Emotion Detection with CLIP

Dylan Griffin and Efthimios Gianitsos

Abstract

Emotion detection plays a crucial role in several fields including psychology, human-computer interaction, and market research [1, 2, 3]. Here, we present the creation of an AI model for identifying a specific emotion in human faces using deep learning. The model is trained on the FER2013 (Facial Expression Recognition 2013) dataset. Images from this dataset were used as inputs into CLIP, an OpenAI model capable of turning both text and images into embeddings [9]. We then used these embeddings as training data for a multilayer perceptron (MLP) architecture to classify them into the target emotion category. The model achieved a classification accuracy of 71.40% on the test dataset, demonstrating its effectiveness in emotion detection.

We then studied the potential applications of CLIP by comparing the FER2013 image vectors to corresponding text vectors. To reduce the number of dimensions for visualization, the dimensionality of the embeddings are reduced into 2 dimensions with the t-SNE technique for observation and comparison [5].

We release our code at https://github.com/XavierTheCreator1/emotion-detection-ai

1. Introduction

1.1 Applications

The ability to accurately identify and understand the emotional states of individuals through facial expressions can lead to significant advancements in several domains [1, 2, 3]. Additionally, the analysis of the emotional content of human facial expressions is essential for the deeper understanding of human behavior [1]. Identifying emotions using machine learning is challenging [4]. This further suggests the need for this research so we are better able to recognize and understand these human emotions through automation.

1.2 CLIP Experiments

We also studied OpenAI's CLIP model and its capacity to generate image and text embeddings. These embeddings were invaluable to not only improve the performance of the emotion detection model simply trained on raw pixel data, but also to visualize these images in a different way, allowing spatial direction to correspond with meaning. The t-SNE technique proved to be invaluable for displaying the embeddings created by CLIP in a way that was easily understood, compared and measured.

1.3 Model Structure

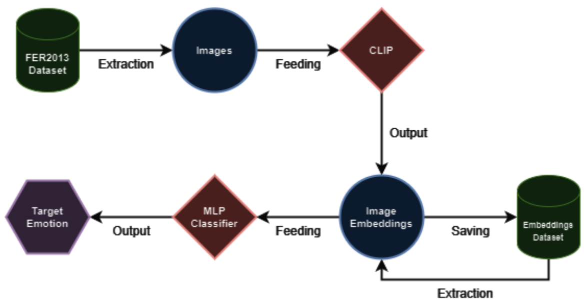

Figure 1: The training structure of the emotion model.

During training, the emotion detection model (shown in Figure 1) first took images from the FER2013 dataset and fed them into OpenAI's CLIP model. CLIP converts these images into vectors (embeddings) which were then saved into a separate dataset. An emotion classifier was then trained on these image embeddings, which are each marked with their target emotion.

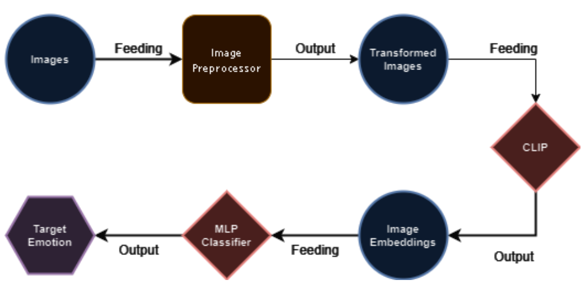

Figure 2: The application structure of the emotion model.

During application, the model shown in Figure 2 took input images and preprocessed them into 48x48 grayscale images (the same image structure in the FER2013 dataset [6]). These preprocessed images were fed into CLIP which then translated the images into embeddings. The embeddings were then directly fed into the emotion classifier to predict the emotion demonstrated in the image.

2. Dataset

The FER2013 dataset consists of 48x48 pixel grayscale images of faces. The faces were centered and occupied approximately the same area in each image [6]. The training set consisted of 28,709 images and the test set consisted of 3,589 images. Consequently, we trained our model using the 28,709 image training set and tested the model using both the training set and the 3,589 image test set.

3. Experimentation

This section describes the development process for the emotion detection model as well as experimentation of the capacities of CLIP.

3.1 Emotion Detection Using Raw Images

The first test examined model performance without the inclusion of CLIP. An initial model was created that fed the raw images from the FER2013 dataset into an MLP classifier.

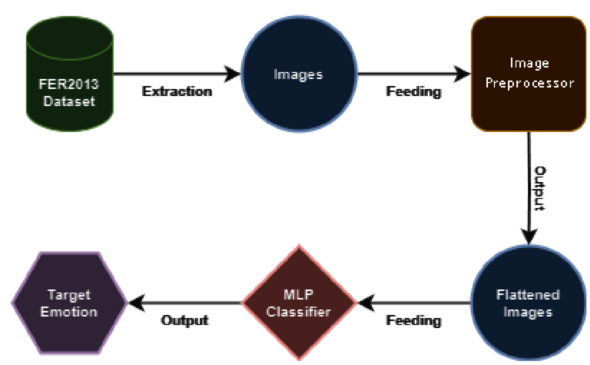

Figure 3: The training structure of the initial MLP emotion model.

