Decoding-enhanced BERT with Disentangled Attention in Speech Emotion Recognition

Răzvan-Gabriel Petec
Department of Computer Science
Babeș-Bolyai University
Cluj-Napoca, Romania
razvan.petec@stud.ubbcluj.ro

***Abstract*—Speech emotion recognition (SER) is a challenging task that involves extracting information about the emotions of some speaker using speech signals. Recently, deep learning models, such as BERT (Bidirectional Encoder Representations from Transformers) [1], have shown promising results in SER [2]. However, these models typically focus on encoding the speech signal and do not explicitly consider the decoding process of generating emotional labels. This paper will showcase the results of using DeBERTa (Decoding-enhanced BERT with Disentangled Attention) [3] on the text data of IEMOCAP dataset [4], enabling it to explicitly capture the relationship between the encoded speech features and the corresponding emotional labels.**

***ACM classification*—Natural language processing (I.2.7)**

***AMS classification*—Natural language processing (68T50), Applications of statistics to psychology (62P15)**

***Index Terms*—affective computing, speech emotion recognition, decoding-enhanced BERT with disentangled attention**

I. Introduction

Natural Language Processing (NLP) is a branch of Computer Science and Artifcial Intelligence that focuses on the computational treatment of human language with the core intent of making machines understand and generate human languages. NLP's favored applications, such as translation systems, search engines, natural language assistants, sentiment, and opinion analysis, are resolving societal issues at an unprecedented rate [5].

The sequential nature of textual data, wherein the order and relationships among words are determinative of complete sentence meaning, underscores a fundamental challenge. Traditional unsupervised machine learning models, which disregard word order and relationships while being confined by fixed input sizes, prompt a paradigm shift toward computationally deep approaches for textual data analysis.

The Recurrent Neural Network (RNN), as a sequential model adept at handling sequential data [6], encounters limitations attributable to protracted training times and difficulties in accommodating long-range sequence dependencies. The Long Short-Term Memory (LSTM), a variant of RNN, addresses some of these challenges by offering a resolution to long-range sequence dependencies. However, the enhancement comes at the expense of neglecting parallel computations and exhibiting slower operational speeds compared to the conventional RNN [7].

Emotion Detection (ED) is a branch of sentiment analysis (SA) that aims to identify subtle emotional nuances from various sources, including speech, images, and text. Despite the abundance of textual data, accurately detecting emotions from text remains challenging [8]. This difficulty stems, in part, from the absence of contextual cues provided by voice modulation, facial expressions, and other non-verbal signals. Additionally, the lack of an effective context extraction method for text poses another obstacle. Moreover, the need to differentiate between emotion-conveying words and their actual emotional meanings presents a significant hurdle, as many texts convey multiple emotional expressions. Recent advancements in state-of-the-art (SOTA) results have been achieved using pre-trained transformer-based models in this field.

Thus, the objective of this paper is to review DeBERTa, one of the most precise transformer-based models at the moment, specifcally in text-based emotion detection. Initially, some of the SOTA models will be described, then, the paper focuses on how the hypothesis was approached, then some of the results will be displayed, following the conclusions and the description of how the approach can be improved with some future work.

II. Related Work

A. CNN and LTSM based Emotion Recognition Models

Yoon et. al [9] proposed 3 text models using 2 different approaches for the IEMOCAP database: Text_Model1 which used 1D convolutions of kernel size 3 each, with 256, 128, 64 and 32 filters using Relu as Activation and Dropout of 0.2 probability, followed by 256 dimension fully connected layer and Relu, feeding to 4 output neurons with Softmax; Text_Model2 which used two stacked LSTM layers with 512 and 256 units followed by a Dense layer with 512 units and Relu Activation; and Text_Model3 which was simillar to Text_Model2, but they used Randomized initialization with 128 dimensions, the first two models being initialized with Glove Embeddings based word-vectors. Text_Model3 obtained similar performance as Text_Model2. The LSTM based models use Adadelta and Convolution based models use Adam as optimizers. The best of the three models was Text_Model3, obtaining an accuracy of 64.78%.

B. BERT based Emotion Recognition Models

Emile et. al [10] proposed a hierarchical transformer-based encoder tailored for spoken dialog. They extend two well-known pre-training objectives to adapt them to a hierarchical setting. They used BERT [1] through the pytorch implementation provided by the Hugging Face transformers library [11]. The pre-trained model was fed with a concatenation of the utterances. Formally given an input context Ck = (u1, ... uT), the concatenation [u1, . . . , uT] was fed to BERT. With this approach, they obtained an accuracy of 66.05%.

