CNNs and Vision Transformers for Emotion Recognization

Grant Liu[1] **Luoning Zhang**[1] **Ge Gao**[1] **William Chu**[1] **Montek Singh Kalsi**[1]
[1]University of California, Irvine
{grantl3,luoninz1,gaog5,wzchu,kalsim}@uci.edu

1 Introduction

In the last few decades, artificial intelligence and computer vision have gone through remarkable advancements, particularly in the domain of recognizing facial emotions. Two prominent techniques have emerged as frontrunners in this field: Convolutional Neural Networks (CNNs) and Vision Transformers.

CNNs, inspired by the human visual system, have revolutionized the field of computer vision. These deep learning models excel at extracting hierarchical features from images, making them highly effective in tasks such as object detection, image classification, and facial recognition. With their ability to learn spatial hierarchies of features through convolutional layers, CNNs have been extensively employed in emotion recognition tasks, where they analyze facial expressions to infer underlying emotions.

Vision Transformers, however, have introduced a major shift in how we approach visual tasks. Unlike CNNs, which rely on convolutions for feature extraction, Vision Transformers process images as sequences of tokens, leveraging self-attention mechanisms to capture global dependencies and long-range relationships within the input data. This novel approach has demonstrated remarkable performance across various computer vision tasks, including image classification and object detection.

Both CNNs and Vision Transformers offer unique advantages and trade-offs. In this project, we delve into the strengths and limitations of both in the context of emotion recognition tasks. By understanding the underlying principles and methodologies of these approaches, we aim to better understand the features that make up human emotions, and how artificial models identify them.

2 Data

We trained our models on the FER-2013 dataset: https://www.kaggle.com/datasets/msambare/fer2013, a popular benchmark dataset for tasks related to facial expression recognition. It contains a large collection of grayscale images of human faces labeled with one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. The dataset consists of 35,887 images of 48x48 pixels, which makes it suitable for training and testing deep learning models efficiently due to the relatively small sizes of each image. This may also be a concern, however, since the lower image resolution means less features for our models to pick up on. Another potential issues is the fact that the number of samples for each emotion is unbalanced, with a disproportionately large number of "happy" faces and a disproportionately small number of "disgust" faces. This could negatively affect how well our models generalize to new data (Figure 1). To combat this, we resampled our training data inversely proportional to the number of images in each class. This upscales the under-represented group of images and downscales the over-represented groups of images to make the frequency balanced before training the model.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

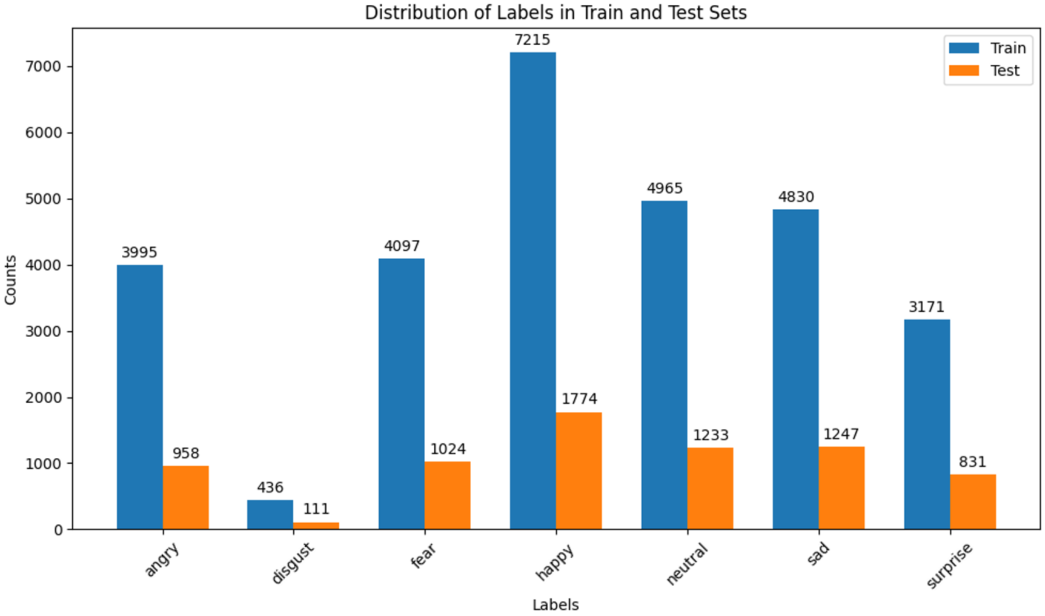

Figure 1: Frequency of different categories in FER-2013 dataset

3 CNN Architecture

3.1 Method

The first intuition for machine learning with images is to use CNNs. CNNs are particularly well-suited for image classification tasks like facial emotion recognition because they automatically learn to detect important features in images. In our initial CNN design (Figure 2), we input the model with the 48 by 48 grayscale image from the FER-2013 dataset, the first convolutional layer captures basic features like edges. Max-pooling layers then reduce the size, focusing on the most significant details. The second convolutional layer could identify more complex patterns in the image. The final fully connected layer progressively combines these features, and outputs a prediction for one of the seven emotions included in our dataset.

