MINDMAP AI: INTEGRATING CUSTOM TRANSFORMER AND CNN ARCHITECTURES FOR EMOTION CLASSIFICATION

Matthew Schulz, Rishikesh Sivakumar, Sai Nikhil Reddy Marreddy, Kaiyuan Lin, Jinyong Han
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA
{mkschulz, rs22010, smarredd, linkaiyu, jinyongh}@usc.edu
Github Repository: https://github.com/marreddysainikhilreddy/emotion-classification

1 INTRODUCTION

Our project began with the innovative concept of a mental health diary application, envisioned to automatically capture images of users as they wrote their text diary entries. The core idea for the application was to analyze this combined data to assess emotional states and provide recommendations for mental health improvement or maintenance. However, insightful feedback prompted us to pivot our focus towards the foundational aspect of this application – the machine learning models responsible for the emotional assessment.

We zeroed in on two models: a transformer-based model for emotion classification in text and a convolutional neural network (CNN) model tailored for facial emotion classification in images. The choice of these models stemmed from their proven efficacy in their respective domains – the transformer model's advanced capabilities in contextual understanding of text and the CNN's proficiency in recognizing and classifying visual and emotional cues.

1.1 GOAL

The primary aim of our project is to significantly enhance the performances of both the text-based transformer and the image-based CNN emotion detection models. We have set a specific goal to surpass established performance benchmarks for each model's dataset, focusing on improving metrics such as the F1 score in the transformer model and the accuracy in the CNN model.

1.2 IMPORTANCE

A thorough review of existing transformer and CNN models for our task revealed that the performances of established benchmarks for the datasets we use, particularly in the context of our application's demands, were suboptimal. Our project is driven by the imperative to improve these performances, ensuring that the models recognize human emotions with higher accuracy and reliability.

This performance gap highlighted the need for models that could more adeptly handle the intricacies of emotional data, prompting us to investigate and develop solutions. Therefore, our project's significance is not just in advancing the technical aspects of machine learning models but in elevating the standard for applying these models in practical, impactful ways in the mental health field.

1.3 CONTRIBUTION

In our project, we significantly enhanced conventional transformer and CNN models by implementing custom architectures and preprocessing steps to address their limitations in emotion classification. With the transformer model, our focused research on emotions relevant to mental health

allowed us to streamline its classification task, reducing the number of emotions classified without losing analytical depth. After evaluating various models, including BERT, DistilBERT, RoBERTa, T5, XLNet, and ELECTRA, we found RoBERTa to be the most effective. Further innovation involved incorporating Bi-LSTM networks into our transformer model, enhancing its ability to process temporal text data. This resulted in our final model achieving a substantial increase in the validation F1 score, reaching 59.03% on the GoEmotions dataset, up from the initial benchmark of 49%.

We experimented with architectures such as Vgg16, Resnet50, MobileNetv2, and EfficientNet for the CNN model, ultimately customizing layers on top of a pre-trained Vgg16 model. This customization included Gaussian Noise, Global Average Pooling, Flatten, Fully Connected layers, batch normalization, and dropout configurations, leading to a robust and accurate model. Our CNN model demonstrated impressive performance on the fer2013 dataset, achieving a validation accuracy of 90.33% and an F1 score of 66.85%. Key preprocessing steps like data augmentation and MTCNN for facial recognition in images further improved the model's performance, particularly in challenging conditions like low light and resolution. These techniques have been instrumental in enhancing the model's generalization to real-world scenarios. Our approach underlines the efficacy of personalized model development for emotion detection, as we've shown that tailored models can surpass general architectures in specific applications like mental health monitoring.

2 Related Work

2.1 Transformer Model

The first paper Zhang et al. (2020) introduces the Knowledge Aware Incremental Transformer with Multi-task Learning (KAITML) for emotion recognition in textual conversations. Although we have not implemented experiments using the methods presented in this paper, we plan to use ideas for future work.

The second paper Murfi et al. (2022) demonstrates how the BERT model, particularly in combination with LSTM-CNN (Long Short Term Memory-Convolutional Neural Network), enhances the performance of hybrid models in sentiment analysis. We used this paper to help guide our hybrid models.

The third paper Lee et al. (2023) presents the Transformer Transfer Learning (TTL) model for emotion detection and contributes a self-reported emotion dataset collected from tweets. We used this paper for inspiration in using self-reported data.

The fourth paper Demszky et al. (2020) is notable for data contribution, introducing the GoEmotion dataset and a benchmark model with an F1-score of 46%. We used this paper as one of our performance benchmarks.

