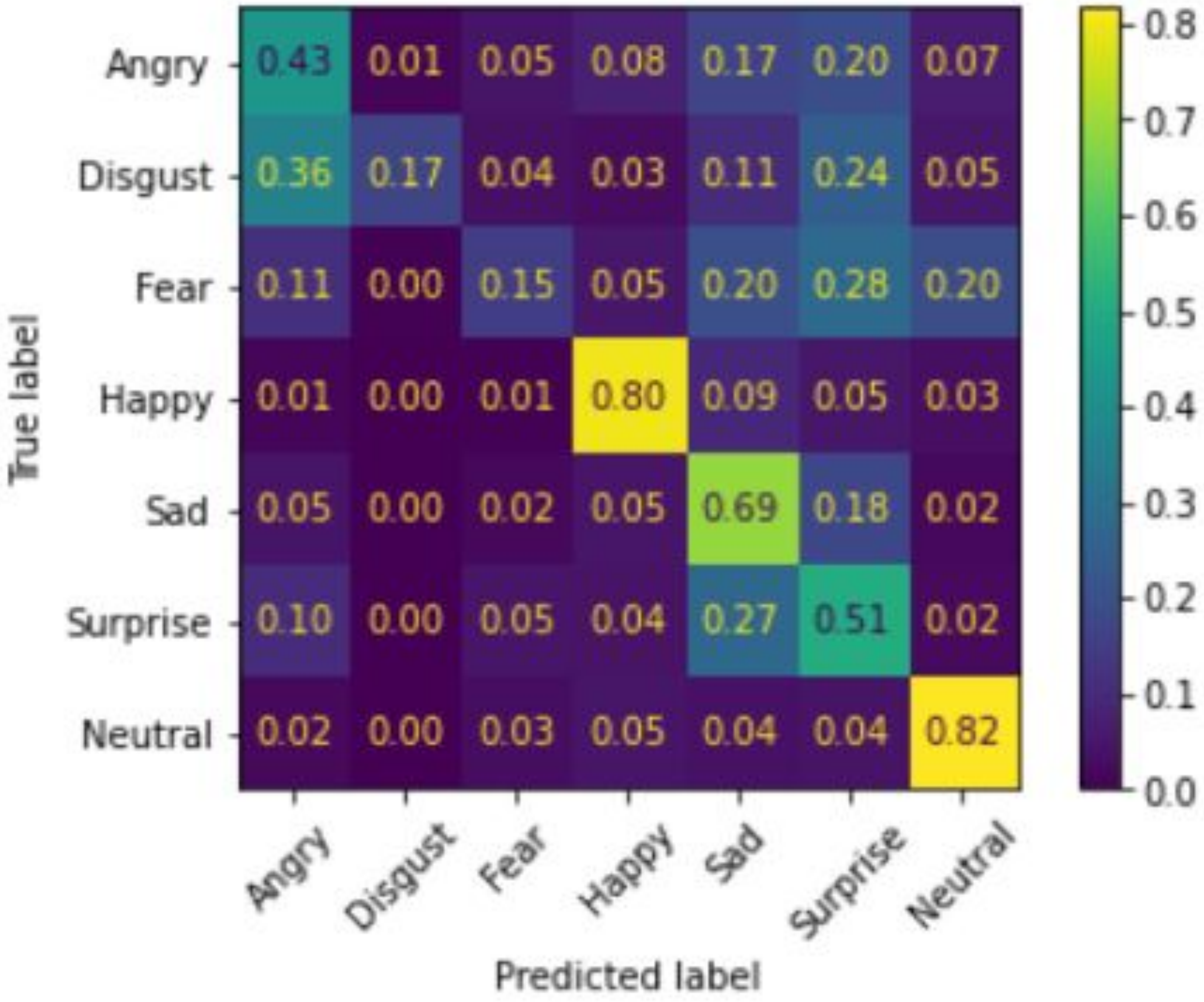
| | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Angry | 0.43 | 0.01 | 0.05 | 0.08 | 0.17 | 0.20 | 0.07 |
| Disgust | 0.36 | 0.17 | 0.04 | 0.03 | 0.11 | 0.24 | 0.05 |
| Fear | 0.11 | 0.00 | 0.15 | 0.05 | 0.20 | 0.28 | 0.20 |
| Happy | 0.01 | 0.00 | 0.01 | 0.80 | 0.09 | 0.05 | 0.03 |
| Sad | 0.05 | 0.00 | 0.02 | 0.05 | 0.69 | 0.18 | 0.02 |
| Surprise | 0.10 | 0.00 | 0.05 | 0.04 | 0.27 | 0.51 | 0.02 |
| Neutral | 0.02 | 0.00 | 0.03 | 0.05 | 0.04 | 0.04 | 0.82 |

