Emo-Gen BART for SCI-CHAT 2024 Shared Task: A Multitask Emotion-Informed Dialogue Generation Framework

Alok Debnath
ADAPT Centre
Trinity College Dublin
debnatha@tcd.ie

Yvette Graham
ADAPT Centre
Trinity College Dublin
ygraham@tcd.ie

Owen Conlan
ADAPT Centre
Trinity College Dublin
owen.conlan@tcd.ie

Abstract

This paper is the model description for the *Emo-Gen BART* dialogue generation architecture, as submitted to the SCI-CHAT 2024 Shared Task. The **Emo**tion-Informed Dialogue **Gen**eration model is a multi-task BART-based model which performs dimensional and categorical emotion detection and uses that information to augment the input to the generation models. Our implementation is trained and validated against the IEMOCAP dataset, and compared against contemporary architectures in both dialogue emotion classification and dialogue generation. We show that certain loss function ablations are competitive against the state-of-the-art single-task models.

1 Introduction

The realm of human conversation is intricately woven with emotions, a fundamental aspect that significantly influences the dynamics of communication (Li et al., 2021). In contemporary research within Natural Language Processing (NLP) and Human-Computer Interaction (HCI), the development of emotion-aware conversational agents has emerged as a focal point. Various methodologies have been employed to handle emotions in conversation, with categorical labels and dimensional ratings being prominent avenues. These labels often find their roots in established emotion theories, such as Ekman's (Ekman and Oster, 1979) or Plutchik's (Plutchik, 1980), as evidenced in datasets like IEMOCAP (Busso et al., 2008) and DailyDialog (Li et al., 2017). Additionally, alternative corpora and models adopt unique lists of emotion words, exemplified by the EMPATHETIC-DIALOGUES dataset (Rashkin et al., 2019).

The "dimensional" approach to handling emotion involves the utilization of characteristics inherent in emotional speech (Buechel and Hahn, 2017). A noteworthy model in this context is the Valence-Arousal-Dominance (VAD) Model, which assesses the positive or negative sentiment, the degree of excitation, and the level of control exerted by the stimulus (Buechel and Hahn, 2016). This model has become a cornerstone in understanding and quantifying the nuanced dimensions of emotions expressed in conversational interactions. As we delve into the intricacies of emotion-aware conversational agents, the utilization of both categorical and dimensional frameworks provides a comprehensive understanding of the emotional landscape within human-machine dialogues.

In the domain of emotion-aware or empathetic conversational agents, diverse methodologies have been employed to augment systems' understanding and responsiveness to emotional cues. Some methods incorporate input augmentation techniques, thereby exposing the conversational agent to various emotional expressions to enhance learning robustness (Goel et al., 2021; Carolus et al., 2021). Simultaneously, alternative approaches integrate common-sense or pragmatic information, drawing upon broader contextual knowledge to enrich the agent's comprehension of emotions within a given conversation (Ghosal et al., 2020; Scotti et al., 2021).

Our system, **Emo-Gen BART** is a modification on BART architecture (Lewis et al., 2019). Our approach uses BART's emotion decoder attention representation to perform emotion classification as well as dimensional emotion detection. We then augment that representation to reinforce signals associated with emotion information. Our strategy implements emotion classification and regression and combines their loss with the emotion-informed generation task. When accounting for contextual information through the conversation, we find that this method makes it competitive with state-of-the-art conversational agents.

Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024), pages 70–74

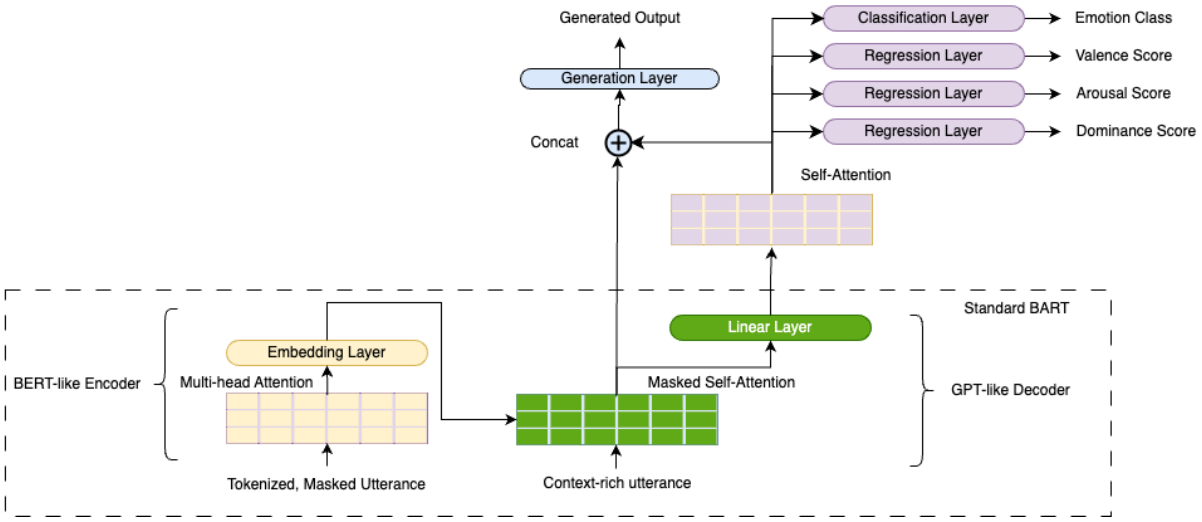

Figure 1: The Emo-Gen BART is a variation over the standard BART architecture with a bidirectional BERT-like encoder and an autoregressive decoder. Note that during fine-tuning for conditional generation, the input sentence is provided to both the encoder and the decoder. Using the decoder output, we perform multiple emotion detection tasks. The final generation layer uses the decoder attention as well as the multitask attention.