III. Approach

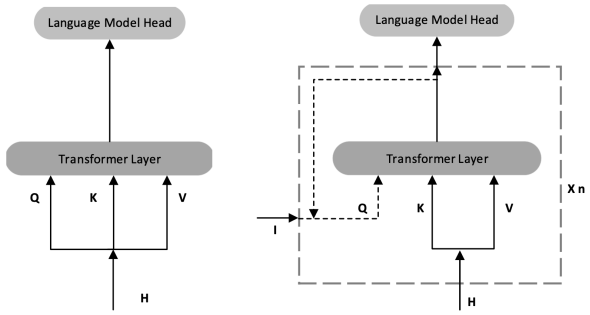
A. IEMOCAP Dataset

Interactive EMOtional dyadic motion CAPture database (IEMOCAP) [4] is a multimodal and multi-speaker database that contains approximately 12 hours of audiovisual data, including video, speech, motion capture of the face, and text transcriptions. The database is annotated by multiple annotators into categorical labels: anger, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral. From all of this data, the text and the labels were extracted to fit through the proposed model.

B. Pre-trained Encoder Model

Similarly to Emile et. al [10], the DeBERTaV3-Small model published by Microsoft [12] from the transforms library provided by Hugging Face [11] was used to try to achieve some good resulst on the text data of IEMOCAP. This version of the model has 44M backbone parameters, having a vocabulary containing 128K tokens.

The decision to use this model instead of other BERT models, such as BERT or RoBERTa, was taken because DeBERTa improves the other two models by using disentangled attention and enhanced mask decoder (Fig. 1). With those two improvements, DeBERTa out perform RoBERTa on a majority of Natural Language Understanding (NLU) tasks with 80GB training data.



(a) BERT decoding layer (b) Enhanced Mask Decoder

Fig. 1. Comparison between the BERT's and DeBERTa's decoding layer.

As a further improvement, DeBERTa V3 used ELECTRA-Style pre-training with Gradient Disentangled Embedding Sharing. Compared to DeBERTa, Microsoft's V3 version significantly improves the model performance on downstream tasks.

C. Training

a) Loss function: Having the labeled IEMOCAP dataset and taking into the consideration that the output of our model was equal to the number of possible classes in the dataset, the cross entropy loss function was used:

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

b) Optimizer: To minimize our loss and for a faster convergence, the weighted Adam optimizer was used, having the following equations:

- $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
- $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
- $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
- $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
- $\theta_t = \theta_{t-1} - \frac{\text{lr} \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \text{lr} \cdot \text{weight\_decay} \cdot \theta_{t-1}$

where:

- m_t is the first moment estimate.
- v_t is the second moment estimate.
- β_1 and β_2 are the exponential decay rates for the moment estimates.
- g_t is the gradient at time step t.
- $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected moment estimates.
- lr is the learning rate.
- ϵ is a small constant to prevent division by zero.
- weight_decay is the weight decay term.

c) Mini-batches: Usually, most of the sequential models are trained with a stochastic approach, because of the fact that the sequential data could have different sizes, but, the stochastic approach is known to take longer to train and, some times, to be worse on generalizing data [13]. To solve this issue, we padded our sequences as a preprocessing step to the longest sequence in the training set with 0s, so the model can learn to "ignore" those numbers.

IV. Results

The model was trained for 9 epochs, which took 450 minutes on my machine. Different metrics were took throughout the training for each batch. Before training, the IEMOCAP dataset was split into a train set (80% of the samples) and a validation set (the rest of the samples), because no official split is provided in the original paper. Before the training environment, all the seeds of the used libraries (python's random, numpy and pytorch) were set to 0, to have consistent data throughout different experiments.

Epoch Start	Epoch End	Accuracy
1	3	51.493%
4	6	51.293%
7	9	51.957%

The hyperparameters choosen when fine-tuning the model were:

- a batch size of 32 samples
- the parameters of the optimizer:
  - $learning\_rate = 2 * 10^{-5}$
  - $\beta_2 = 0.999$
  - $\beta_1 = 0.9$
  - $weight\_decay = 10^{-2}$
  - $epsilon = 10^{-8}$

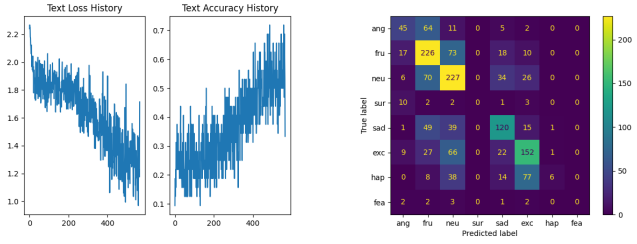


Fig. 2. Loss and accuracy evolution for the training set and confusion matrix for the evaluation set, epochs 1-3.