Normalized confusion matrix for ResNet



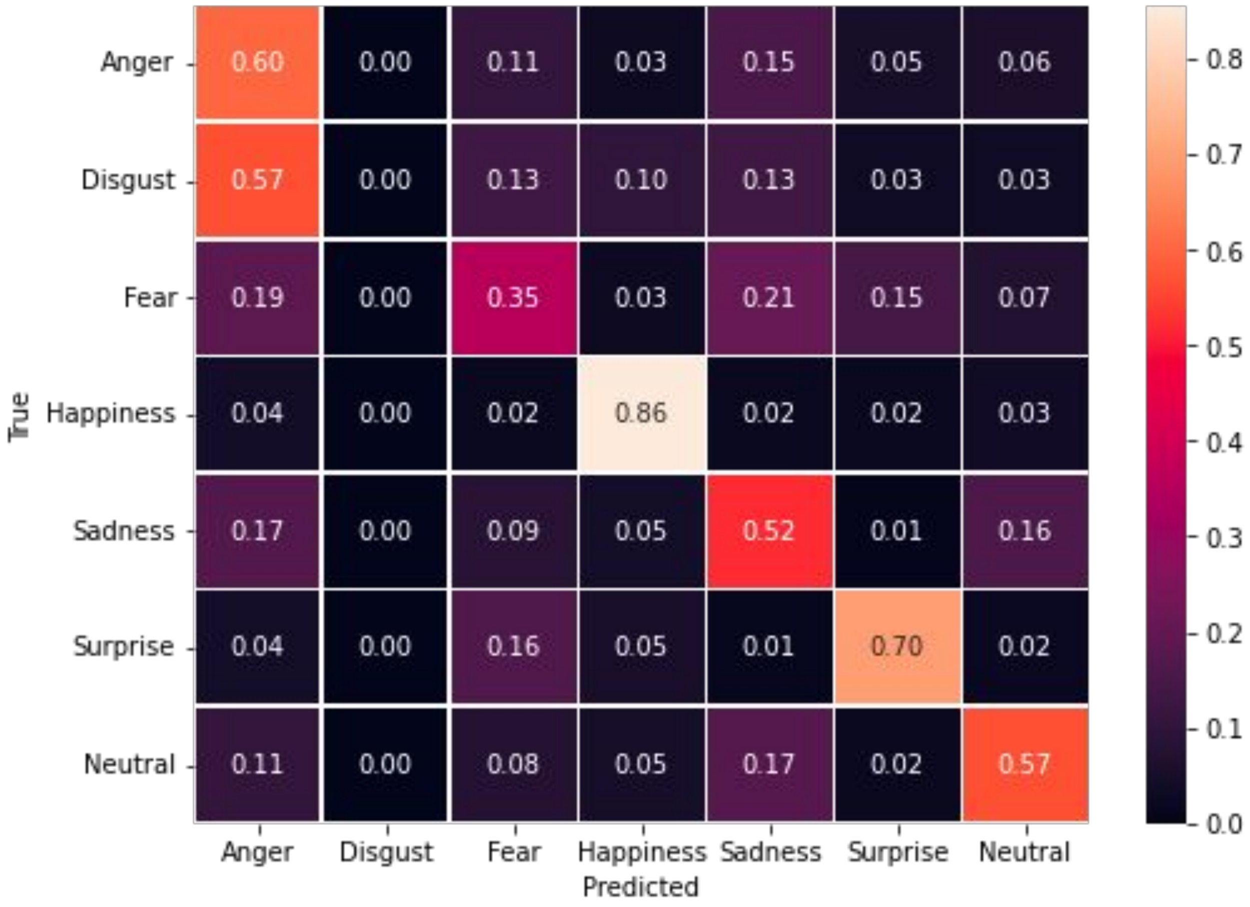
| | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Angry | 0.32 | 0.00 | 0.07 | 0.21 | 0.17 | 0.18 | 0.05 |
| Disgust | 0.36 | 0.00 | 0.07 | 0.24 | 0.12 | 0.18 | 0.03 |
| Fear | 0.09 | 0.00 | 0.26 | 0.17 | 0.17 | 0.20 | 0.11 |
| Happy | 0.04 | 0.00 | 0.02 | 0.78 | 0.08 | 0.06 | 0.02 |
| Sad | 0.05 | 0.00 | 0.05 | 0.18 | 0.52 | 0.17 | 0.02 |
| Surprise | 0.09 | 0.00 | 0.07 | 0.20 | 0.24 | 0.38 | 0.02 |
| Neutral | 0.04 | 0.00 | 0.09 | 0.10 | 0.08 | 0.05 | 0.65 |