Finally, the fifth paper Corti (2021) compares various transformer language models (BERT, DistilBERT, RoBERTa, XLNet, and ELECTRA) on the GoEmotion dataset Demszky et al. (2020), evaluating them using the F1 score. We also wanted to conduct a transformer comparative analysis and evaluate the models on more metrics.

2.2 CNN Model

The paper, Khaireddin & Chen (2015), improves facial emotion recognition(FER) using CNNs, mainly focusing on the VGGNet architecture. It sets a new benchmark for accuracy in FER on the FER2013 dataset, achieving 73.28% without additional training data. Since it includes extensive experiments on optimization techniques and hyperparameters, it helped us to fine-tune VGG based on FER2013 data.

The second paper, Pramerdorfer & Kampel (2018), reviews the state of the art in CNN-based FER, highlighting differences in methodologies, CNN architectures, preprocessing, and performance impacts. Through it, we can figure out the overall process required for face detection and recognition of emotional expressions using CNN.

The third paper, Khanzada et al. (2022), presents customized deep learning models for FER, achieving 75.8% accuracy on the FER2013 dataset, an improvement over existing models. Techniques like

transfer learning, data augmentation, and ensembling were key to this success. Based on this paper, we planned how to construct customized layers and how to finetune ResNet.

The fourth paper, Kusunose et al. (2022), focuses on improving facial expression emotion recognition. It uses transfer learning and a generative model, StyleGAN2, for data augmentation. The approach significantly improves recognition accuracy, especially in scenarios with limited training data. We have not completed the experimental results using this method, but it would be helpful for future work.

The final paper, Li & Deng (2018), is a comprehensive survey on deep learning techniques for FER. Because it deals with the overall process of FER, it has helped us to approach what data sets exist in terms of faces, how deep learning deals with FER in the low-level stage, and how High-Level architectures should be chosen.

3 Model Information

3.1 Task

3.1.1 Transformer Model

Our project employs a transformer model for the intricate task of multi-label emotion classification in text, adeptly capturing human emotions' overlapping and diverse nature. This model, particularly suited due to its self-attention mechanism, adeptly handles the nuances of language and context, effectively assigning multiple emotion labels to textual data. Despite challenges like class imbalance and the complexity of multi-emotion correlation, the transformer model proficiently generates a probability distribution for each emotion class. This capability enables a nuanced analysis of emotional states conveyed in text, providing essential insights for applications requiring detailed emotion recognition.

3.1.2 CNN Model

The primary objective of this study is to develop a model that effectively detects user's emotional states by analyzing their images using our CNN-based model. We have collected datasets containing photos of users depicting various emotional states, which helped us train the CNN model. We have added custom layers on top of pre-trained vgg16 and finetuned it to make better predictions. Finally, we analyzed and interpreted the emotion detection results of our trained model and assessed its effectiveness and limitations.

3.2 Datasets

3.2.1 Transformer Model

The GoEmotions dataset offers various genuine emotional expressions with its 58,000 human-annotated Reddit comments across 28 categories. It is ideal for training our transformer model to detect and analyze emotions in text accurately.

3.2.2 CNN Model

We have trained our model on the Fer2013 dataset, comprising 35,886 images separated into 7 emotion categories. The primary reason for using the Fer2013 dataset is that it contains relatively low-resolution photos, making it a challenging and valuable dataset to train our image emotion classifier model to detect emotions. The CK+ images-based dataset contains 981 images separated into 7 emotion categories. Training on the Fer2013 dataset and testing a more controlled, specific dataset like CK+ has helped us understand the model's capability to generalize unseen and real-world data.

3.3 Metrics

3.3.1 Transformer & CNN Model

In assessing our transformer and CNN model for emotion classification, we employed key metrics like the F1 Score and ROC AUC, each providing vital insights into the model's performance. The F1 Score, combining precision and recall, is crucial for imbalanced datasets typical in emotion classification, ensuring a balanced evaluation of the model's ability to identify each emotion accurately. ROC AUC complements this by measuring the model's proficiency in distinguishing between overlapping emotion classes at various thresholds. Collectively, these metrics offer a detailed assessment of the model's classification capabilities, which is crucial for evaluating the performance and generalization of our models.

4 Approach

4.1 Experiments Conducted

4.1.1 Transformer Model

A. Removing Non-target Emotions

In this experiment, we refined our dataset by removing emotions that do not directly correlate with mental health based on our research. Subsequently, we eliminated instances from the dataset that lacked any positive classification with the target emotions. This process inherently streamlined the classification task for the model by reducing the number of emotion classes it needed to distinguish. Reducing class complexity directly contributed to improved model performance, as the model could focus on a more targeted set of emotionally relevant labels.

