Improving Multimodal Sentiment Analysis Using Multilogue-Net and Transformer Architecture

Michelle Feng
Carnegie Mellon University
msfeng@andrew.cmu.edu

Marian Qian
Carnegie Mellon University
marianq@andrew.cmu.edu

1 Introduction

This project aims to perform multi-modal sentiment analysis. Sentiment analysis as a general field involves the extraction of features to determine underlying emotional tones. As an open ended field, both the set of potential inputs and outputs to problems in this space vary. Sentiment analysis problems can categorize towards binary outputs (think positive vs. negative), numerous categories, or even take the form of sentiment regression problems with continuous outputs. Sentiment analysis can target text, audio, and video inputs, as well as any other data forms that can reflect connotations.

Multimodal sentiment analysis has many modern applications. [1] With tons of user focused data made available by the Internet of Things and the internet, the use of sentiment analysis has the ability to shape the future of many products and services. [2] Since sentiment analysis is a relatively new research area, multimodal applications have much room for improvement.

This project is geared towards improving existing models on multi-categorical emotion classification for video/text applications. Using the CMU-MOSEI dataset [3] (shortened to MOSEI for the rest of this paper), we intend on leveraging sentence utterance videos and transcriptions to classify towards sentiment values in the range $[-3, 3]$. Note that although this paper classifies from MOSEI, we also draw motivation from previous work targetting the CMU-MOSI dataset (MOSI), which is a smaller scale multimodal sentiment analysis dataset that also includes video, audio, and text modalities.

2 Background

The MOSEI dataset we use involves text, acoustic, and visual information. Multimodal research involve how to best encode and combine (fuse) these different modalities together so that the model can better extract useful information during training and inference. Other deep learning models such as convolutional neural networks, multi-layer perceptron layers, gated recurrent units, and transformers are used often as encoders to take the basic raw input features and project them to a lower dimensional feature dimension that would ideally be a better representation of the input data. Regarding the fusion of modalities, different fusion paradigms like early fusion (concatentation before passing inputs through a model) and late fusion (concatentation of features passed through an encoder) are an area of study as well [8]. The question of how to best combine the different modality features and how to process them after they are combined, whether through attention or another deep learning model also impacts how well the multimodal model performs.

3 Related Work

Existing literature using the CMU-MOSI and CMU-MOSEI datasets have applied a wide range of technologies towards assigning accurate sentiment values towards video clips.

For example, the space can be an application of recurrent neural networks (RNN). An example of such an architecture is *Multilogue-Net* [1], which is a network comprised of GRUs for state and

10617 Project Final Report (2023).

emotion for every modality and GRUs for context for every modality. The novelty of Multilogue-Net is the better leveraging of all available modalities resulting in improved context representation (there is a GRU for context that only takes in text representation, which may be the reason for improved context representation, since context of a video is often captured better through text). Multilogue-Net then successfully captures modality representations through using individual GRUs for each modality and successfully captures the relationship between emotion, context, and states of the modalities themselves by concatenating state vectors with context vectors before passing through the context GRU, performing simple attention across all context vectors before passing through the state GRU, and passing through state vectors and emotion vectors through the emotion GRU. To determine the final prediction of sentiment analysis, Multilogue-Net uses pairwise attention across the emotion vectors for each modality and concatentates them with the original emotion vectors for each modality before passing them through a multilayer perceptron. Multilogue-Net has been tested on the MOSI dataset and has reached an accuracy of 81.19; on the MOSEI dataset, it reached an accuracy of 82.10 and a mean absolute error (MAE) of 0.59.

The *UniMSE* [7] model is one of the first models to unify the features used for emotion recognition and sentiment analysis. The architecture includes two LSTMs for audio and video modality inputs and uses a T5 (text to text transformer) that uses the features from the other modalities. UniMSE also uses pretrained modality fusion by injecting the audio and video features into deeper T5 layers so that the "acoustic and visual signals can participate in text encoding and are fused with multiple levels of textual information" [7] and so that the audio and video features can "probe relevant information in the pre-trained text knowledge" [7] of the T5. In addition, the T5 performs contrastive learning on the features right before the decoder. Through UniMSE's ablation studies, the authors found that pretrained modality fusion and contrastive learning both help improve the perofrmance of the model. UniMSE has been tested on the MOSEI dataset has an accuracy of 87.50 and a mean absolute error of 0.523 (the highest ranking on the MOSEI dataset on Papers with Code website).

