Expanding Facial Emotion Recognition: Contrastive   
Learning on a Custom FER-2013 Dataset

Jacqueline Torres

St. John’s University, College of Professional Studies  
jacqueline.torres19@my.stjohns.edu

Abstract

Emotion recognition plays a pivotal role in human-computer interaction, affecting areas such as user experience, mental health monitoring, and personalized services. This study explores the application of contrastive learning techniques to enhance emotion recognition using the Facial Expression Recognition 2013 (FER2013) dataset, augmented with two additional emotions: "Interested" and "Disappointed." The proposed approach leverages a Siamese network architecture with a MobileNetV2 backbone to learn rich representations of facial expressions. By training the network to discriminate between similar and dissimilar pairs of images, the model learns to capture subtle differences in facial features associated with each emotion. Results demonstrate the contrastive learning approach's effectiveness, achieving some performance in recognizing the expanded emotions. The findings suggest contrastive learning can be a valuable tool for enhancing emotion recognition systems, particularly when applied to datasets with additional emotional categories.

Introduction*[[1]](#footnote-1)*

As stated by Prinz (2012), there is no set number of emotions. Emotions are an essential aspect of expressing one’s feelings to others. As humans, we can understand facial expressions, but a computer may not be able to understand them. Due to this, there have been various approaches to train machines to understand emotions from images of facial expressions. In a literature review done by Ko et al. (2018), there are three main components that are composed of FER approaches which are face and facial detection, feature extraction and expression clarification.

Some notable and innovative studies that have provided advancements in FER analysis are Ko et al. (2018), Khaireddin and Chen (2021), and Roy et al. (2023). Ko et al. (2018) compared multiple algorithms used for FER and the accuracy produced on an MMI dataset, which contains six basic emotions: happiness, sadness, surprise, anger, disgust, and fear. Khaireddin and Chen (2021) focused on improving the prediction accuracy of the specific emotion recognition dataset called FER2013, containing 65-68% human accuracy, by using CNNs while adopting the VGG network. The FER 2013 is another dataset that contains seven emotions: happiness, sadness, neutral, anger, surprise, disgust and fear. Lastly, Roy et al. (2023) uses a novel contrastive learning technique named ViewFX to aid in recognizing emotions from images that may be in different angles or intensities on KDEF and DDCF datasets. The KDEF contains around seven different emotions while the DDCF contains eight different emotions, two of them being happy with mouth open and closed. Most datasets range from six to eight emotions that primarily contain: happiness, sadness, neutral, anger, fear, disgust, and surprise. There is a gap of knowledge when it comes to adding different emotions that don’t fall within these categories into numerous FER techniques.

Contrastive learning is a powerful technique in self-supervised learning. It allows models to learn meaningful representations from unlabeled data by distinguishing similar data pairs from dissimilar ones as seen in Figure 1. This approach has led to cutting-edge results in various fields, including computer vision, natural language processing, and bioinformatics (Shen et al., 2023). The main focus on the study that will be reviewed in this report is the usage of a contrastive learning architecture called Siamese networks using a MobileNetV2 base model and applying it with a contrastive loss to distinguish between similar and dissimilar pairs. The hope is that the model will be able to maximize the distance between dissimilar pairs and minimize the distance between similar pairs.

A diagram of a model

Description automatically generated

Figure 1. Explanation of a General Contrastive Learning Process

**Method**

Dataset

FER-2013 contains 35,685 images that have seven different emotions. Those emotions include happiness, sadness, neutral, anger, surprise, disgust and fear. A Kaggle dataset contained two different emotions that were going to be added on which were ‘interested’ and ‘disappointed’. The emotion interested contained 10,281 images while the emotion disappointed contained 17,210 images.

Preprocessing

To apply the model, there must be preprocessing techniques applied to the dataset to allow for a smoother process.

The first step in the preprocessing process is removing the corrupted images from the dataset. This is essential due to corrupted images can greatly impact the results as well as interrupt any other preprocessing techniques.

The second step was to reduce the dataset. The dataset contains too many images and there is an imbalance of the amount per class as seen in Table 1.