B. Dataset Cleaning & Augmentation

Our dataset was subjected to cleaning and augmentation processes. We used regex operations to remove special characters and converted emojis to text for clearer emotional context, aiming to refine the data for model training. Synonym replacement was also used to augment underrepresented emotions. However, these methods did not enhance performance; data cleaning might have eliminated subtle emotional nuances, and synonym replacement could have reduced the original text's emotional intensity or context misalignment, thus not effectively improving the model's learning.

C. Transformer Comparative Analysis

We conducted a comparative analysis of various transformer models to identify the most effective architecture for our task. We evaluated several models, including BERT, DistilBERT, RoBERTa, T5, XLNet, and ELECTRA. Our findings indicated that RoBERTa outperformed the others, introducing a 1% improvement over the second highest performing model. RoBERTa's architecture, presumably due to its optimized attention mechanisms and extensive pretraining on diverse language corpora, demonstrated enhanced proficiency in capturing the nuances of emotional expression, making it the optimal choice for our emotion classification task.

D. Hybrid Models

In this experiment, we experimented with "DistilBERT + Bi-LSTM" and "RoBERTa + BiLSTM" hybrid architectures, aiming to enhance transformer models by integrating Bi-LSTM and attention layers to utilize pre-trained transformer weights for effective sequential context processing. However, these hybrids didn't surpass the performance of standalone DistilBERT or RoBERTa models, possibly due to increased complexity and overfitting, redundancy in sequence processing by transformers, and integration challenges with the hybrid architectures.

4.1.2 CNN Model

A. Dataset Augmentation and Preprocessing

We use data augmentation techniques like randomly rotating images by 10 degrees, shifting images horizontally and vertically, and zooming images in and out by 10%, respectively. These techniques

will help the model learn spatially invariant features and enable the recognition of features at different scales. We perform a normalization step on training, testing, and validation data to rescale image pixel values from the [0, 255] to [0, 1] range to help make the model training more stable and converge faster. We are converting class labels to one-hot encoded vectors as input for our model. We have tried advanced face recognition algorithms like the haar cascade classifier and MTCNN (Multi-task Cascaded Convolutional Network) as part of our preprocessing to locate and isolate facial features from images.

B. Custom VGG-16 architecture with Transfer Learning

We use a transfer learning technique with a pre-trained Vgg16 on ImageNet as our base model and add our own custom layers for image emotion classification on the Fer2013 dataset. Pretrained VGG16 has learned a robust set of features from ImageNet. Using a pre-trained model, we leverage the learned features, reducing the need for a large dataset and extensive computational resources. We are setting the first 11 layers of the VGG 16 model to be non-trainable. After our base model, we added batch normalization, which normalizes the activations from the previous layer.

Gaussian noise with a standard deviation of 0.01 was added to make the model more robust to noise in input data. The Global Average pooling 2D layer performs spatial average pooling on the previous layer's feature maps. The flatten layer converts the pooled feature maps into a single vector. After this, We added the First Fully connected layer with 256 neurons and L2 regularization with a penalty coefficient of 0.001 on both the kernel and bias weights to prevent overfitting. After this, we added batch normalization and dropout layer with a rate of 0.5 to improve generalization. The Second Fully Connected layer has 128 neurons and uses the same L2 regularization. After this, we have added batch normalization, a dropout layer with a 0.5 rate, and a final softmax activation layer, which gives the probability distribution over 7 classes.

C. Training and Inference

We are training our model for 30 epochs using categorical cross-entropy loss and Adam optimizer with a learning rate of 0.0001 along with hyperparameters like beta 1 as 0.9 and beta 2 as 0.999 that control the decay rates of the moving averages of the gradient and its square, which are used for momentum and scaling the step size respectively. We are creating a dictionary to handle class imbalance with each class ID mapped to a weight. The weights are calculated based on the frequency of each class. Our idea is to give more weight to less frequent classes so that our model does not become biased towards more frequent classes.

D. Experiments (other approaches)

We have tried different model architectures like Resnet50, custom cnn, vgg16, vgg19 architectures. The two Resnet50 architectures gave validation accuracies of 46.61% and 40.86%, custom CNN architecture gave a validation accuracy of 63.62%, vgg19 with custom layers gave a validation accuracy of 32.35%. Our custom fine-tuned Vgg16 with the base model as pre-trained Vgg16 and custom layers on top of it has achieved a validation accuracy of 90.33% and an F1 score of 66.85%.

4.2 RESULTS

For the transformer model, the first benchmark we are using to evaluate our model is from the paper Demszky et al. (2020), which gave an F1-score of 46%, and the second benchmark from the paper Cortiz (2021), which gave an F1 score of 49%.

For the CNN model, the first benchmark we are using to evaluate our model is from the paper Khanzada et al. (2022), which gave an accuracy of 75.8%. The second benchmark is from the paper Khaireddin & Chen (2015), which gave an accuracy of 73.28%.

