CMPE 492

Improving Facial Emotion Recognition with Word Embeddings

Furkan Bülbül

Advisor:

İnci Meliha Baytaş

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Broad Impact

Facial emotion recognition is an important aspect of human-computer interaction, with applications in a variety of fields such as healthcare, marketing, and security. Its effective implementation promotes a better understanding of human emotions, resulting in more personalized interactions with technology.

The capacity to correctly interpret facial emotions has profound implications in a variety of fields, including mental health and education. Recognizing emotional information at the right time can help with early treatments in mental health illnesses and improve the effectiveness of educational interventions focused on to certain emotional states. However, existing facial emotion recognition models struggle to accurately detect. One of the mostly used dataset for emotion recognition is that FER2013[1] which have limited accuracy.

1.2. Ethical Considerations

Most significant worry is privacy and permission, as the use of face data for emotion recognition may violate people's privacy rights. Nowadays, GDPR issues arises one by one, inspecting the face images seems to be risky about GDPR rights of people. Explicit consent from participants regarding the collection and use of their facial data are critical. Even if the consent taken from participant, public usage of facial emotion dataset have its own risk factors. Furthermore, training model on the facial data have more to consider. Training model that may result in discriminative behaviour seems possible, and should be prevented somehow.

[1] `https://www.kaggle.com/datasets/msambare/fer2013/data`

2. PROJECT DEFINITION AND PLANNING

2.1. Project Definition

We realized that, even with the advances in neural networks for image classification, there is still a lack of accuracy on some facial expression data, particularly in the FER2013 dataset. The project will primarily focus on integrating word embeddings into the existing facial emotion recognition models, CNN (VGG and ResNet). The projects employs the pre-trained word embedding model GloVe. The main motivation of ours is to construct a model within the domain of deep learning, combining image feature extraction techniques with word embeddings.

I have the following deliverables:

- Implementation of base image classification models VGG, and ResNet.
- Implementation of multimodal model for classification.
- Evaluation of those image classifier models on the FER2013 dataset.
- Documentation and implementation of architecture of the proposed model that integrates word embeddings for facial emotion recognition.
- Documentation of experimental setup, including data pre-processing steps, hyperparamater tuning, and evaluation in general.
- Evaluation of gain by utilizing word embeddings in image classification task.

2.2. Project Planning

2.2.1. Project Time and Resource Estimation

I estimate the time and resources required for each step of the project.

- Researching on image classifier models and literature review: 3 weeks

- Implementation and training of base image classifiers: 2 weeks
- Developing the architecture with word embedding and image embedding: 3 weeks
- Rvaluation and analysis: 3 weeks
- Documentation and reporting: 2 weeks

Resource that I need throughout the project development is a high performance computing that is provided by the department.

2.2.2. Success Criteria

Achieving an improvement in facial emotion recognition accuracy on FER2013 dataset compared to the state of art models. Also, receiving positive feedback from my advisor is a valuable criteria for the project's success.

2.2.3. Risk Analysis

There are the following problems encountered:

- Data quality issue: It could be possible that FER2013 dataset may be insufficient to build a highly acceptable model because of the size of the dataset. However, augmentation techniques are planned to use in the project.
- Technical challenges: In addition to deep learning implementation and training, I faced other tasks such as Dockerization, Wandb integration, and running Slurm jobs were the technical side of the project.
- Resource constraints: I have accessed enough GPU power to train models for this project. But since it is a common resource for the department, there was times that I had to wait for the queue.

3. RELATED WORK

There are few public datasets for face expression due to privacy issues. Despite this limitation, the FER2013 dataset offers researchers a useful resource for furthering the science of facial expression recognition. Several cutting-edge models in facial emotion recognition have been trained on the FER2013 dataset, making it a general benchmark, as I intend to do.

Advancements in deep learning models is reflected on FER2013 dataset. There are different deep learning model architectures that aims to improve accuracy further.