Normalized confusion matrix for AlexNet



| | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Angry | 0.43 | 0.01 | 0.05 | 0.08 | 0.17 | 0.20 | 0.07 |
| Disgust | 0.36 | 0.17 | 0.04 | 0.03 | 0.11 | 0.24 | 0.05 |
| Fear | 0.11 | 0.00 | 0.15 | 0.05 | 0.20 | 0.28 | 0.20 |
| Happy | 0.01 | 0.00 | 0.01 | 0.80 | 0.09 | 0.05 | 0.03 |
| Sad | 0.05 | 0.00 | 0.02 | 0.05 | 0.69 | 0.18 | 0.02 |
| Surprise | 0.10 | 0.00 | 0.05 | 0.04 | 0.27 | 0.51 | 0.02 |
| Neutral | 0.02 | 0.00 | 0.03 | 0.05 | 0.04 | 0.04 | 0.82 |

Normalized confusion matrix for ResNet



| | Anger | Disgust | Fear | Happiness | Sadness | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Anger | 0.60 | 0.00 | 0.11 | 0.03 | 0.15 | 0.05 | 0.06 |
| Disgust | 0.57 | 0.00 | 0.13 | 0.10 | 0.13 | 0.03 | 0.03 |
| Fear | 0.19 | 0.00 | 0.35 | 0.03 | 0.21 | 0.15 | 0.07 |
| Happiness | 0.04 | 0.00 | 0.02 | 0.86 | 0.02 | 0.02 | 0.03 |
| Sadness | 0.17 | 0.00 | 0.09 | 0.05 | 0.52 | 0.01 | 0.16 |
| Surprise | 0.04 | 0.00 | 0.16 | 0.05 | 0.01 | 0.70 | 0.02 |
| Neutral | 0.11 | 0.00 | 0.08 | 0.05 | 0.17 | 0.02 | 0.57 |

Normalized confusion matrix for ViT

| | **AlexNet (FERNet)** | **ResNet** | **ViT** |
|---|---|---|---|
| Training time | ~13 hours | ~5 hours | 37 minutes (smaller dataset) |
| Number of parameters | 29,069,064 | 223,847 | 86,394,631 |
| Final test accuracy (Human: 65±5%) | 50.4% | 60% | 61% |
| Final training accuracy | 72.4% | 63.8% | N/A |
| Number of epochs | 30 | 30 | 6 |

AlexNet's, ResNet's, and ViT's  Training Results Table

| | **AlexNet (FERNet)** | **ResNet** |
|---|---|---|
| Training time | ~13 hours | ~5 hours |
| Number of parameters | 29,069,064 | 223,847 |
| Final test accuracy<br>(Human: 65±5%) | 50.4% | 60% |
| Final training accuracy | 72.4% | 63.8% |
| Number of epochs | 30 | 30 |

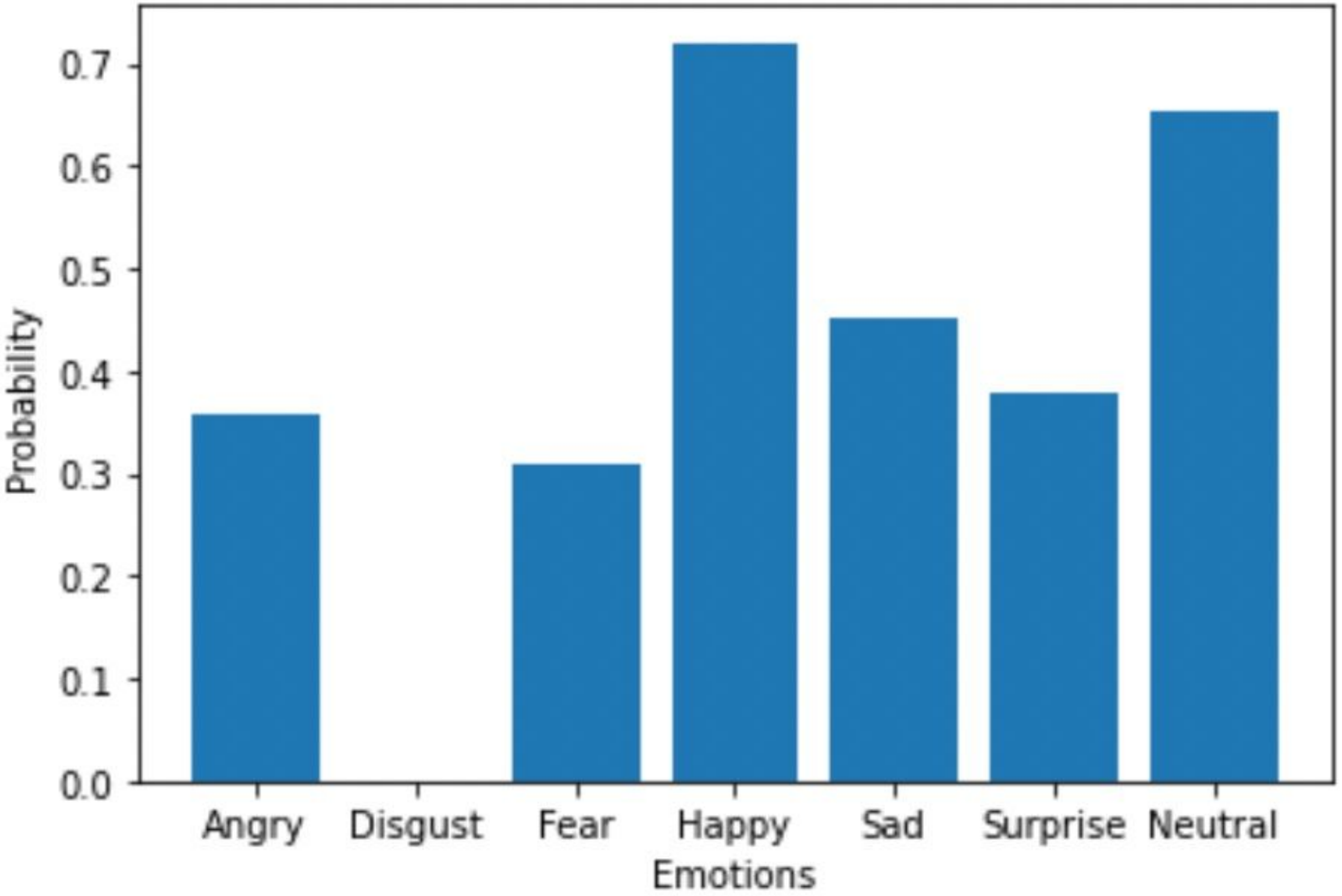
AlexNet's and ResNet's Training Results Table