This initial design, though theoretically correct, did not reach our intended accuracy, likely due to underfitting, where the model is not complex enough to capture all the details in the images. So we looked to improve the architecture and increase the prediction accuracy. In our final model (V2.1), each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function to introduce non-linearity, and batch normalization to stabilize and accelerate the training process. The network architecture includes double the convolutional layers compared to the initial design, with an increasing number of filters (32, 64, 128, and 256), allowing the model to learn a hierarchy of features, from simple edges to complex textures and patterns. Max-pooling layers with a kernel size of 2x2 are also applied intermittently, reducing the feature map size while retaining the most critical information.

To prevent overfitting and ensure robust learning, dropout layers are integrated with a dropout rate of 0.5, which randomly sets a portion of the input units to zero during training. This technique makes the network not overly reliant on a few weights, thus improving its generalization capabilities. Finally, the network culminates with fully connected layers that flatten the feature maps into a 1-dimensional vector, followed by a fully connected layer that outputs the final class probabilities for the seven emotion categories in the FER-2013 dataset. The combination of these additional convolutional layers, batch normalization, max-pooling, and dropout ensures that the model can effectively capture and discriminate subtle facial features, leading to improved accuracy in emotion recognition.

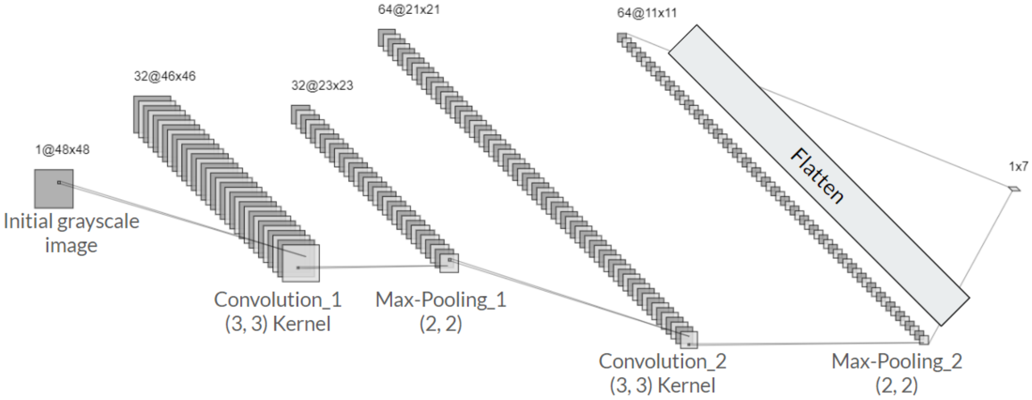

Figure 2: CNN Architecture V1

3.2 Results

After updating our model's architecture, we observed a notable improvement in the training and validation loss curves between the initial and revised models (Figure 3). The initial model (left graph) demonstrated a sharp decrease in training loss from approximately 1.6 to 1.0 over eight epochs, but exhibited signs of overfitting, as evidenced by the validation loss plateauing and slightly increasing after the third epoch. This behavior suggests that while the model learned the training data well, its generalization to unseen data was suboptimal. The revised model (right graph) showed a more extended training period with a significant reduction in training loss from 1.8 to 0.8 over seventeen epochs. Importantly, the validation loss, although initially fluctuating, trended downward more consistently, indicating improved generalization and stability. These enhancements underscore the effectiveness of our adjustments in mitigating overfitting and achieving a more robust model performance.