Model	Val F1	Val ROC AUC	Val Acc	Train Runtime (sec)
DistilBERT	58.07	74.76	49.80	7,580.53
DistilBERT + Preprocessing	54.37	72.31	45.63	7,602.56
BERT	58.63	75.00	50.54	14,832.89
RoBERTa*	59.03	75.18	50.97	15,301.16
T5	58.50	74.45	48.79	36,087.32
XLNet	57.60	74.59	49.74	15,099.63
ELECTRA	58.18	74.34	49.37	3,171.95
DistilBERT + Bi-LSTM (Hybrid)	57.68	74.36	49.98	5,927.86
RoBERTa + Bi-LSTM (Hybrid)	56.63	92.76	47.79	15,473.79

*Note: "Preprocessing" refers to cleaning data, converting emojis to text, and data augmentation. Additionally, all models remove non-target emotions.

Table 1: Transformer Model Performance Comparison

Model	Train Acc	Validation Acc	Train Runtime (sec)
Fine Tuned VGG16 + Custom layers	91.75	90.33	6,176
Fine Tuned VGG16 + Custom layers + MTCNN	91.57	90.49	4,980
Custom CNN Architecture	82.82	63.62	68,400
Resnet50 + custom layers (v2)	51.85	47.26	7,827.80
Resnet50 + custom layers (v1)	51.52	40.86	7,200
VGG19	32.19	32.35	2,800

*Note: "MTCNN" refers to an advanced facial recognition algorithm used during preprocessing steps

Table 2: CNN Model Performance Comparison

5 Conclusion

In conclusion, our study involved comprehensive evaluations and enhancements of transformer and CNN models for emotion classification in text and facial expressions. For the transformer model, we focused on architectures like BERT, DistilBERT, RoBERTa, T5, XLNet, and ELECTRA, integrating Bi-LSTM to improve temporal text data processing. A key strategy was reducing emotion classes in the GoEmotions dataset to those most relevant to mental health, leading to more effective training. Our optimized RoBERTa model notably achieved a 59.03% validation F1 score, a significant 10% increase from the benchmark. In the CNN models, we experimented with various architectures on fer2013, incorporating data augmentation, MTCNN facial recognition, and advanced transfer learning. Our fine-tuned Vgg 16 model with custom layers emerged as the top performer, achieving a validation accuracy of 90.33%, considerably higher than the 73% benchmark noted in previous studies. This model also recorded an F1 score of 66.85% and a validation F1 score of 60.33%, underscoring its efficacy in practical applications.

6 Future Work

In our future work, we aim to advance the capabilities of transformer and CNN models and develop a multimodal application. For the transformer model, we'll explore alternative hybrid architectures, attention visualization for better interpretability, the creation of a novel, diary-style emotional dataset, and cross-domain validation. These efforts will deepen our understanding of complex language patterns in mental health contexts. For the CNN model, we plan to employ techniques like StyleGAN2 and stable diffusion for data augmentation, utilize the AffectNet dataset, and implement advanced facial recognition algorithms to enhance data quality. Ensemble learning with multiple CNN architectures and testing diverse image datasets will improve performance and generalization. Lastly, a future multimodal application will fuse the transformer and CNN models to analyze emotional cues from both text and visual data, offering a comprehensive mental health assessment tool that leverages both models' strengths for a richer, more holistic understanding of emotional states.

REFERENCES

Diogo Corti. Exploiting joint learning of chest radiographs and clinical data for improved covid-19 diagnosis. 2021.

Diogo Cortiz. Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. 2021.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. 2020.

Y. Khaireddin and Z. Chen. Facial emotion recognition: State of the art performance on fer2013. 2015.

Amil Khanzada, Charles Bai, and Ferhat Turker Celepcikay. Facial expression recognition with deep learning: Improving on the state of the art and applying to the real world. 2022.

Tomoki Kusunose, Xin Kang, Keita Kiuchi, Ryota Nishimura, Manabu Sasayama, and Kazuyuki Matsumoto. Facial expression emotion recognition based on transfer learning and generative model. 2022.

Sanghyub John Lee, JongYoon Lim, Leo Paas, and Ho Seok" Ahn. Transformer transfer learning emotion detection model: synchronizing socially agreed and self-reported emotions in big data. 2023.

Shan Li and Weihong Deng. Deep facial expression recognition: A survey. 2018.

Hendri Murfi, Syamsyuriani, Theresia Gowandi, Gianinna Ardaneswari, and Siti Nurrohmah. Bert-based combination of convolutional and recurrent neural network for indonesian sentiment analysis. 2022.

Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: State of the art. 2018.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. 2020.