|  | **AlexNet (FERNet)** |
|---|---|
| Training time | ~13 hours |
| Number of parameters | 29,069,064 |
| Final test accuracy<br>(Human: 65±5%) | 50.4% |
| Final training accuracy | 72.4% |
| Number of epochs | 30 |

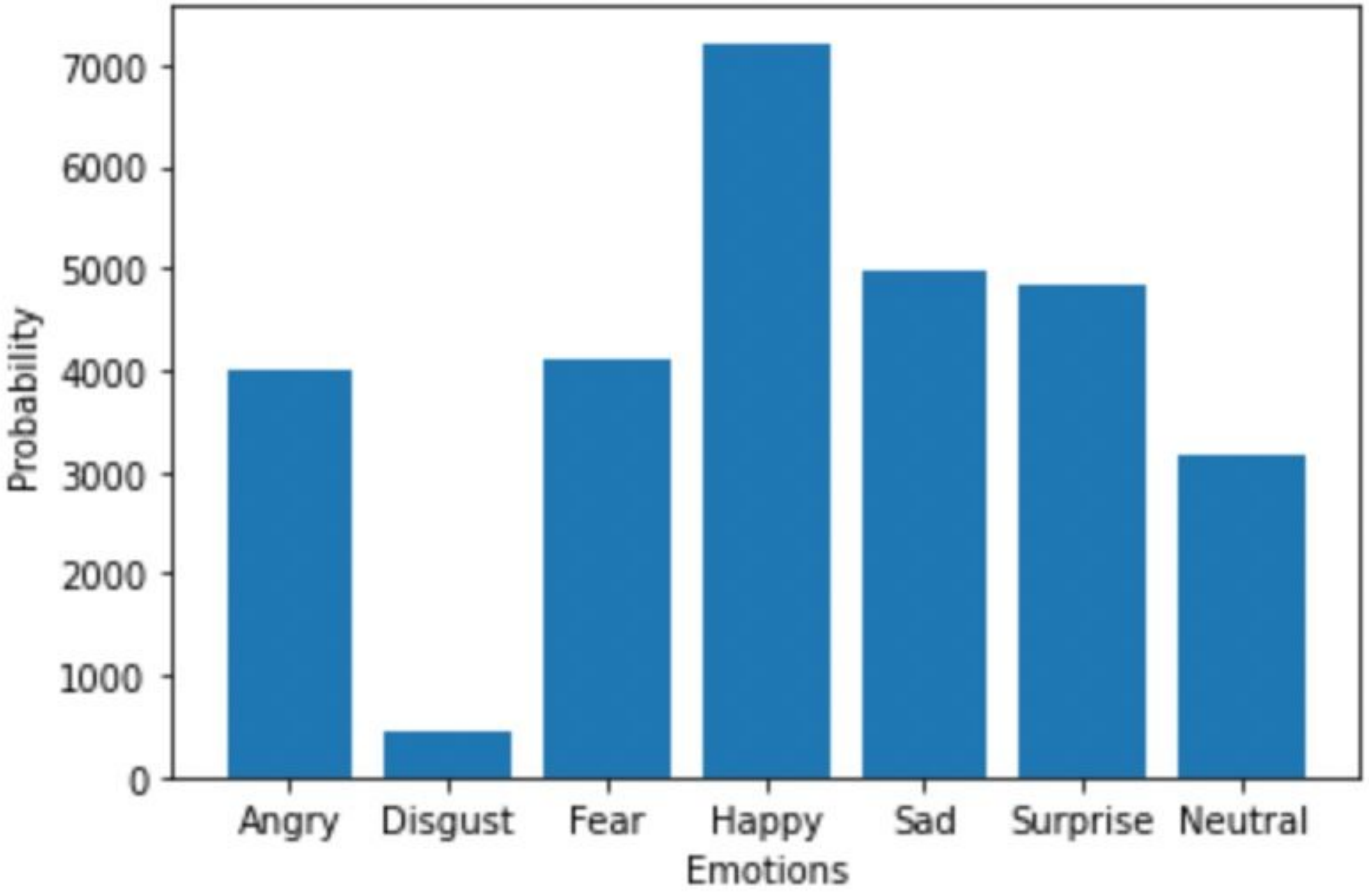
AlexNet's Training Results Table

|  | **ViT** |