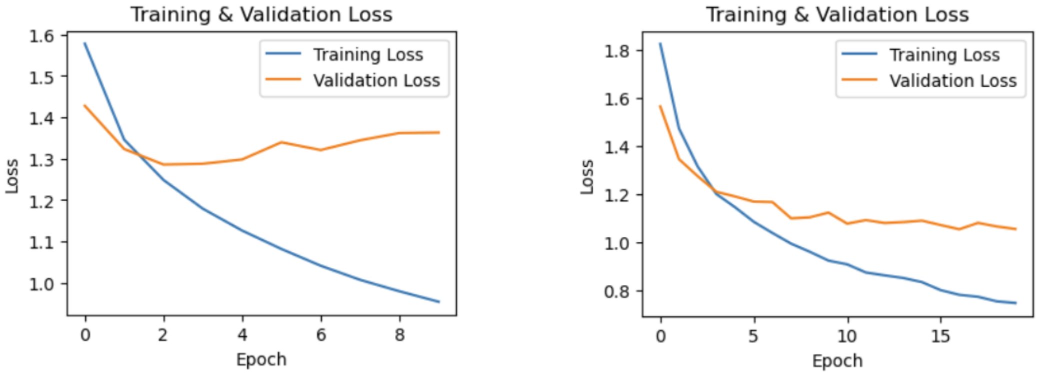

Figure 3: Comparison of CNN loss curves between V1 and V2.1

4 Vision Transformer Architecture

4.1 Method

Vision transformers (ViT), introduced by the paper "An Image is Worth 16x16 Words," combine the scalability and efficiency of transformers commonly used in natural language processing, along with image recognition techniques seen in CNN architectures. Vision transformers are able to utilize these techniques effectively by splitting an image into patches (16x16 pixel squares from the original image), and linearly embedding these patches [1].

Our initial attempt was to mimic the transformers used in that very paper from scratch. However, initial results were discouraging and extremely ineffective, which are mainly due to two reasons. First, we only had one level of encoders, whereas other baseline ViTs have multiple. Perhaps even more importantly, was the modest amount of data in the FER dataset. Again in the same paper, "Transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data." Typically to solve this issue, much more data was used, from between 14M - 300M images, which was orders of magnitude higher than the amount in the FER dataset.

Realizing that it is proper to use a pre-trained model to fit the ViT on a small dataset like FER-2013, we used Google ViT pre-trained models that are initially trained on ImageNet 21k known as "google/vit-base-patch16-224," a baseline model for 224x224 pixel images and splitting into patch sizes of 16x16 pixels. We downloaded this transformer model through Huggingface's Python library "transformers" and then made adjustments on the last few fully connected MLP head that classifies into the number of classes to make the model fit our purpose. Originally, the ViT classifies nearly one thousand classes, but we replaced it for our case of a layer of logits for 7 emotions. This approach allows us to fine-tune an already effective model for our emotion classification task, and preserve weights that draw attention well, resulting in applying those same attention layers to a different image recognition problem.

Initially, we replaced the classifier head with three fully connected layers and observed a relatively low accuracy. Through testing, we realized the model was not powerful enough, and decided to add 2 additional fully connected layers for a total of 5 layers.

We also optimized the training rate through experiments. After trying different initial learning rate, we found the model didn't converge well with the learning rate of 1e-3. The model converged slowly with the learning rate of 1e-5. It converged relatively well with a learning rate of 1e-4 (see the middle graph of figure 4). So, we adopted this initial learning rate.

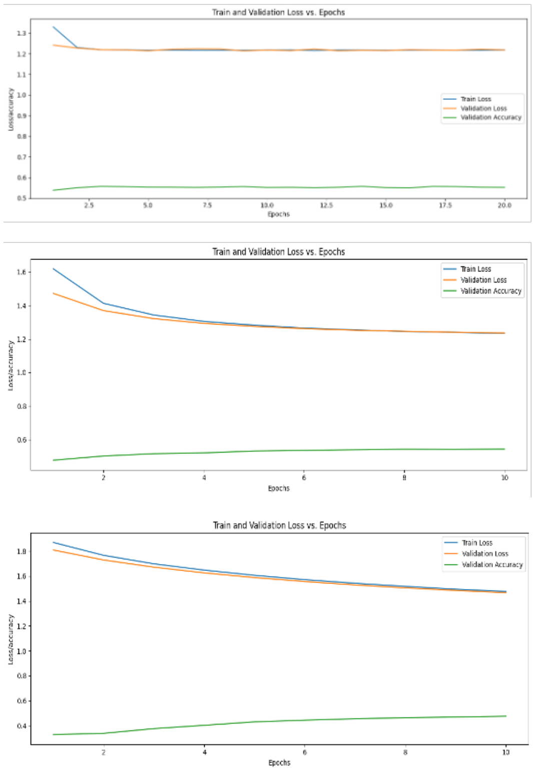

Figure 4: Initial Learning Rate: 1e-3, 1e-4, and 1e-5, respectively from top to bottom

With 24 epochs of training, it was evident that this model, though achieving high accuracy on the training data, is overfitting to the training data (see figure 5). To combat this, we added dropout layers between the fully connected layers of our MLP head. Furthermore, we decided to freeze many of our layers, in particular layers of the pre-trained ViT model. The intuition is that we want to leverage the features of the pre-trained network that was learned from larger datasets and prevent overfitting on these layers. We then gradually unfroze these layers, as our model becomes more accurate, and these layers will be trained with smaller errors.

