Affective Emotional Layer for Conversational LLM Agents

Stanford CS224N Custom Project

Aditya Bora
Department of Computer Science
Stanford University
adibora@stanford.edu

Nikhil Suresh
Department of Computer Science
Stanford University
ncsuresh@stanford.edu

Abstract

One important element to engaging conversations in building interactive systems is the ability to understand and express emotional responses and reflect a correspondingly accurate specified emotion, which remains quite difficult. As a result, our goal is to enable conversational agents to effectively produce various emotional metrics during a conversation, akin to human responses and foster more natural and realistic responses from AI. In this study, we propose a modification to existing architectures to optimize performance by exploring using alternative attention mechanisms, specifically convolutional attention. We proposed the use of convolutional attention in a transformer based language model due to its theoretical ability to better capture local dependencies in a given input. Ultimately, our results showed that multi-headed attention out performs convolutional attention in every metric regarding response quality as well as emotional realism.

1 Key Information to include

- Mentor: Tathagat Verma
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Emotion plays a pivotal role in human communication, acting as the undercurrent that shapes our interactions and conveys the depth and nuance of our messages. With the recent advancements in Large Language Models, we have models which are now capable of speaking with human level coherence. One area, however, that current AI systems have not matured in is their ability to comprehend and react appropriately to emotional subtleties. This restriction hinders their ability to achieve completely natural responses, which are crucial for a variety of different applications.

This research aims to address this limitation by exploring advanced NLP techniques to enhance the emotional intelligence of language models.

Early research regarding emotion primarily focused on sentiment analysis, where the goal was to determine which basic emotions like happiness, sadness, anger, etc. were conveyed in a text input. The learning's from research in sentiment analysis eventually led to the development of more complex models that were able to extract greater levels of emotional nuance from a given text. As focus on text generation increased, these principles from sentiment/emotional analysis were used to train models which included emotional context in generated text via a variety of different methods from fine tuning, modifications in training, additional conditioning, etc.

Our approach to enhance emotional realism involves the modification of an existing mixture of experts architecture with the use of convolutional attention. We introduce this convolutional attention

modification to the neural network architecture in order to better capture local dependencies and contextual patterns related to emotional expressions when training the model on textual data.

The successful implementation of emotionally intelligent conversational agents has far-reaching implications in domains where nuanced text-based communication is critical, such as online customer support and mental health counseling. By addressing the emotional aspects of communication, we can enhance the effectiveness and quality of these interactions, leading to improved user experiences and outcomes.

3 Related Work

The study of emotion in natural language processing originally concentrated on identifying emotions in given input texts. Initially, this involved sentiment analysis, where basic emotions like happiness or sadness were detected. Over time, the focus has shifted to more advanced techniques. Today, the primary focus of modern research is to identify a wider range of nuanced emotions from text, allowing for a deeper understanding of the emotional content in written material.

Deep learning is one approach that has proven to be effective at parsing emotion. Guo et. al implemented a deep learning assisted semantic text analysis (DLSTA) approach for human emotion detection in text and was able to reach a 97.7 percent classification accuracy. Guo (2022)

As large language models have become more advanced, there's has been growing emphasis on their ability to understand and express emotions . Several innovative methods have been developed to integrate this emotional understanding into language models. These approaches range from prompt modification to alternate model architecture and training methodologies.

Li et. al's study compellingly demonstrates that prompt modification can significantly improve language model outputs, particularly in terms of generation performance. Their innovative approach, termed "EmotionPrompt", involves augmenting original prompts with emotional stimuli, leading to notable improvements. Li et al.

Other methodology uses different learning and fine-tuning techniques to alter the model architecture itself to produce emotionally apt outputs. Shah et. al updates pre-trained network weights using contrastive learning so that the text fragments exhibiting similar emotions are encoded nearby in the representation space, and the fragments with different emotion content are pushed apart. Shah et al. (2023)

Casas et al. fine-tuned a GPT-2 model using a dataset with emotionally varied and altered data, focusing on improving the model's ability to modify the emotional intensity of input sentences. Their approach integrated a paraphrasing model with the modified GPT-2, enabling the system to alter the emotional tone of sentences while maintaining their semantic content.Casas et al. (2021)

For evaluation, these methods use a combination of different natural language processing scores including BLEU, Meteor, etc. Some papers train emotion classifiers to determine emotional accuracy of their models while others use human subjects to conduct surveys measuring emotional accuracy.