|---|---|
| Training time | 37 minutes (smaller dataset) |
| Number of parameters | 86,394,631 |
| Final test accuracy<br>(Human: 65±5%) | 61% |
| Final training accuracy | N/A |
| Number of epochs | 6 |

ViT’s Training Results Table

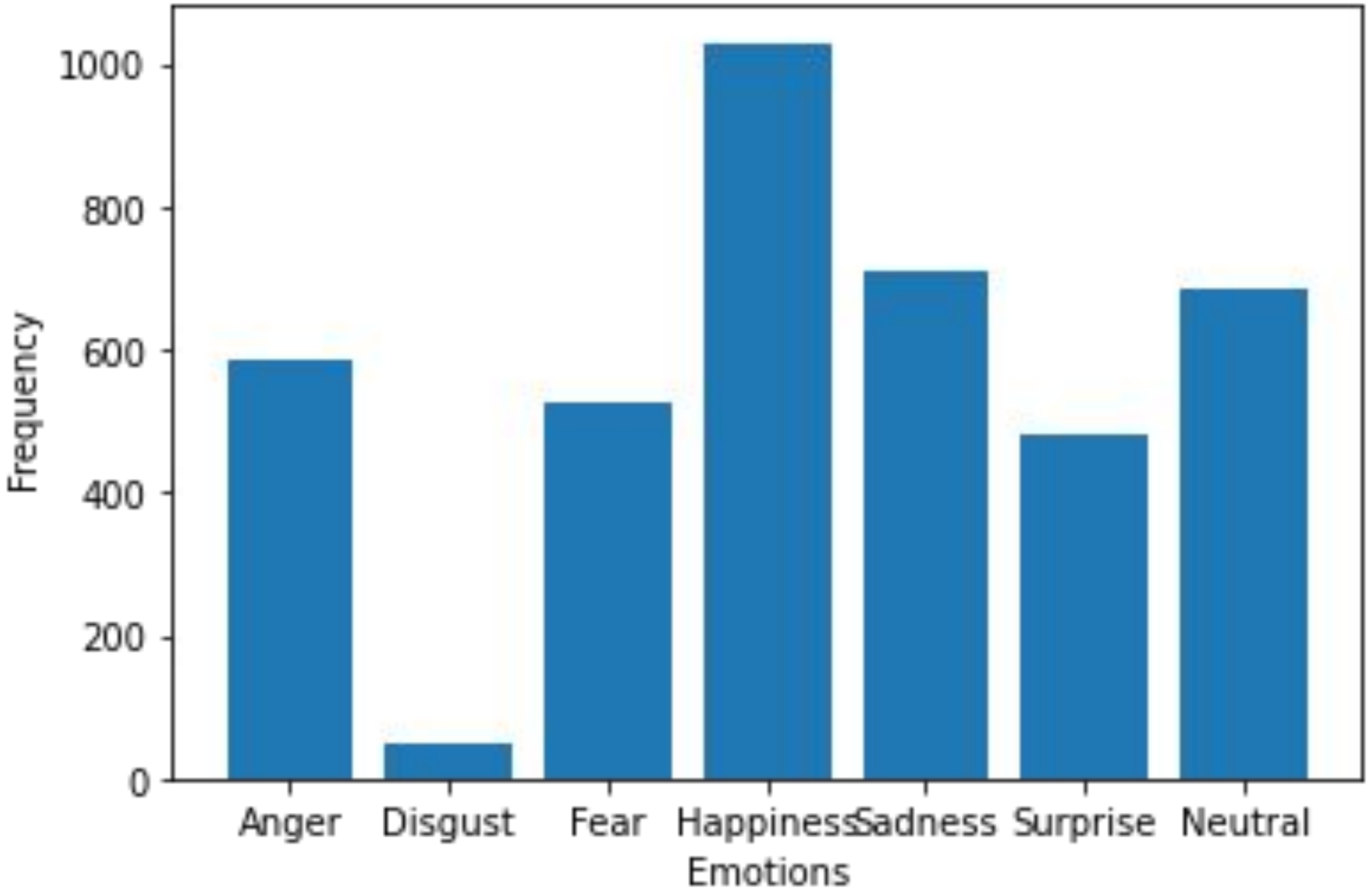


Test accuracy on each label (AlexNet)