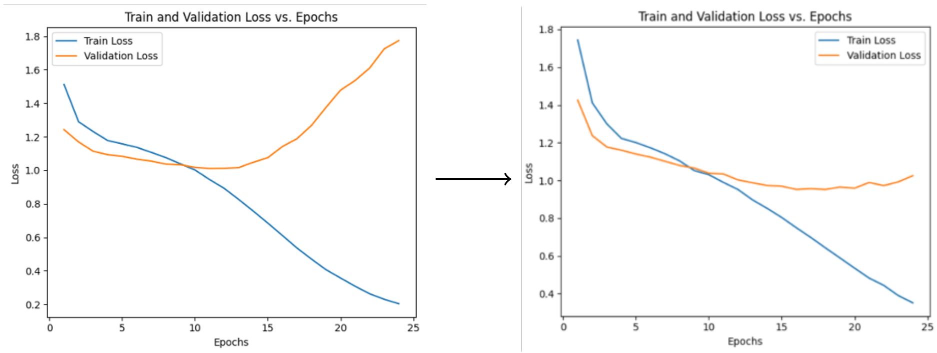

Figure 5: Comparison of loss curves between V1.6 and V1.6.7

4.2 Result

4.2.1 Performance evaluation

After running 24 epochs, we get the lowest validation loss of 0.9520 and the highest validation accuracy of 0.6634. Among the 7 categories, the model predicts "happy" the best with an accuracy of 0.8326, whereas the "Fear" category has the lowest accuracy, 0.4814 (see figure 6). This is likely because fear is such a complex emotion that even humans have trouble distinguishing.

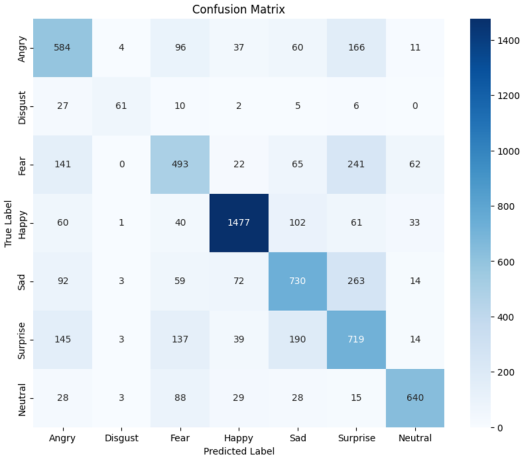

Figure 6: Confusion matrix of ViT model

4.2.2 Generalization

We've also discovered that the ViT model has the capability to generalize to data outside of the FER-2013 dataset. We tested this by inputting real-world images into the model (Figure 7). By resizing the image to 224x224 pixels, the model gave reasonable predictions of the person's emotion in these images.