4 Approach

Main Approaches:

Baseline: The baseline architecture we are leveraging uses a basic transformer for encoding of a given input. This transformer uses standard multi-headed attention. The novelty in the emotional expressions results from the fact that the model consists of n decoders, further denoted as listeners, which are optimized to react to each context emotion accordingly. The listeners are trained along with a Metalistener that softly combines the output decoder states of each listener according to the emotion classification distribution. All the listeners are modeled by a standard transformer decoder layer block, denoted as T RSDec, which is made of three sub-components: a multi-head self-attention over the response input embedding, a multi-head attention over the output of the emotion tracker, and a position-wise fully connected feed-forward network.

Modifications to Model: We propose to modify this architecture to try and preserve quality but optimize performance. To do so we explore using different attention mechanisms which have lower

computational complexity but retain effectiveness. Specifically we replace the existing multi-head attention mechanism with a convolutional attention mechanism.

Convolutional attention mechanisms integrate the strengths of convolutional neural networks (CNNs) into attention models. Unlike traditional attention mechanisms that focus on dependencies regardless of their distance, convolutional attention leverages the local structure of data, typical in CNNs. This is particularly useful in processing data where local context significantly contributes to the understanding of each part, such as in images or in certain types of sequential data.

The convolutional attention can be described by the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + C\right)V \tag{1}$$

Here, Q, K, and V represent query, key, and value matrices respectively, similar to standard attention mechanisms. d_k is the scaling factor, typically the dimension of the keys.

In the convolutional attention mechanism, C signifies the convolutional features acquired by applying a set of convolutional filters, denoted as $F = \{f_1, f_2, \ldots, f_n\}$, on the input matrix I. Each filter f_i in F convolves across I, facilitating the extraction of localized features. The convolution operation of each filter f_i yields a corresponding feature map, given by $M_i = f_i * I$, where $*$ represents the convolution operation. These individual feature maps M_i are then cumulatively aggregated to construct a comprehensive representation of local contextual information. This aggregation, which could be a summation or other methods, results in $C = \sum_{i=1}^{n} M_i$, encapsulating the collective local contextual cues extracted by all filters.

The motivation for using convolutional attention is that it can explicitly capture local context due to its inherent nature of working with a window of tokens. This feature can be particularly beneficial for understanding emotional context, as emotions in text often depend heavily on local word groupings and their nuances.

5 Experiments

5.1 Data

As used in many studies in the topic space, we use the *Empathetic Dialogues* Rashkin et al. (2019) dataset which includes 25,000 speaker-listener open domain conversations that are grounded in emotional situations. The speakers were each given an emotion that they embodied in their speech with a listener responding accordingly. We note also that the dataset includes 32 evenly distributed emotion labels which include basic emotions like surprised, excited, annoyed, lonely, joyful and afraid. The model accesses these emotional labels to learn associations between the context and the emotional response. This approach facilitates a more nuanced understanding of emotional dynamics in conversations, allowing the model to generate responses that are not only contextually relevant but also emotionally resonant. Furthermore, the balanced distribution of emotion labels within the dataset ensures a comprehensive exposure to a wide range of emotional states, thereby enriching the model's ability to engage in empathetic dialogue across diverse situations.

5.2 Evaluation methods

In our study, we rigorously evaluate the performance of two models: one using a baseline multi-headed attention mechanism, and the other incorporating convolutional attention. For each model, we generate a sample of 30 outputs and compute their BLEU (Bilingual Evaluation Understudy) scores. The BLEU metric assesses the quality of machine-generated text by comparing it to reference translations, focusing on the precision of n-grams (word sequences of various lengths) in the generated text against those in the reference.

Furthermore, we utilize the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score to measure the overlap of n-grams between the model-generated texts and the reference texts. ROUGE is particularly useful for evaluating the extent to which the models are capable of reproducing human-like responses, as it considers both the precision and recall of n-grams.

By employing both BLEU and ROUGE metrics, we aim to provide a comprehensive assessment of each model's ability to replicate human-like responses.