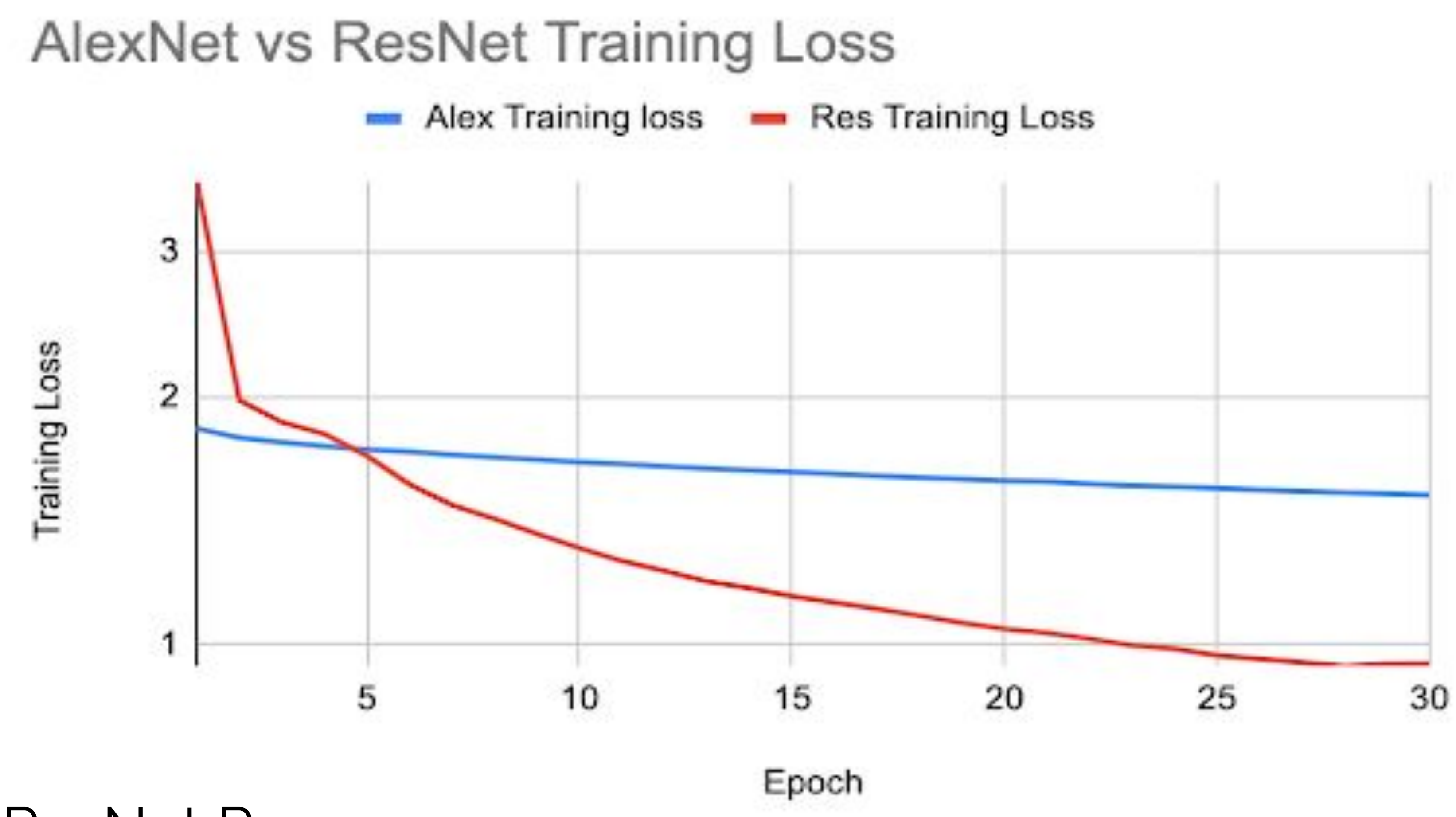
Full FER2013 Dataset