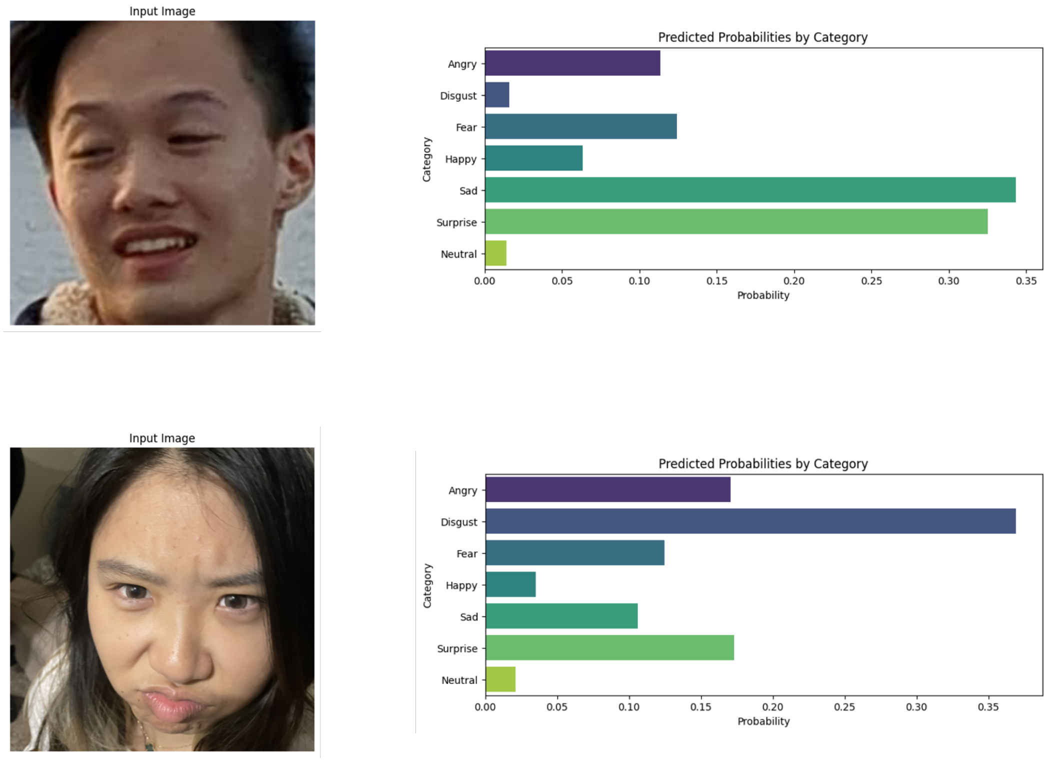

Figure 7: Data out of the training and test dataset

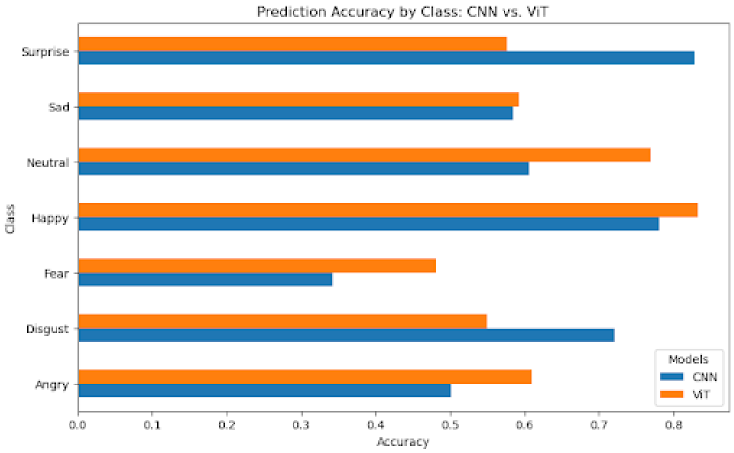

Figure 8: Accuracy of each model over all categories

5 Conclusion and Discussion

Our project found several interesting takeaways. We discovered that ViTs performed better in 5 out of 7 emotion categories (See Figure 8). However, CNNs excelled in categories with fewer training samples, such as disgust. This suggests that CNNs may be more effective when dealing with limited data and lower image resolution, whereas ViTs are advantageous for more complex and diverse datasets. It's important to note that Google's ViT was pre-trained and fine-tuned on a larger and more diverse dataset than our CNN architecture, making it unfair to conclusively state that ViTs are superior for this task. Additionally, the ViT's pre-training on higher resolution, colored images may have impacted its performance on the grayscale FER-2013 dataset, potentially contributing to lower accuracies. Regardless, the pros and cons of each approach are still shown: the CNN had a much simpler architecture and was faster to train with less data, while the ViT showed robustness with high accuracy on data it wasn't trained on. Both models demonstrated the capability to generalize to new, real-world data outside of the dataset, shown by predictions on our images. This indicates their potential applicability in practical scenarios, such as real-time emotion detection in human-computer interaction systems or mental health monitoring tools.

Future research could involve investigating a hybrid ensemble approach that attempts to leverage the strengths of both CNNs and ViTs together. Additionally, addressing the imbalance in the FER-2013 dataset with more advanced techniques such as image augmentation and incorporating more diverse and higher-resolution images could further improve model performance. In conclusion, our findings indicate that the choice between CNNs and ViTs should be guided by the specific characteristics and constraints of the dataset in use, but both architectures are promising for facial emotion recognition tasks.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Contribution

Luoning Zhang: wrote and optimized the ViT code, compiled the ViT section of the slide, wrote the result subsection of the ViT section of this report, organized group meetings, and did goal decomposition and progress tracking;
Grant Liu: Project idea formation, Model proposals, Introduction and Data sections, Presentation slides, Proof reading paper;
Ge Gao: Designing and optimizing the CNN Architecture, handling class imbalance in the dataset, writing the final report (CNN section, final touches), making the final presentation (CNN section);
William: Initial ViT code, additional pretrained Vit models and parameters, writing transformer architectures in report and presentation;
Montek: contributed background information on intro slides about architectures, performed analysis of results in both slides and the paper.