The aforementioned test allows us to measure the "humaness" of the responses; however, it does not quantity its performance with regard to emotional representation. To assess the effectiveness of using convolutional attention versus multi-headed attention in this aspect, we conduct a comparative analysis using datasets with emotional data labelling and without. We specifically compare the performance gaps as indicated by the BLEU and ROUGE recall scores between the two models across both dataset types. When analyzing the results on the dataset with emotional content, we note the ability of each model to capture the nuanced emotional aspects in response and is essential because it validates our hypothesis in testing the ability of convolutional attention to be better suited for emotional data. Conversely, when analyzing the results on the dataset without emotional labels, we assess performance in a neutral context and focus on language understanding.

Comparison of the gaps in scores between the models across the datasets is essential as a smaller gap in emotional versus the non-emotional is indicative that the convolutional attention is more adept at handling emotional data as we hypothesize. As such, it should give us a measure of how the mechanisms perform in both the emotional and language quality factors.

For further qualitative evaluation, we conduct a survey sampling from 5 participants who were asked to rank 50 samples outputs on a numerical scale from one to five on two metrics - emotional intensity and response relevancy - where five indicates high emotion or high relevancy for the output. Additionally, we also aim to directly compare the generated outputs for both models with each other. We randomly sample 50 dialogues and the five participants were given randomly ordered responses and prompted to choose the response they felt was better. Both qualitative experiments provide a more detailed method of assessing the ability to replicate human-like expression.

5.3 Experimental details

In our experiments, we used a consistent setup with 300-dimensional word embeddings and a hidden size of 300. We trained our models with a batch size of 16 and tested with a batch size of 1 to simulate real-world, individual query responses. We train our model using Adam optimizer and used a learning rate of 1e-4 via standard practice for training. For the calculations of the BLEU and the Rougue scores, we generate 20 samples from each model and calculate the scores using the ground truth response.

5.4 Results

The following scores were calculated from the generations from each of the two models and on the emotional data and the non-emotional data, as seen in Table 1 and Table 2, respectively.

(Emotional Data)	BLEU Score	ROUGE Recall Score
Multi-Headed Attention	0.31	0.1245
Convolutional Attention	0.23	0.1111

Table 1: Evaluation of Multi-Headed Attention and Convolutional Attention mechanisms on emotional data

(Non-emotional Data)	BLEU Score	ROUGE Recall Score
Multi-Headed Attention	0.28	0.1135
Convolutional Attention	0.17	0.0905

Table 2: Evaluation of Multi-Headed Attention and Convolutional Attention mechanisms on non-emotional data

In the tables above, the data illustrates that the multi-headed attention mechanism and the convolutional attention mechanism both perform better when using the emotional data as reflected in

BLEU and ROUGE scores. We also note that the multi-headed model consistently outperformed the convolutional attention model in both the emotional and non-emotional data cases. However, both models show a decline in both evaluation metrics when using non-emotional data over emotional data. In the multi-headed attention model, the BLEU score decreases by 0.03 and the ROUGE score decreases by 0.0110. In the convolutional attention model, the BLEU score decreases by 0.06 and the ROUGE score decreases by 0.0195.

The survey results provide a qualitative comparison of the outputs between each of the models on the emotional data. Participants were asked to rate the intensity of the emotion as well as the relevancy of the responses on a scale from one to five. The results of the trial are listed below in Table 3 which gives the score evaluations of the 50 outputs from each model and Table 4 which gives a direct comparison of the percentage preference for each model across 50 different outputs.

	Emotional Intensity	Response Relevancy
Multi-Headed Attention	3.44	3.70
Convolutional Attention	3.38	3.55

Table 3: Survey FINISH this

	Conv-Pref	MH-Pref	Tie
Convolutional vs. Multi-Headed Attention	25.65%	30.57%	43.78%

Table 4: Direct comparison of both models across 50 outputs with Conv-Pref representing a preference toward output of a convolutional attention based model and MH-Pref representing a preference toward output of multi-headed attention based model.