However, the results of the model shown in Figure 3 were inaccurate. The best iteration of this model, after a rigorous fine-tuning of hyperparameters, had a testing accuracy of 41.40%, a result far lower than many other models trained on the FER2013 dataset which boasted an accuracy over 60% [7, 8].

To improve this performance, a different model was proposed utilizing a random forest classifier, opposed to the previously used MLP classifier.

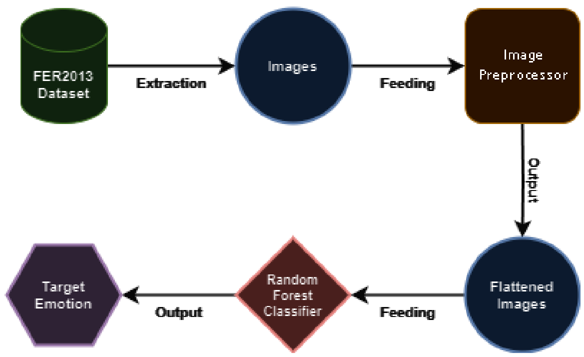

Figure 4: The training structure of the initial random forest emotion model.

The model shown in Figure 4, outperformed the previously tested MLP model, but was still inaccurate. An ideal iteration of this random forest model with 10,000 decision trees presented a testing accuracy of only 48.90%.

3.2 Emotion Detection With CLIP

OpenAI's CLIP model was effectively trained to convert image and text data into embeddings. We speculated that these embeddings provided higher-quality training data to a classifier, resulting in a higher-performing model.

Since the random forest classifier outperformed the MLP classifier without CLIP, it was originally assumed that even with CLIP, a random forest will outperform an MLP.

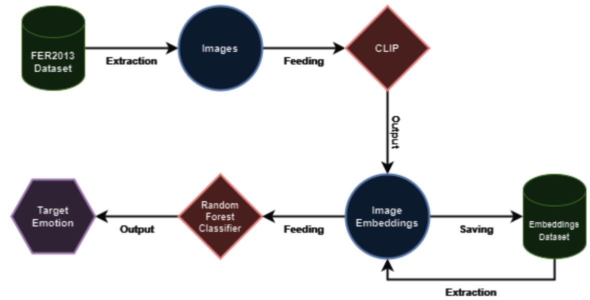

Figure 5: The training structure of the emotion model using CLIP and a random forest classifier.

The addition of CLIP into the model structure (Figure 5) improved performance. This structure, using a random forest classifier with 10,000 decision trees, had an accuracy of 68.17%, much higher than other emotion classification models trained on FER2013 [8].

To further improve the model we tested an MLP classifier trained on the CLIP-generated image embeddings. As seen in Figure 1, the FER2013 image dataset was fed into CLIP to generate a dataset of image embeddings, which were then used to test an MLP classifier over many different combinations of hyperparameters. By testing many hyperparameters of an MLP classifier, we adapted the model to emotion detection. After optimization, the model had an accuracy of 71.40%, one of the best accuracies resulting from a model trained on the FER2013 dataset, performing better than many other well-researched and trained emotion classification models [7, 8].

3.3 Embedding Comparison With CLIP

The capacities of CLIP were further explored with a different kind of experiment. We visualized both text and image embeddings created by the CLIP model.

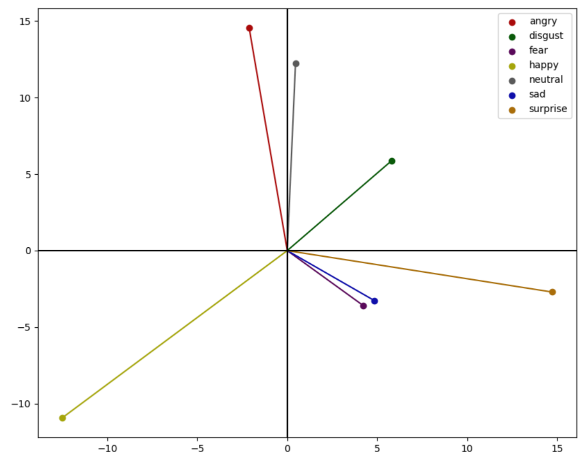

Figure 6: The image embeddings generated using CLIP and visualized through the t-SNE technique, averaged and categorized by target emotion.

The image embeddings displayed in Figure 6 present a conceptually interesting dilemma. Generally for humans, the opposite emotion of “happy” would be “sad” [10]. However, the image embeddings depict the opposite of “happy” to be “disgusted” as evident through the two vectors being seemingly reflected on both the x- and y- axes. When considering the raw images themselves, this conclusion drawn from these embeddings seemed to be supported. The FER2013 faces labeled as “disgusted” demonstrated the most negatively connotated and exaggerated facial features out of any other set of emotions. Oppositely, the faces labeled as “happy” had much simpler facial features with a much more positive connotation.

