



Decoding-enhanced BERT with Disentangled Attention in Speech Emotion Recognition

Research Project

Coordinator: Prof. Phd. Istvan-Gergely Czibula

Student: Răzvan-Gabriel Petec

Why?

Such an algorithm can help in different scenarios, such as a public speech where the speaker might not be fully visible, the impact of the attendee's experience is heavily influenced by what the speaker says. An AI algorithm designed for this purpose can ensure that speakers effectively communicate the intended sentiments in their message and could connect better with the audience.

Goals

- to create such an algorithm that could help speakers tweak their speeches in such a way that they could transmit the desired sentiments to the audience.
- robust performance across different speakers, minimizing dependency on demographic factors such as gender, age, or cultural background.
- to try to improve the current state of the art.

The background of the image is a dense, abstract pattern of dark gray, three-dimensional geometric shapes. These shapes, which include cubes and triangles, are arranged in a way that creates a strong sense of depth and perspective. The lighting is directional, coming from the upper left, which causes the faces of the shapes to be shaded differently, with some appearing lighter and others in deep shadow. This creates a complex, textured effect that resembles a crystalline or architectural surface.

Dataset

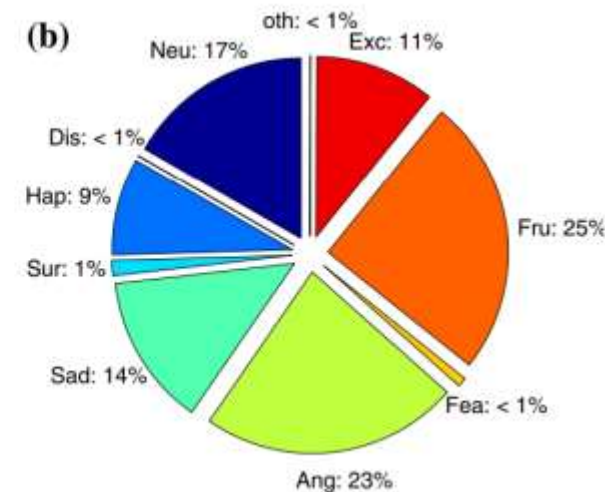
IEMOCAP: Interactive Emotional Dyadic Motion CAPture Database

**Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower,
Samuel Kim, Jeannette N. Chang, Sungbok Lee, Shrikanth S. Narayanan**

IEMOCAP 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.

The IEMOCAP data was collected from 10 actors, five males and five females, who were all native English speakers. Each actor performed in approximately 120 dyadic sessions, with each session lasting approximately 5 minutes. The sessions were recorded in a controlled environment, and the actors were asked to perform improvisations or scripted scenarios that were designed to elicit emotional expressions.



IEMOCAP: Interactive Emotional Dyadic Motion CAPture Database

The IEMOCAP database is a valuable resource for researchers in the field of emotion recognition and human-computer interaction. It has been used and is still used to this date in numerous studies to develop and evaluate emotion recognition algorithms. The database is also used to study how humans express emotions through their facial expressions, speech, and body language.

Usage 

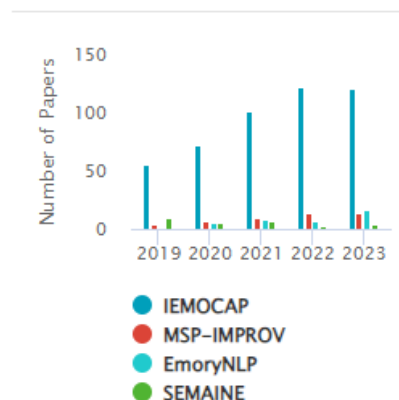


Table 3 Example of the annotations for a portion of a spontaneous session (third scenario in Table 1)

Seg. (s)	Turn	Transcription	Labels	[v,a,d]
[05.0–07.8]	F00:	Oh my God. Guess what, guess what, guess what, guess what, guess what, guess what?	[exc][exc][exc]	[5,5,4][5,5,4]
[07.8–08.7]	M00:	What?	[hap][sur][exc]	[4,4,3][4,2,1]
[08.9–10.8]	F01:	Well, guess. Guess, guess, guess, guess.	[exc][exc][exc]	[5,5,4][5,5,4]
[11.1–14.0]	M01:	Um, you–	[hap][neu][neu]	[3,3,2][3,3,3]
[14.2–16.0]	F02:	Don’t look at my left hand.	[exc][hap][hap;exc]	[4,3,2][5,4,3]
[17.0–19.5]	M02:	No. Let me see.	[hap][sur][sur]	[4,4,4][4,4,4]
[20.7–22.9]	M03:	Oh, no way.	[hap][sur][exc]	[4,4,4][5,4,3]
[23.0–28.0]	F03:	He proposed. He proposed. Well, and I said yes, of course. [LAUGHTER]	[exc][hap][hap;exc]	[5,4,3][5,5,3]
[26.2–30.8]	M04:	That is great. You look radiant. I should’ve guess.	[hap][hap][exc]	[4,4,3][5,3,3]
[30.9–32.0]	F04:	I’m so excited.	[exc][exc][exc]	[5,4,3][5,5,3]
[32.0–34.5]	M05:	Well, Tell me about him. What happened	[hap][exc][exc]	[4,4,3][4,4,4]