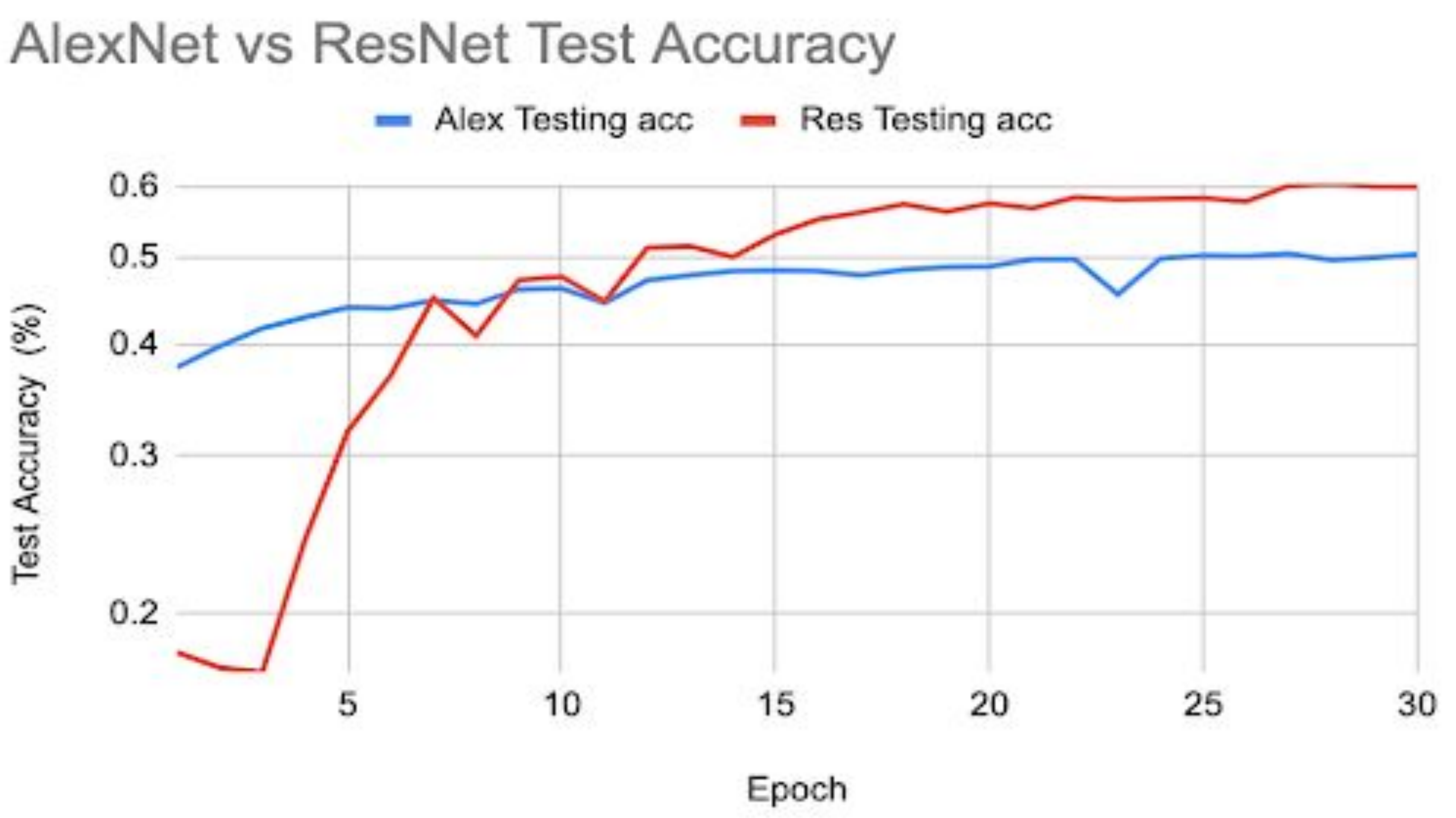
*Y-axis is “Frequency”