CNNs have performed remarkably in many image classification problems. Researches have benefitted CNN-based models to achieve significant improvement of recognition accuracy. Pramerdorfer and Kampel [1] compares and points out the different performances of CNNs. They mention major bottlenecks in that is encountered in FER problem, the most concerned issue is that there is not publicly available large dataset. They perform various tests on different CNNs. According to their results, even though VGG model has less parameters and less depth, it performs better compared to ResNet and Inception. That behaviour is observable in my results.

Khaireddin et al., [2] employs VGG11 architecture. The researches highly used data augmentation to overcome the issue with the dataset size. They conduct fine-tuning to find optimal hyperparamaters and optimizers. It is claimed that they could get 73.28% accuracy rate by training only on the FER2013 dataset. Additionally, Negara et al. [3] adopts VGG architecture to develop the accuracy on the FER2013 dataset. They uses more convolutional layers in the feature extractor and only one classifier layer. After conducting many experiments, they achieve 69.46% accuracy rate on the dataset.

While word embeddings have shown promise in a variety of natural language

processing tasks, their application to facial emotion recognition (FER) remains largely unexplored. One study proposes a similar architecture to CLIP for FER. FER-former [4] model is proposed FER focused transformer mechanism.

Akata et al. [5] proposed a function that measures the compatibility between an image and label embedding. While they work with textual attribute embeddings, they also used label embeddings. This article was one of the support point of the idea to use label embeddings.

As mentioned in methodology, I adopted contrastive loss that is introduced in SimClr [6]. They train a self-supervised model using only visual dataset. I employed a multimodal variant of this loss function, which may result in an asymmetric loss across modality similarity. It is used in the ConVirt [7] which is a self-supervise multimodal learning for contrastive learning for medical paired images and texts.

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}[k \neq i] \exp(\text{sim}(z_i, z_k)/\tau)} \tag{3.1}$$

4. METHODOLOGY

4.1. Datasets

FER2013 data is publicly available it will be used as a benchmark in the evaluation process. It is commonly used as a benchmark in facial emotion recognition task.

Training a model with FER2013 dataset suffers from overfitting. As mentioned in the literature review section, data augmentation is employed by many studies. So, it is crucial to implement effective data augmentation. Porcu, Simone et al. [8] have studied data augmentation techniques that improves facial emotion recognition systems. So, I have adapted their proposal except Generative Adversarial Network since it becomes very costly to have GAN model to increase the data size. I plan to use data augmentation for further. Those techniques are random rotation, horizontal reflection, cropping, translation, and resizing.

4.2. Model Architecture

Figure 4.1 represents the high level architecture of our model. An image embedding takes an image as input, extracts its key features and then converts those features into a numerical code (embedding vector). Pre-trained word embedding model works as the same way, it creates the vector representation of the words fed into the model. Output of those two models are represented in the different spaces. Therefore, model requires mapping models to have those models into the same space. Once, they are represented in the same space, loss calculation and back propagation will train the Image embedding and mapping models.

I used two different image embedding model, VGG and ResNet, that is trained on our data. As pre-trained word embedding model, I used [9] GloVe.