We leverage *MultiBench* [8], a large-scale benchmark that provides an end-to-end machine learning pipeline designed for standardized data loading and experiments. MultiBench offers a standard for the baseline and architectures for further development and improvements to our model.

4 Methods

Baseline: The baseline that we implemented is a standard fusion model (a neural network model that involves "fusing" together the inputs or features of different modality inputs). We found a Google Colab notebook using MultiBench [6] that implemented this model and trained it on the CMU-MOSI dataset. This fusion model in particular concatenates the inputs together (called early fusion) and then has sequential layers of a GRU (gated recurrent unit) and a regular MLP on top of the inputs. We trained the model for 100 epochs and used the Adam optimizer with a starting learning rate of 0.001. The results of the model are in the next section.

Method 1: Multilogue-Net with Transformers (TransMultilogue)

Our main method combines the GRU network from Multilogue-Net with the use of a pretrained transformer. The pretrained transformer takes takes in the emotion vectors from each modality, concatenates them, and outputs a hidden state. We then combine the transformer's hidden state with the original emotion states, pass the features through a multi-layer perceptron, and output our final sentiment regression value.

The general motivation of this method is to utilize the pretrained text knowledge of a pretrained transformer. UniMSE uses LSTMs to encode data from audio and visual modalities and uses the shallower layers of the T5 to encode the text modality, and allowing all of these encoded features to interact with each other and with previous textual information, which resulted in UniMSE listed as the highest performing model on MOSEI. We decided to use Multilogue-Net's network of GRUs, where the encoder GRUs are specific to not just the modality, but also to a certain aspect of the modality like emotion or context. To include behavior of UniMSE of modality features interacting with each other, we included the use of a pretrained transformer with the input to the transformer as the fusion of the emotion vectors from the emotion GRUs (so we replaced the pairwise attention mechanism with this pretrained transformer). We did not perform pretrained modality fusion directly where we injected layers of the transformer with audio and visual features due to time and resource constraints, but our

method of feeding in a concatenated input of features of different modalities achieves relatively the same idea as concatenating these features later on into the deeper layers of the transformer.

The UniMSE paper mentioned a concern about only injecting audio and visual features into deeper layers of the transformer, as the model also used the T5 transformer as an encoder for the text modality. Their concern was if audio and visual features were injected too early, then the text encodings would be disturbed too much by other modalities. For our method, we keep the state, emotion, and context GRU for the text modality, so we already encoded the text modality before concatenating it with other modalities and feeding it into the transformer.

UniMSE used the commonly known T5-base transformer with over 220 million parameters and 12 transformer layers. Due to resource limitations as we were using Google Colab to train our model, we decided to use a smaller pretrained model instead; we used the pretrained Reformer model found on Hugging Face [9] which has around 3 million parameters. This transformer model was also fine-tuned during training and it's last hidden state was used to predict our target sentiment analysis value.

During training of our method, we used the same hyperparameters to train as used for Multilogue-Net.

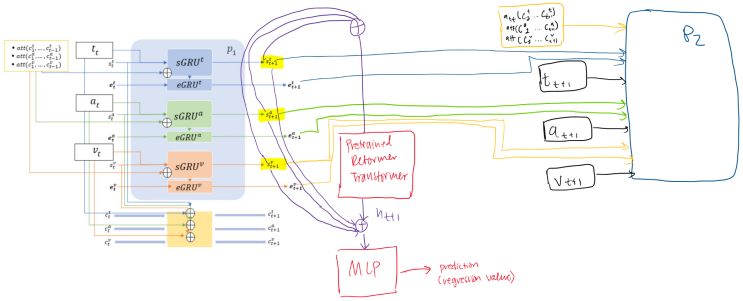

Figure 1: Representation of Multilogue-Net GRUs combined with a pretrained Reformer transformer model

5 Results

We evaluated and compared models using mean squared error loss, mean absolute error, and for TransMultilogue, the Pearson correlation coefficient, which indicates the strength of linear the linear relationship (shown by values with further from 0).