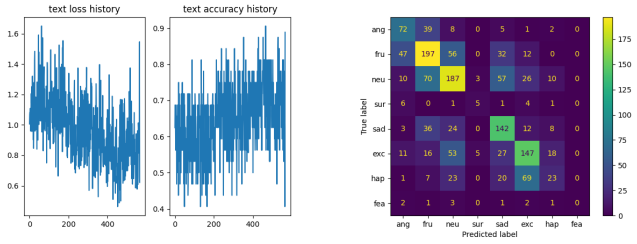


Fig. 3. Loss and accuracy evolution for the training set and confusion matrix for the evaluation set, epochs 4-6.

We can see that the loss function converges very well for the first epochs (Fig. 2, Fig. 3), and for the last ones it seems like it converged and there is no need for further training (Fig. 4), but the accuracies on the validation dataset throughout the epochs doesn't seem to change, meaning that it could be harder for the model to generalize. Seeing the confusion matrix, it's visible that many of the

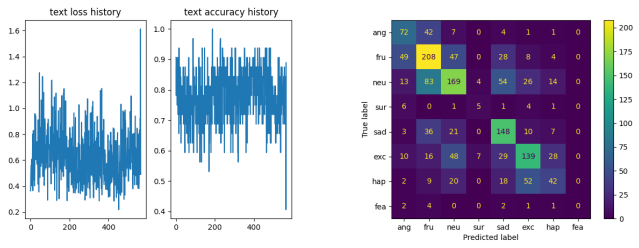


Fig. 4. Loss and accuracy evolution for the training set and confusion matrix for the evaluation set, epochs 7-9.

emotions the model is confusing are similar, so it could be harder for the model to distinguish between similar emotions.

Model	***Accuracy***
bc-LSTM + Attention (2017)	58.540%
Attention-BLSTM (2018)	62.900%
Dialogue-CRN + RoBERTa (2021)	67.530%
EmoCAPS (2022)	71.770%
DeBERTa (mine)	**51.957%**

V. Conclusions and Future Work

This paper demonstrates the effectiveness of applying DeBERTa, a pre-trained language model architecture that integrates disentangled attention and enhanced mask decoder into the BERT framework, for the task of speech emotion recognition. The findings in this paper suggest that DeBERTa's disentangled attention mechanism and enhanced mask decoder are beneficial for understanding and processing the text from different speeches. The fact that DeBERTa has not been previously applied to speech emotion recognition highlights the potential for further research in this area.

In future work, some more hyperparameter tuning could be done to try to achieve better results, as not much was done in this paper. Also, the experiments were done using the small version of DeBERTa due to hardware limitations, but bigger versions of the same model could be used to achieve better results.

References

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." North American Chapter of the Association for Computational Linguistics. (2019)

[2] Acheampong, F.A., Nunoo-Mensah, H. & Chen, W. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif Intell Rev 54, 5789–5829. (2021)

[3] He, Pengcheng, Xiaodong Liu, Jianfeng Gao and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. (2020)

[4] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359. (2008)

[5] Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. Knowl Inf Sys 60(2):617–663. (2019)

[6] Du K-L, Swamy MN. Neural networks and statistical learning. Springer Science & Business Media, Berlin. (2013)

[7] Sundermeyer M, Schlüter R, Ney H. Lstm neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association, p 4. (2012)

[8] Acheampong FA, Wenyu C, Nunoo-Mensah H. Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports e12189. (2020)

[9] S. Yoon, S. Byun and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text" IEEE Spoken Language Technology Workshop (SLT), pp. 112-118. (2018)

[10] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. Hierarchical Pre-training for Sequence Labelling in Spoken Dialog. In Findings of the Association for Computational Linguistics: EMNLP, pages 2636–2648. (2020)

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. (2019)

[12] He, Pengcheng & Gao, Jianfeng & Chen, Weizhu. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. (2021)

[13] Geiping, Jonas & Goldblum, Micah & Pope, Phillip & Moeller, Michael & Goldstein, Tom. Stochastic Training is Not Necessary for Generalization. (2021)