Scaled-down dataset used for ViT

## Our Results


AlexNet vs ResNet Training Loss


AlexNet vs ResNet Test Accuracy

## ResNet Paper

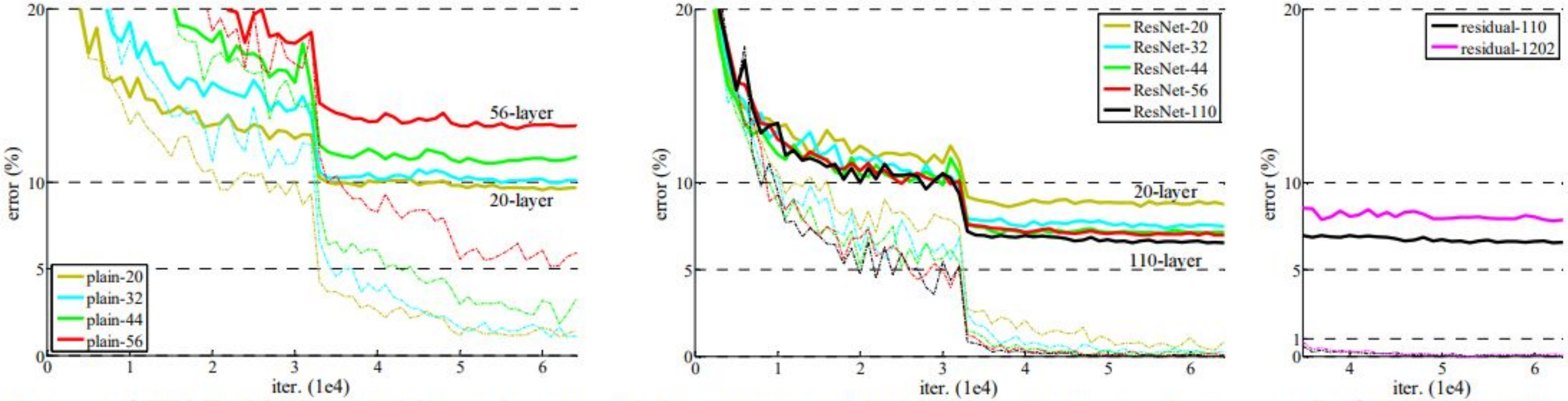
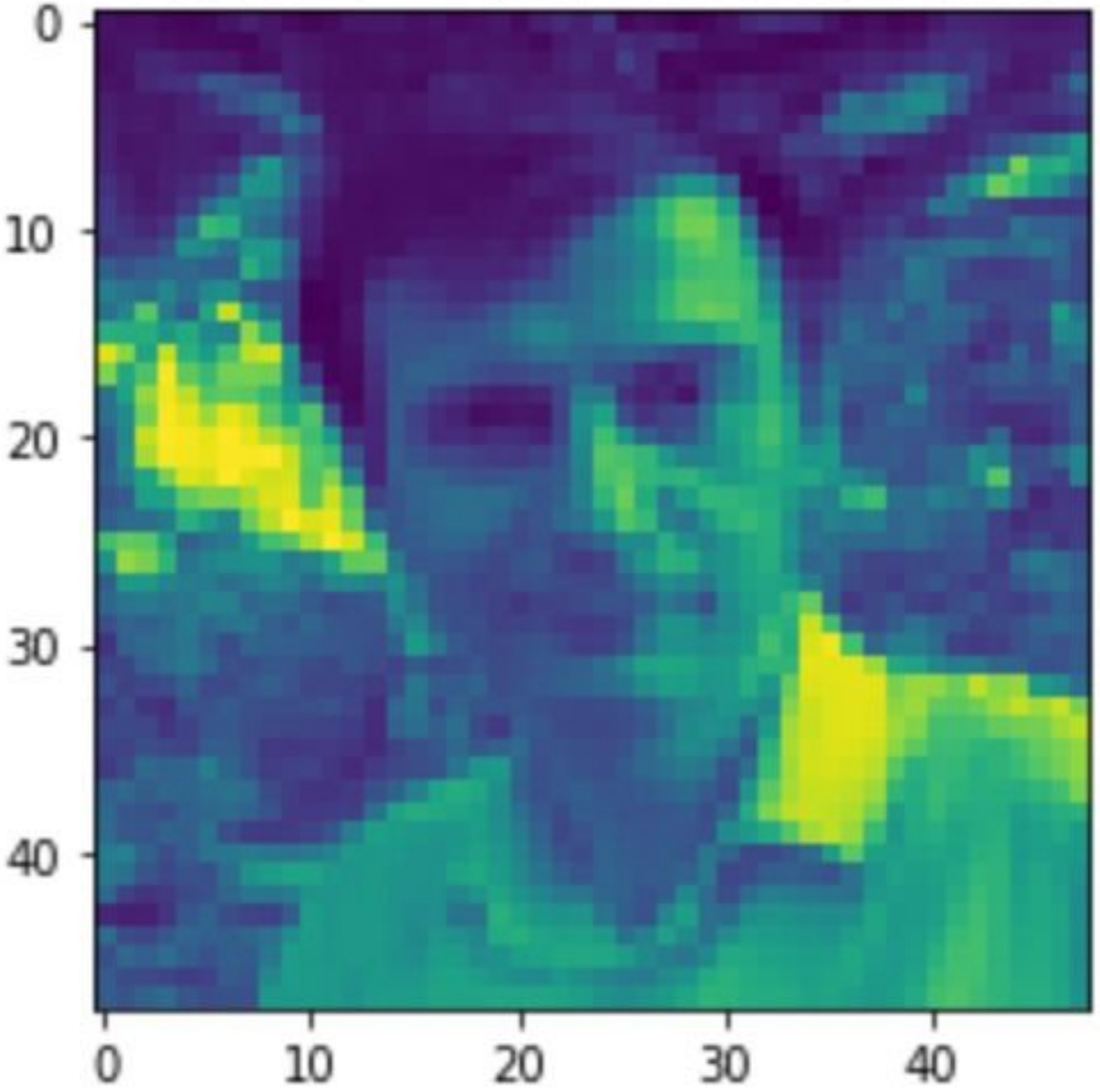


Figure 6. Training on **CIFAR-10**. Dashed lines denote training error, and bold lines denote testing error. **Left**: plain networks. The error of plain-110 is higher than 60% and not displayed. **Middle**: ResNets. **Right**: ResNets with 110 and 1202 layers.

# REFERENCES

arXiv:1307.0414
arXiv:1512.03385
arXiv:2010.11929



Best label = Fear, with Score: 15.18