2 Model Architecture and Implementation

Emo-Gen BART, a customized version of the BART language model, is specifically tailored for conditional dialogue generation. The architecture of BART involves a bidirectional encoder processing tokenized and masked input sentences. During fine-tuning, this encoder utilizes the denoised input along with the encoder representation to generate subsequent sentences. For conditional generation tasks, a randomly initialized encoder precedes the bidirectional encoder during training.

During fine-tuning, Emo-Gen BART modifies the BART architecture by extracting the last hidden layer, employed to predict emotion class and Valence-Arousal-Dominance (VAD) attention scores, illustrated in Figure 1. Emotion-aware information is incorporated by concatenating the multitask and decoder attention outputs before the generation phase.

2.1 Loss Functions and Training Objectives

Emo-Gen BART, a modification of the BART encoder-decoder model, incorporates three key refinements during fine-tuning. Firstly, a multitask classification and regression model employs the decoder output for prediction. Secondly, attention outputs from the multitask model are concatenated with the decoder attention outputs during the generation phase. Thirdly, in fine-tuning for conditional generation, the decoder receives input as the sentence with the preceding context truncated at the input length.

Consider an utterance $\mathbf{u} = u_1, \ldots, u_M$ the model parameters θ, which update based on each task.

Classification The objective of the classification layer is to minimize cross-entropy loss between predicted and actual emotion class values. For a batch of N samples, we compute classification loss as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j} \log p(c_{i,j}|\mathbf{u};\theta)$$

where C is the number of classes, $y_{i,j}$ is the binary indicator of where j is the correct class and $p(c_{i,j})$ is the predicted probability distribution of the model for the i^{th} utterance $\mathbf{u}$.

Regression The three regression tasks, i.e. valence, arousal, and dominance detection, are trained with the objective of minimizing the mean-squared error loss between the predicted and actual values, which is computed as:

$$\text{MSE}(\hat{y}, y_{\text{true}}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_{\text{true}_i})^2$$

for any predicted value $\hat{y}_i$ and any true value y_{true_i} for any i of N samples. The predicted and actual regression value for each utterance is summed up

for all utterances across a batch. So, the dimensional emotion loss can be computed as:

$$\mathcal{L}_{\text{reg}} = \sum_{d \in D} \lambda_d \cdot \text{MSE}(\hat{d}, d_{\text{true}})$$

where D are the emotion dimensions, λ_d is the weight for each regression task. For our purposes, $\forall d \in D; \lambda_d = 1$.

Generation The generation layer was implemented analogously to the BART decoder. The outputs of the final layer from the decoder and the multitask self-attention layers are concatenated and passed through a linear layer for generation. The input to the encoder is the current utterance tokenized, while the decoder input includes the context of the conversation.

Note that the input to the encoder and the decoder differ. For every utterance $\mathbf{u}$, there is a context $\mathbf{c} = \{c_1, \ldots, c_N\}$, which is comprised of previous utterances and responses. Therefore, the input to the generation layer may be computed as:

$$\mathbf{x} = \text{Attn}_{\text{decoder}}(\mathbf{c} \cdot \mathbf{u}) \oplus \text{Attn}_{\text{multitask}}(\mathbf{u})$$

For every input $\mathbf{x}$, the model generates a response $\mathbf{y} = \{y_1, \ldots, y_n\}$. The training objective here is also to minimize cross-entropy loss between the generated sequence and the actual dialogue response, which may be computed as:

$$\mathcal{L}_{\text{gen}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \log P(y_t | \mathbf{x}, y_1, \ldots, y_{t-1}; \theta)$$

wherein N is the number of samples per batch, T is the length of the generated sequence, $y_{i,t}$ represents the predicted probability distribution over the vocabulary for the t^{th} token in the i^{th} sequence.

Combined Training Objective The training objective of the model is to minimize the total loss, computed as a weighted sum of the regression, classification, and generation losses.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{cls}} + \beta \cdot \mathcal{L}_{\text{reg}} + \gamma \cdot \mathcal{L}_{\text{gen}}$$

For our purposes, $\gamma = 1$ and $0 < \alpha = \beta \leq 1$. We find that varying the relative importance of the non-generation tasks impacts generation, but causes sensitivity to initial hyperparameters.

3 Experiments

In this section, we describe the dataset, experimental setup, and hyperparameter information for reproducing these experiments.