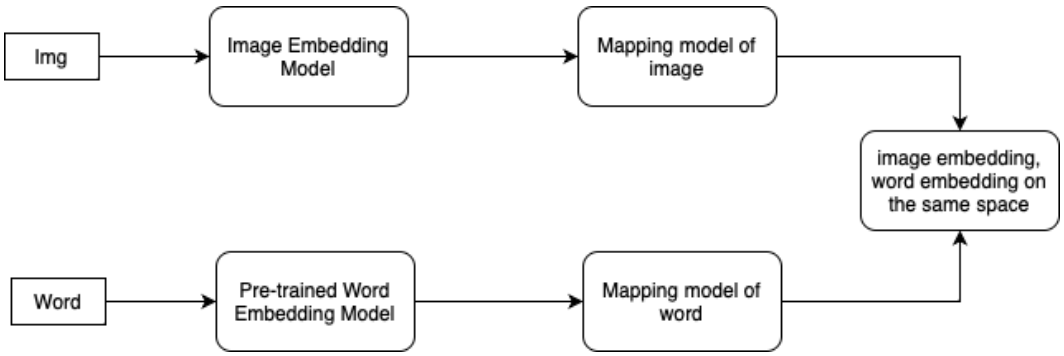

Figure 4.1. High Level Schema of the model

4.3. Training Procedure

First, the dataset will be divided into three subsets: training, validation, and test sets. Model is trained using training set. Adam is used as optimizer with different hyperparamaters during experiments.

4.3.1. Contrastive Multimodal Training

The process of training is shown in Figure 4.2. Steps are as follows:

- Image-text pairs are fed into encoders.
- Image encoder extracts feature vector.
- Text encoder extracts the word embedding of text data.
- MLP layer maps feature vector to common space.
- Cross entropy loss calculated based on similarities.

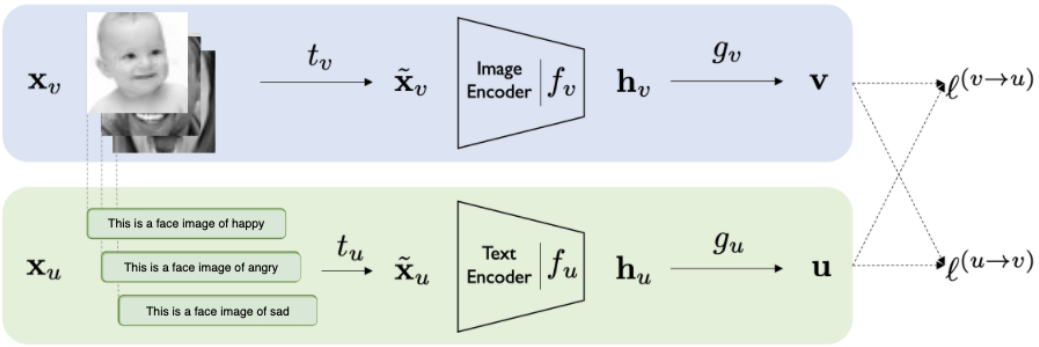

Figure 4.2. Multimodal Contrastive Learning

The loss function is consists of two components. Total loss is weighted average of text-to-image contrastive loss and image-to-text contrastive loss.

$$\ell_i^{(v\to u)} = -\log \frac{\exp\left(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau\right)} \tag{4.1}$$

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{(v\to u)} + (1-\lambda) \ell_i^{(u\to v)} \right) \tag{4.2}$$

4.3.2. Classification

During multimodal contrastive training, image feature extractor is already trained. I also, wanted to train the classifier part of the original image classification model (VGG and ResNet). I fine tuned the model, with cross entropy loss and completed the training. I have experimented with two variations. Firstly, freezing the CNN layers and training only FC layer. Also, fine tuned the CNN after multimodal training. I found the almost identical results for both options. Figure 4.3 shows the overall architecture including the classification (fine tuning).