Furthermore, we found that images labeled “sad” and “fear” were quite similar, even when dimensionally reduced. This relation can be seen as self explanatory as humans often have similar facial expressions when feeling sad and fearful.

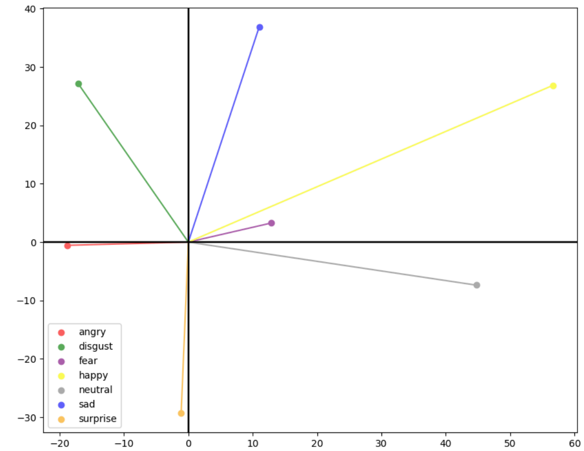

Figure 7: The text embeddings generated using CLIP and visualized through the t-SNE technique.

Text embeddings were created by sentences describing each human emotion. For example, the embedding for “happy” was created using the string “A happy human face”. The text embeddings demonstrated none of the patterns found in the image embeddings. The vector for “happy” in Figure 7 did not have opposing x- and y- axes as it did in Figure 6. To further grasp the differences between the image and text dimensionally-reduced embeddings, an overlay including Figure 6 and Figure 7 was created.

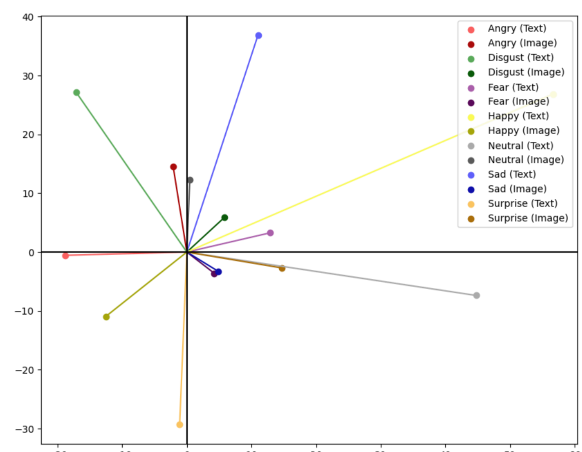

Figure 8: Figure 6 and Figure 7 overlayed.

An overlay between Figure 6 and Figure 7 was created to find a pattern between the image embeddings and their text counterparts. However, no such patterns were found as the image embedding visualizations were almost always unrelated to their text counterparts, evident by their near-perpendicular relationship between the t-SNE representations of the image and text embeddings for the majority of the emotions. This lack of relationship is likely reflected by the t-SNE technique which has limitations when reducing the dimensionality of data for visualization because of data loss.

4. Conclusion

When utilizing CLIP in tandem with an MLP classifier, an accurate model can be formed to predict human emotion using images of faces. Comparatively, CLIP - in cases such as those described within this paper - has the capacity to significantly improve the performance of models originally trained on raw text or image data.

Furthermore, it was found that CLIP's capacity to generate embeddings can be used in other applications, such as the comparison of emotion expressed as both text and image embeddings.

References

[1] Zhao, Jian, et al. "Cognitive psychology-based artificial intelligence review." *Frontiers in neuroscience* 16 (2022): 1024316.

[2] Zad, Samira, et al. "Emotion detection of textual data: An interdisciplinary survey." *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021.

[3] Kusal, Sheetal, et al. "AI based emotion detection for textual big data: techniques and contribution." *Big Data and Cognitive Computing* 5.3 (2021): 43.

[4] Giannopoulos, Panagiotis, Isidoros Perikos, and Ioannis Hatzilygeroudis. "Deep learning approaches for facial emotion recognition: A case study on FER-2013." Advances in hybridization of intelligent methods: Models, systems and applications (2018): 1-16.

[5] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

[6] Sambare, Manas. "FER-2013." *Kaggle*, 19 July 2020, www.kaggle.com/datasets/msambare/fer2013.

[7] Kusuma, Gede Putra, J. Jonathan, and A. P. Lim. "Emotion recognition on fer-2013 face images using fine-tuned vgg-16." Advances in Science, Technology and Engineering Systems Journal 5.6 (2020): 315-322.

[8] Zahara, Lutfiah, et al. "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi." *2020 Fifth international conference on informatics and computing (ICIC)*. IEEE, 2020.

[9] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[10] Solomon, Robert C., and Lori D. Stone. "On "positive" and "negative" emotions." Journal for the theory of social behaviour 32.4 (2002).