3.1 Dataset

We fine-tune Emo-Gen BART on the IEMOCAP corpus (Busso et al., 2008). This is a benchmark corpus of recorded conversations which have been transcribed into dialogue sessions, annotated with both categorical and dimensional emotion. The IEMOCAP dataset includes video data of impromptu performances or scripted scenes of about 10 actors. There are in total 7433 utterances and 151 dialogues in the IEMOCAP dataset. At the same time, it contains audio and text transcription to meet the needs of multimodal data. In this data set, multiple commentators set the emotional labels of the utterances into nine categories: including happy, sad, neutral, angry, excited frustrated, surprised, and afraid

3.2 Hyperparameter Tuning

We fine-tune the model over 64 epochs with a learning rate of 10^{-5} and a batch size of 16. The data is preprocessed to include context of every utterance alongside the utterance to the generation layers, the input length set at 256. The multitask self-attention layers follow the dimensions of the decoder layers, i.e. 768 hidden dimensions with 4 attention heads per layer for 6 layers. For generation, we constrain the model to generate sentences with a minimum of 2 tokens, with a temperature of 1.6, a high top-k vocabulary spread of 400 tokens and the top-p probability sum of 0.95. Training and generation are performed on an NVIDIA-RTX2080ti.

3.3 Baseline Models for Comparison

We compare our results against the following baseline models:

BC-LSTM, introduced by Poria et al. (2017) employs a Bidirectional LSTM structure to capture contextual semantic information. However, it lacks the capability to recognize speaker relationships within the encoded content.

DialogueGCN, presented by Ghosal et al. (2019), organizes a conversation into a graph structure, converting the speech emotion classification task into a graph-based node classification problem. The method employs a graph convolutional neural network to effectively classify the outcomes.

Ide and Kawahara (2021), introduced a BART-based multitask framework as well. The difference between our model and their implementation is the

Model	Avg F1 Score
BC-LSTM	59.19
DialogueGCN	64.18
Ide and Kawahara (2021)	62.42
Emo-Gen BART	**69.49**

Table 1: The comparative performance results for emotion classification of our model against the baselines.

Model	BLEU	*dist-1*	*dist-2*
Ide and Kawahara (2021)	32.55	6.00	**30.77**
Emo-Gen BART	**36.46**	**6.46**	30.65

Table 2: The comparative performance results for emotion-aware generation.

use of only a categorical label for their multitask generation, and that it does not adopt the context input.

4 Results and Findings

4.1 Emotion Classification Results

By leveraging the BART pre-trained language model, our model adeptly encodes sentences to enhance the representation of utterances. Simultaneously, our multitask attention framework integrates both the inherent emotional tendencies of the utterance and contextual information. This approach proves more effective in discerning the speaker's emotion, as affirmed by experimental results. Our assumptions regarding the emotional factors within ERC find validation through these findings.

4.2 Dialogue Generation Results

Initially, we assess the relevance of output responses to the correct response using BLEU (Papineni et al., 2002). Subsequently, we examine lexical diversity by evaluating distinctiveness, as proposed by Li et al. (2016). This distinctiveness measure is calculated through *distinct-1* and *distinct-2*, which focus on unigrams and bigrams, respectively. We find that the *distinct-2* value for our method is lower than the state-of-the-art multitask model, which warrants further investigation.

The model has been submitted to the SCI-CHAT shared task for human evaluation and benchmarking.

5 Conclusion and Future Work

In this paper, we introduce Emo-Gen BART, an architecture that employs a modified BART language model to enhance the capabilities of emotion-aware conversational agents. Our approach integrates a multitask attention framework, acting as an emotion capsule, to improve the model's proficiency in identifying emotional cues during dialogue generation.

We find that this approach of accounting for several tasks including emotion classification and regression, can inform the model and improve upon baseline results. We use only a single model variation where all the loss functions are weighted equally, however model ablations which form a hyperparameter relationship between the various tasks. Finally, with multitask setups which change the nature of the architecture itself, it would be interesting to leverage LLM predictions using dataset specific signals.

Acknowledgments

The work presented in this paper is supported the and is supported by the Science Foundation Ireland Research Centre, ADAPT at Trinity College Dublin under Grant Agreement No 13/RC/2106_P2. This work has received research ethics approval by Trinity College Dublin Research Ethics Committee (Application no. 20210603).

References

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016*, pages 1114–1122. IOS Press.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *EACL 2017*, page 578.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Astrid Carolus, Carolin Wienrich, Anna Törke, Tobias Friedel, Christian Schwietering, and Mareike Sperzel. 2021. 'alexa, i feel for you!'observers' empathetic reactions towards a conversational agent. *Frontiers in Computer Science*, 3:682982.

Paul Ekman and Harriet Oster. 1979. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Raman Goel, Sachin Vashisht, Armaan Dhanda, and Seba Susan. 2021. An empathetic conversational agent with attentional mechanism. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4. IEEE.

Tatsuya Ide and Daisuke Kawahara. 2021. Multi-task learning of generation and classification for emotion-aware dialogue response generation. *NAACL-HLT 2021*, page 119.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 1204–1214.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Vincenzo Scotti, Roberto Tedesco, and Licia Sbattella. 2021. A modular data-driven architecture for empathetic conversational agents. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 365–368. IEEE.