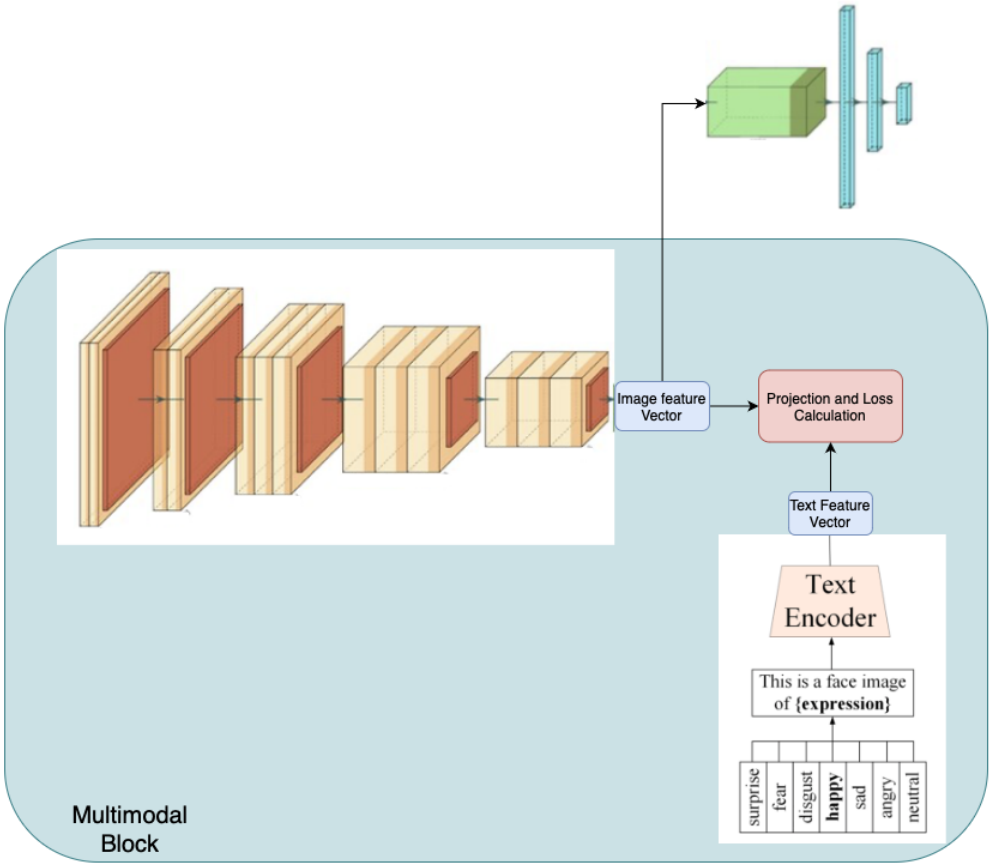

Figure 4.3. Complete architecture

4.4. Evaluation Metrics

I have evaluated the trained model using standard metrics such as accuracy and confusion matrix on test set. I have tried different hyperparameters.

- Learning Rate
- λ : weight paramater of loss (equation 4.2)
- dropout
- scheduler
- L2 Regularization
- MLP dimensions
- Word Embedding Dimension

5. REQUIREMENTS SPECIFICATION

5.1. Functional Requirements

- **Input Requirements:** The system should accept input data in the form of images containing human faces and corresponding textual descriptions of emotions.
- **Output Requirements:** The system should provide predicted emotion labels for each input image.
- **Processing Requirements:** The system should perform data preprocessing steps, such as resizing and normalization, before feeding the data into the model for inference. It should also handle the integration of word embeddings into the model architecture.
- **Feature Requirements:** The system should support multi-modal input processing, allowing for the simultaneous processing of image and text data.

5.2. Non-Functional Requirements

- **Performance Requirements:** The system should achieve high accuracy in emotion recognition, catching the state-of-the-art.

6. IMPLEMENTATION AND TESTING

6.1. Implementation

Pytorch is used as the primary implementation tool in the project. Source code is mainy consist of the following

- Dataset classes: used to create data loader.
- Model classes: represents the architecture of Neural Network.
- Trainer classes: handles forward pass, loss calculation and back-propagation.
- Runner script: uses whole classes to train and test the model.

For inspecting the behaviour of the process of training created log files are visualized with Wandb

6.2. Testing

Models are tested with the test dataset that is partitioned from the whole data. Test dataset is not fed into the model until training completes.

6.3. Deployment

Models are trained on the HPC of the department. Docker image created, and model is run in the container of the image in the job submitted to the Slurm workload manager.

7. RESULTS

I have experimented with VGG11, VGG16, ResNet18, and ResNet50. They follows a similar training phase in terms of training and validation plots. Figures below provide an example of how the validation and training processes operate. I have searched for the optimum hyperparamaters by running the model with different variations. These are the ones I used to maximize the test accuracy.