6 Analysis

6.1 Impact on Emotional vs. Non-emotional Data Handling

The reported BLEU and ROUGE scores indicate two key results. The first being that the multi-headed attention exhibits better performance than the convolutional attention over both types of datasets (emotional and non-emotional) as indicated by the higher scores in both metrics. The observed discrepancy in performance between models using multi-headed and convolutional attention mechanisms, especially pronounced in the context of emotional data, underscores the intrinsic differences in how these models process and prioritize textual information. The multi-headed attention mechanism, with its capacity to concurrently attend to information from different representation subspaces at different positions, appears to offer a more holistic comprehension of the textual context. This broader perspective likely facilitates a superior integration of emotional nuances within the generated text, as evidenced by the higher BLEU and ROUGE scores in datasets laden with emotional content. In contrast, the convolutional attention mechanism, designed to capture local dependencies more effectively, might be expected to excel in discerning the finer emotional nuances embedded within close textual proximities. However, the results indicate that its localized focus does not necessarily translate into a better overall grasp of the emotional context, possibly due to its relative inadequacy in synthesizing broader contextual cues that are pivotal in understanding and generating nuanced emotional responses.

6.2 Decline in Non-emotional Data Performance

The decline in performance metrics for both models when engaging with non-emotional data reveals a critical aspect of model sensitivity to emotional content. This sensitivity, while beneficial for processing emotional texts, suggests a potential over-specialization that impairs the models' ability to generalize across a broader spectrum of textual data. The more pronounced decline in performance observed with the convolutional attention model on non-emotional data further emphasizes this model's heightened sensitivity to local emotional cues. Such a trait, while potentially advantageous

in highly emotional contexts, may detract from the model's versatility and efficacy in neutral settings, where emotional cues are sparse or non-existent. This finding underscores the need for a balanced attention mechanism that adeptly navigates both emotionally charged and neutral texts without compromising on performance.

6.3 Participant Surveys: Emotional Intensity and Response Relevancy

Participant surveys were conducted in order to assess the emotional intensity and relevancy of responses generated by each model offer additional insights into the nuanced capabilities of these attention mechanisms. Although the multi-headed attention model appears to be marginally preferred for both emotional intensity and response relevancy, we note that this slight preference might reflect the model's ability to leverage the broader global contextual awareness to produce responses that are more resonant and contextually appropriate. However, again we see that the gap between the ratings for the emotional intensity is smaller than that of the response relevancy which indicates that the convolutional attention mechanism might have performed better in that area as a result of the local dependency patterns in the data. Additionally, due to its localized focus, the convolutional attention model may also provide better precision with emotional context into the generated outputs which may also explain the consistent gap changes we observe.

7 Conclusion

In this research we focused on increasing the emotional realism on language model outputs. Specifically, we investigated the use of a convolutional attention mechanism in a unique transformer based language model architecture. We choose to use a convolutional attention mechanism because of its ability to better capture local dependencies (specifically for parsing emotion) in training.

In our evaluation of our attention mechanism versus multi-headed attention we found that multi-headed attention consistently outperforms convolutional attention in both emotional and non-emotional contexts, as indicated by higher BLEU and ROUGE scores. This superiority is particularly pronounced in processing emotional content, likely due to the multi-headed approach's ability to simultaneously process various aspects of textual information, resulting in a more comprehensive understanding of context, including emotional nuances. However, a notable decline in performance for both models in non-emotional data suggests a sensitivity to emotional content, which could limit their applicability in neutral contexts. The convolutional model, while proficient in recognizing local emotional cues, shows a marked decrease in effectiveness in non-emotional settings, highlighting a potential over-specialization. Participant surveys further reinforce the multi-headed model's edge in creating contextually resonant responses, though the convolutional model's focus on local dependencies suggests its utility in specific scenarios where fine-grained emotional nuances are crucial.

While our hypothesis regarding convolutional attention proved to be less effective, our future work will explore additional attention mechanisms. By continuing to refine and test different attention-based approaches, we aim to develop more versatile and effective language models that can adapt to various emotional and contextual nuances present in human language.

References

Jacky Casas, Samuel Torche, Karl Daher, Elena Mugellini, and Omar Abou Khaled. 2021. Emotional paraphrasing using pre-trained language models. *ArODES (HES-SO (https://www.hes-so.ch/))*.

Jia Guo. 2022. Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, 31(1):113–126.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Qiang Yang, and Xing Xie. *Large Language Models Understand and Can Be Enhanced by Emotional Stimuli*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. I know the feeling: Learning to converse with empathy.

Sapan Shah, Sreedhar Reddy, and Pushpak Bhattacharyya. 2023. Retrofitting light-weight language models for emotions using supervised contrastive learning.