Baseline

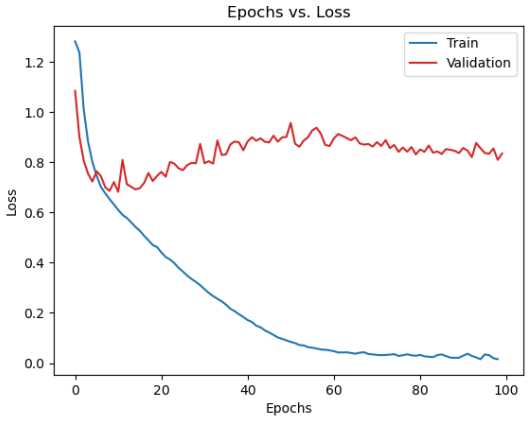

Figure 2: Baseline loss values over 100 epochs of training for train and validation datasets

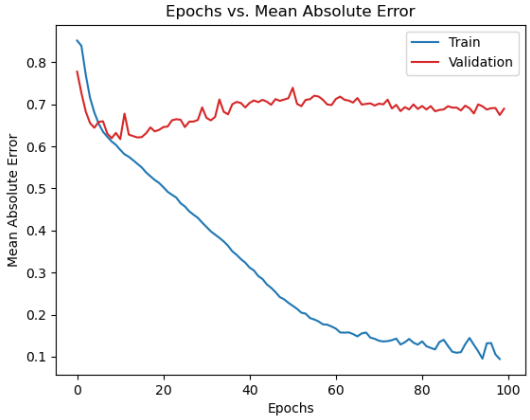

Figure 3: Baseline mean absolute error values over 100 epochs of training for train and validation datasets

The baseline model achieved a test loss of 0.8349181413650513 and mean absolute error of 0.6897171933927467. The baseline model demonstrates clear overfitting, as shown by the improvement of loss and error for training data, but quick rise and plateau of the same metrics for validation sets. Furthermore, the training data demonstrates smooth decrease, while validation results are relatively noisy, indicating further weakness of this model. This aligns with expectations of this model, as a relatively simple approach to capturing the nuances of multimodal data.

Multilogue-net with Transformers (TransMultilogue)

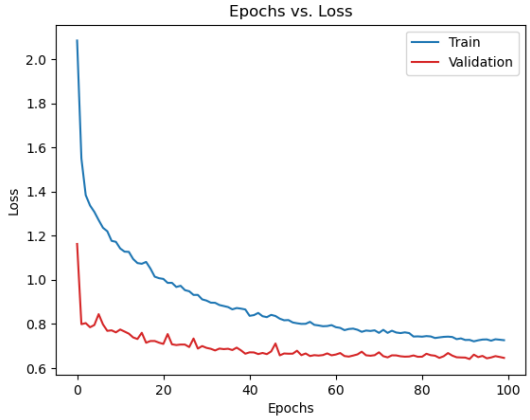

Figure 4: TransMultilogue loss values over 100 epochs of training for train and validation datasets

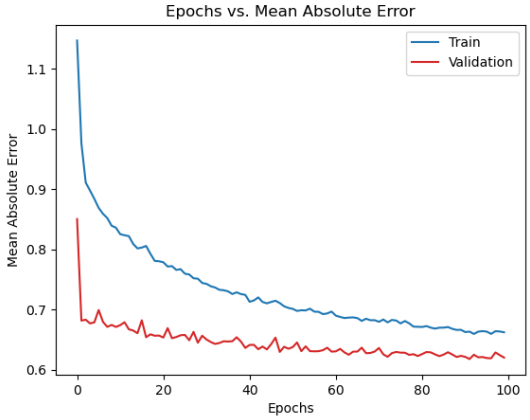

Figure 5: TransMultilogue mean absolute error values over 100 epochs of training for train and validation datasets