- Initial Learning rate: 0.001
- Scheduler: Cosine Annealing Learning Rate
- λ: 0.75
- Dropout: 0.25
- MLP Dimensons: 256 in hidden layer, 64 output layer
- Word Embedding Dimension: 100

Every model has been trained over 300 epochs. I added 100 epochs to the 200 epochs of mutlimmodal training for fine tuning. I also trained the base CNN models for 300 epochs.

7.1. Multimodal Contrastive Training

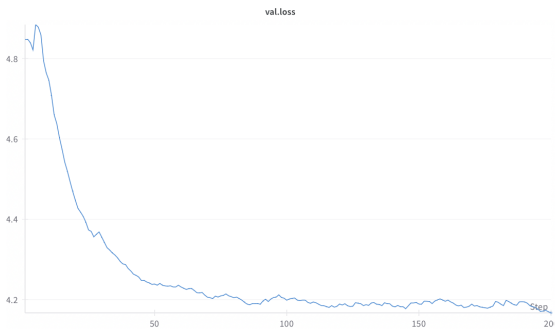

Figure 7.1. Validation Loss

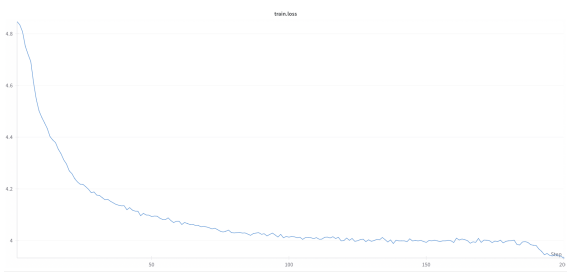

Figure 7.2. Train Loss

7.2. CNN Fine Tune

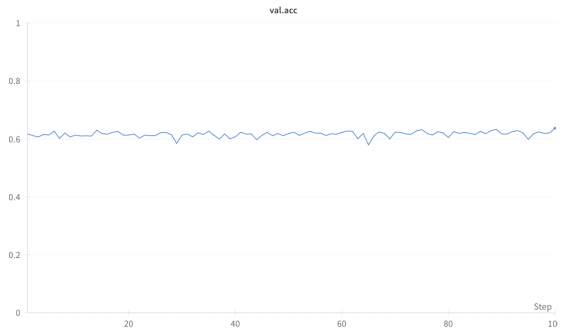

Figure 7.3. Validation Accuracy

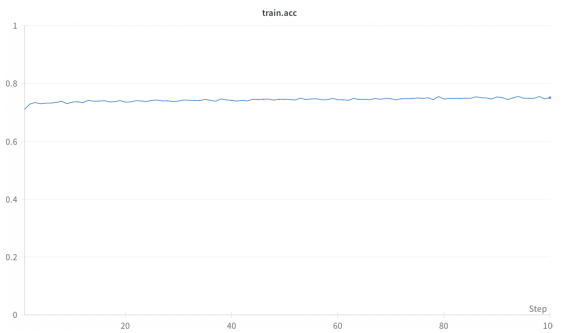

Figure 7.4. Training Accuracy

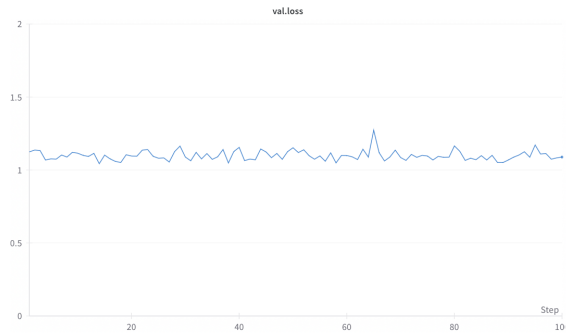

Figure 7.5. Validation Loss

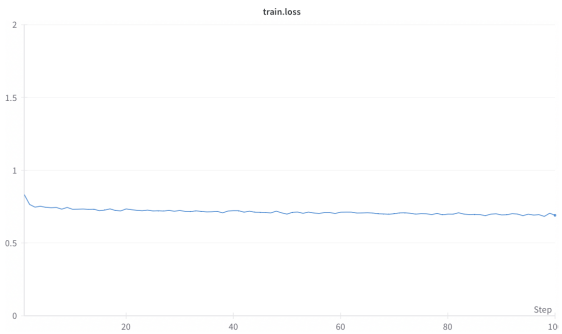

Figure 7.6. Training Loss

CNN	Type	Acc (%)
VGG11	Multimodal	65.25
	Base	63.58
VGG16	Multimodal	64.34
	Base	63.12
ResNet18	Multimodal	62.20
	Base	61.04
ResNet50	Multimodal	60.04
	Base	60.89

Figure 7.7. Comparison of accuracy improvement

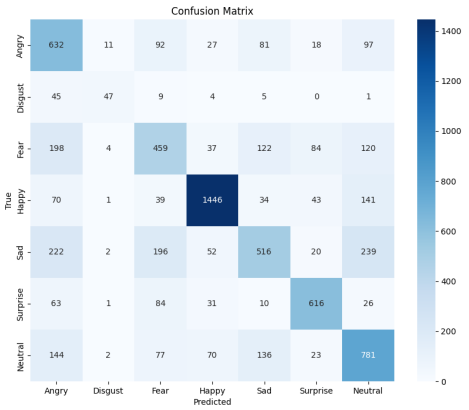

Figure 7.8. Confusion matrix of multimodal training with VGG11

The results show that multimodal training slightly improves CNN models' accuracy. The integration of word embeddings, increases the meaningful data to the model. Most of the tested models clearly showed this effect, except ResNet50. VGG models improved relatively better. As, depth of the image embedding model increases, meaninful improvement vanishes. Considering the results, they show that improvement is possible by integrating the word embeddings.

8. CONCLUSION

In this project, I aim to enhance facial emotion recognition using word embeddings to improve accuracy and robustness in emotion classification tasks. I wanted to replicate the results the studies done. Moreoever, I want to see if improvement is possible.