The third step would be to rename the images in the dataset since many of these images have names that are random so will create less confusion with the class they belong to when creating pairs.

The fourth step would be to create a validation set that contains 30% of the images from the training set.

The last step in the preparation would be the creation of similar and dissimilar pairs with designated labels. During this same process, there will be the process of augmenting the images to create variability to help train the model.

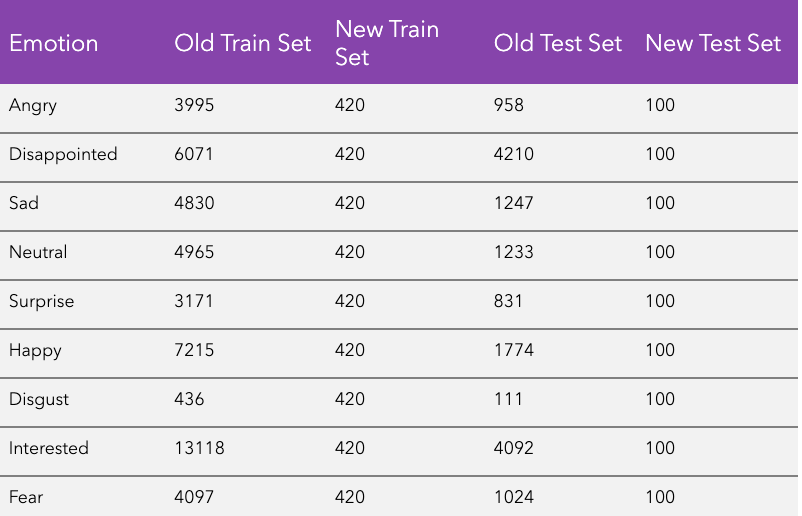


Table 1. Amount of images per class in the old training set and testing set compared to the reduced training and testing set.

Siamese Network

First proposed by Bromley et al. (1993) for signature verification, Siamese networks consist of twin subnetworks that accept separate inputs. These subnetworks share weights and are joined by a top layer computing a metric between their highest-level feature representations. This weight-tying enforces that highly similar inputs are mapped to close locations in the feature space, regardless of the input order due to the network's symmetry.

This model was used instead of other models such as SimCLR (Chen et al., 2020) due to Siamese networks being designed for metric learning tasks. This goes hand in hand with what the model is trying to do as in determining how similar and dissimilar pairs are.

Model Layout

The proposed model leverages a Siamese network architecture with a MobileNetV2 backbone for facial emotion recognition. The MobileNetV2 pre-trained model serves as the feature extractor for the Siamese network. MobileNetV2 is particularly suitable due to its efficient architecture, making it computationally feasible for training the Siamese network.

The Siamese network consists of two identical branches, each processing a separate facial image input. These branches share the weights of the MobileNetV2 model, ensuring they learn consistent feature representations for the faces. Each branch passes the input image through the MobileNetV2 model, which extracts high-level features that capture the essential characteristics of the face. A fully-connected layer with 256 neurons and a ReLU activation function further processes the extracted features from each branch. This step can be seen as projecting the features into a lower-dimensional space suitable for comparison.

A Lambda layer calculates the absolute difference between the encoded representations (embeddings) generated by the two branches. This absolute difference represents the distance between the features in the embedding space, with a smaller distance indicating a higher similarity between the two faces.

A final dense layer with a single neuron and a linear activation function transforms the distance value into a single output score. Ideally, this score will be lower for similar faces (same emotion) and higher for dissimilar faces (different emotions).

The model is trained using the contrastive loss function. The RMSprop optimizer helps guide the training process by adjusting the model's weights to optimize the contrastive loss.

A learning rate scheduler is implemented to dynamically adjust the learning rate during training. This helps ensure the model converges effectively and avoids overfitting to the training data.

Finally, there is an early stopping function that stops the model from going over 5 epochs if the val\_loss results fall within a certain limit.

Contrastive Loss

The contrastive loss is a function defined by two distance metrics. The first distance metric is the squared distance for similar pairs. This minimizes the distance between similar pairs. The second measure is the hinge loss for dissimilar pairs. This term penalizes the model only when the predicted distance for a dissimilar pair is less than the margin. The larger the difference (i.e., the farther apart the dissimilar points are predicted), the smaller the penalty.