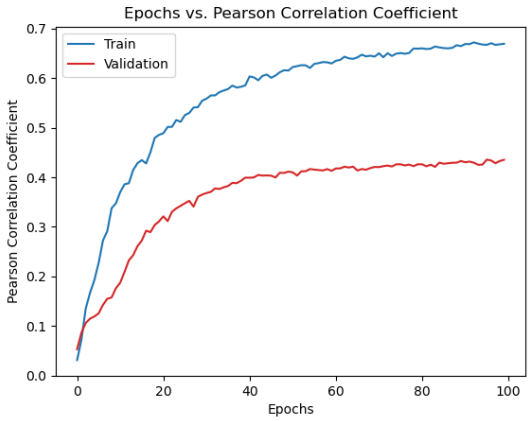

Figure 6: TransMultilogue Pearson correlation coefficients over 100 epochs of training for train and validation datasets

Our method achieved a test loss 0.6414 and mean absolute error of 0.6176999807357788. For loss and mean absolute error, our validation demonstrated consistently lower values compared to the training values. This can be attributed to factors related to data diversity and representation. This could indicate that the validation data is well represented by the training data. Especially since the drop in error and loss happens almost immediately to a point where it decreases much more gradually, it is possible that the validation data happened to be easily predictable (less fluctuation/outliers). Since the training and validation splits were produced by the MOSEI dataset explicitly, this is a factor out of the authors' control.

6 Discussion and Analysis

Our method did not achieve as low of a mean absolute error as Multilogue-Net, which achieved a value of 0.59, compared to our value of 0.617. Using a larger, and therefore more comprehensive, transformer model, may help with producing features that have better textual understanding as the Reformer model we chose was only trained on the Crime and Punishment novel by Fyodor Dostoyevsky. However, that would be assuming lack of textual understanding is the reason why our model did not perform as well as Multilogue-Net; there could be other reasons, like how fusing the layers before passing through the transformer did not lead to enough interactions between modality as compared to pretrained modality fusion where audio and visual features are injected directly to

individual transformer layers. Future steps would be to implement pretrained modality fusion as described in UniMSE, even though that process would be more involved.

Another note is that our last hidden state from the transformer was not passed along to the next timestep; we only used the interactions of the modalities at the current timestep to calculate our sentiment analyis prediction. Including the transformer's previous hidden state during the computation for our sentiment analysis could potentially represent context information better as well as we are keeping the interactions of modalities from previous timesteps.

7 References

[1] Shenoy, A., & Sardana, A. (2020). Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. In Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML) (pp. 19–28). Association for Computational Linguistics.

[2] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, 424–444. doi:10.1016/j.inffus.2022.09.025

[3] Zadeh, A., Liang, P., Poria, S., Vij, P., Cambria, E., & Morency, L.P. (2018). Multi-attention recurrent network for human communication comprehension. In Thirty-Second AAAI Conference on Artificial Intelligence.

[4] Cai, Z., Ghosh, S., Stefanov, K., Dhall, A., Cai, J., Rezatofighi, H., Haffari, R., & Hayat, M. (2023). MARLIN: Masked Autoencoder for facial video Representation LearnINg. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1493-1504). IEEE.

[5] Vasco Lopes, Antonio Gaspar, Luis A. Alexandre, & Joao Cordeiro (2021). An AutoML-based Approach to Multimodal Image Sentiment Analysis. In 2021 International Joint Conference on Neural Networks (IJCNN). IEEE.

[6] MultiBench Example Usage Colab. MultiBench 1.0.0 documentation. (n.d.). https://colab.research.google.com/github/pliang279/MultiBench/blob/main/examples/Multibench_Example_Usage_Colab.ipynb

[7] Hu, G., Lin, T.-E., Zhao, Y., Lu, G., Wu, Y., & Li, Y. (2022, December). UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 7837–7851. Retrieved from https://aclanthology.org/2022.emnlp-main.534.

[8] Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., ... Others. (2021). MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1).

[9] Reformer Model. (n.d.). Retrieved 2 December 2023, from https://huggingface.co/docs/transformers/v4.35.2/en/model_doc/reformertransformers.ReformerModel

8 Appendix

Code for the project