My results, do not seem to catch the state-of-the-art models in terms of accuracy performance. However, when compared to the performance of only CNN models, multimodal tend to perform better. Further optimization and experimentation with training techniques may result even better results. It shows promising side of the self supervised learning of multimodal models in emotion recognition tasks.

REFERENCES

1. Pramerdorfer, C. and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art", *CoRR*, Vol. abs/1612.02903, 2016, http://arxiv.org/abs/1612.02903.

2. Khaireddin, Y. and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013", *CoRR*, Vol. abs/2105.03588, 2021, https://arxiv.org/abs/2105.03588.

3. Kusuma Negara, I. G. P., J. Jonathan and A. Lim, "Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16", *Advances in Science, Technology and Engineering Systems Journal*, Vol. 5, pp. 315–322, 01 2020.

4. Li, Y., M. Wang, M. Gong, Y. Lu and L. Liu, "FER-former: Multi-modal Transformer for Facial Expression Recognition", , 2023.

5. Akata, Z., F. Perronnin, Z. Harchaoui and C. Schmid, "Label-Embedding for Image Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 7, p. 1425–1438, Jul. 2016, http://dx.doi.org/10.1109/TPAMI.2015.2487986.

6. Chen, T., S. Kornblith, M. Norouzi and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations", , 2020.

7. Zhang, Y., H. Jiang, Y. Miura, C. D. Manning and C. P. Langlotz, "Contrastive Learning of Medical Visual Representations from Paired Images and Text", , 2022.

8. Porcu, S., A. Floris and L. Atzori, "Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems", *Electronics*, Vol. 9, 11 2020.

9. Pennington, J., R. Socher and C. D. Manning, "Glove: Global Vectors for Word

Representation.", *EMNLP*, Vol. 14, pp. 1532–1543, 2014.