Image Generator

Since the image pairs are saved in the format of “firstinputclass\_1.jpg\_secondinputclass\_2.jpg”, there is a difficulty for the model to be able to identify what the two inputs are. The image generators allow for the modeel be able to classify that the first image in the pair is the first input and the second image in the pair is the second input. This generator also allows for the labels to relate to the images as they become separated when becoming inputs.

Results

The results aren’t the best due to limiting factors but indicate potential with improvements to the model. The training set loss is slowly decreasing, which may indicate the model may be learning (Lehn et al, 2023). The validation set loss is also slowly decreasing, which indicates that the model can generalize to unseen data. The decreasing validation set accuracy may indicate that the model is overfitting so creating measures to prevent overfitting would be a great advancement to the current model. The final accuracy of the model for the test set is around 1% with a loss of 0.0751. Due to the loss being closer to 0, there is a potential that the model can evaluate how similar and dissimilar two inputs are when inputted into a dataset that the model was not trained on previously.



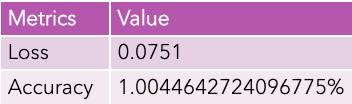


Table 2. a) Summary of Results per Epoch for Training Set and Validation Set b) Loss and Accuracy of a Test Set

Conclusion

Our study demonstrates the potential of contrastive learning for enhancing facial emotion recognition on the custom FER2013 dataset. By augmenting the dataset with additional emotions and leveraging a Siamese network architecture, we achieved a potential backbone in a model that can decrease loss which allows for demonstrating how well the model is in recognizing the "interested" and "disappointed" emotions, as well as the original FER2013 emotions. These findings suggest that contrastive learning can be a valuable tool for expanding the range of detectable emotions in facial expression recognition systems, paving the way for more nuanced and accurate emotion recognition in real-world applications. Limitations that were encountered throughout the method were primarily due to the amount of cost that the large dataset required and was not achievable. Future research will require to create a better model to apply to the whole dataset by adding on more layers as the amount of images increases.

References

Bromley, Jane, Bentz, James W, Bottou, Leon, Guyon, ´Isabelle, LeCun, Yann, Moore, Cliff, Sackinger, Eduard, and Shah, Roopak. Signature verification using a siamese time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence, 7 (04):669–688, 1993.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. https://arxiv.org/abs/2002.05709

Koch, G., Zemel, R. & Salakhutdinov, R. (2015). Siamese Neural Networks for One-shot Image Recognition. Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2105.03588.

Ko, B.C. (2018). “A Brief Review of Facial Emotion Recognition Based on Visual Information”. MDPI, 18(401). doi:10.3390/s18020401.

Lehn, F. vom. (2023, November 8). Interpreting training/validation accuracy and loss. Medium. https://medium.com/@frederik.vl/interpreting-training-validation-accuracy-and-loss-cf16f0d5329f

Prinz, Jesse, ' Emotions: How Many Are There?', in Eric Margolis, Richard Samuels, and Stephen P. Stich (eds), The Oxford Handbook of Philosophy of Cognitive Science, Oxford Handbooks (2012; online edn, Oxford Academic, 1 May 2012), https://doi.org/10.1093/oxfordhb/9780195309799.013.0008.

Sambure, M. 2020. FER-2013, Version 1. from https://www.kag

Gle.com/datasets/msambare/fer2013/data

Roy, S. & Etemad, A. (2023). Contrastive Learning of View-Invariant Representations for Facial Expressions Recognition. https://arxiv.org/pdf/2311.06852

Sarkar, K. 2020. Facial-Expression-Classification-Dataset (3Classes), Version 1. from https://www.kaggle.com/datasets/nig

htfury007/fercustomdataset-3classes/data

Tensorflow. 2024. Tf.Keras.Applications.VGG16: Tensorflow V2.16.1.https://www.tensorflow.org/api\_docs/python/tf/keras/applications/VGG16

1. Copyright © 2023 by the authors.

   This open access article is published under the Creative Commons Attribution-NonCommercial 4.0 International License. [↑](#footnote-ref-1)