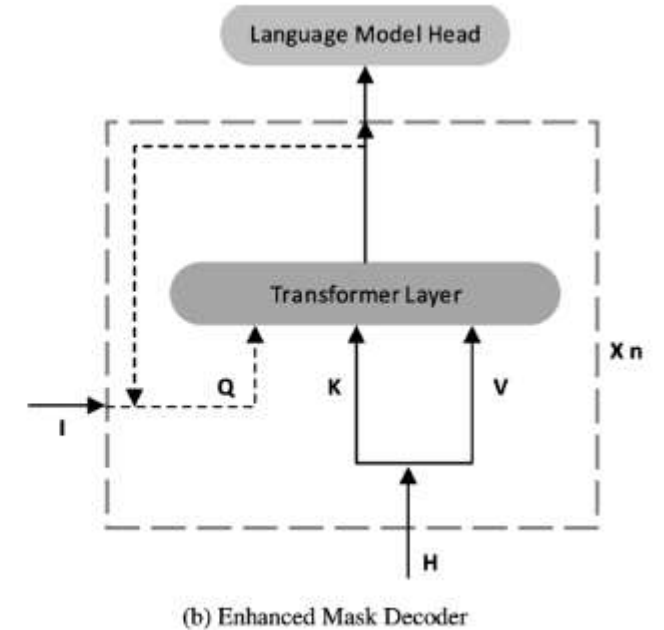
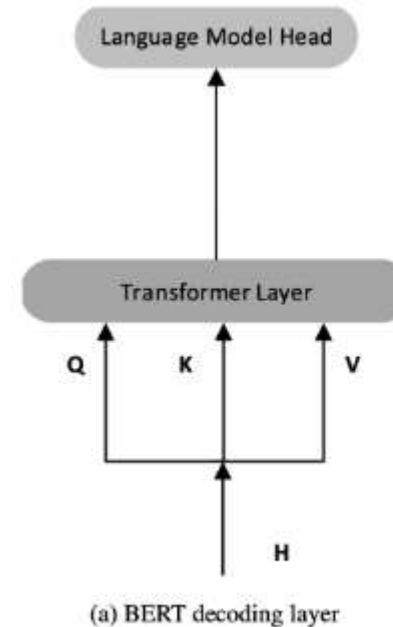
The example includes the turn segmentation (in seconds), the transcription, the categorical emotional assessments (three subjects) and the attribute emotional assessment (valence, activation, dominance, two subjects)



Methodology

Transfer learning using a pretrained Model

Similarly to Emile et. al, the DeBERTaV3-Small model, published by Microsoft, from the transforms library provided by Hugging Face was used to try to achieve some good results 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 enhancedmask decoder. With those two improvements, DeBERTa outperforms RoBERTa on a majority of Natural Language Understanding (NLU) tasks with 80GB training data.



Training

Loss and Optimizer

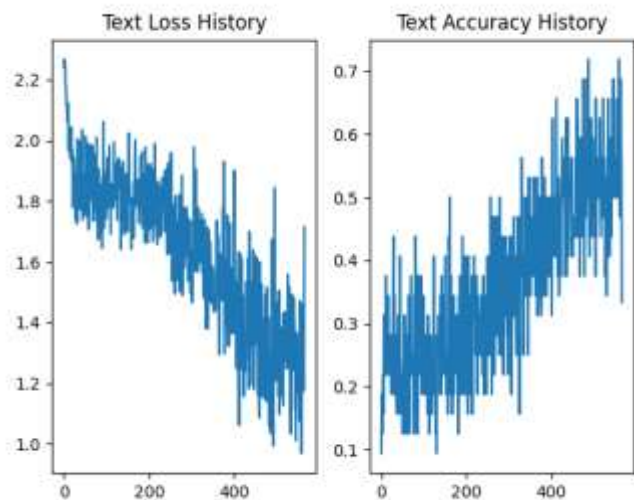
For the loss, the crossentropy loss was used, being one of the most used losses for classification tasks. The chosen optimizer was ADAM (Adaptive Moment Estimation), because of the fast convergence.

Mini-Batch Training

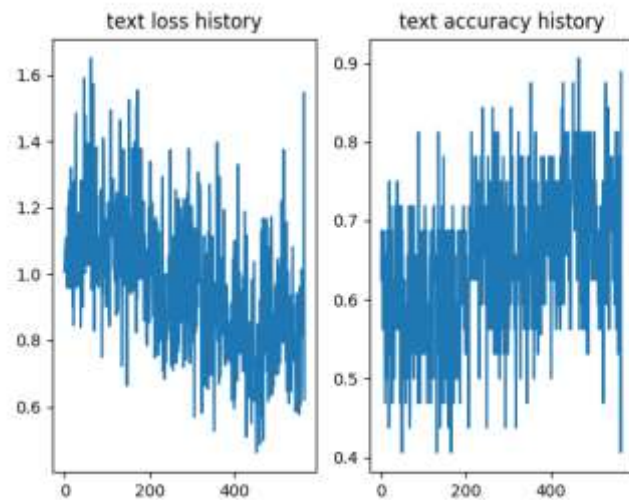
Unlike most of the sequential models which are trained with a stochastic approach (because of the fact that the sequential data could have different sizes, the stochastic approach is known to take longer to train and, some times, to be worse on generalizing data) this model was trained on mini-batches with batch size 128, this was achieved by padding the sequences with 0s as a preprocessing step to the longest sequence in the training set, so the model can learn to "ignore" those numbers.



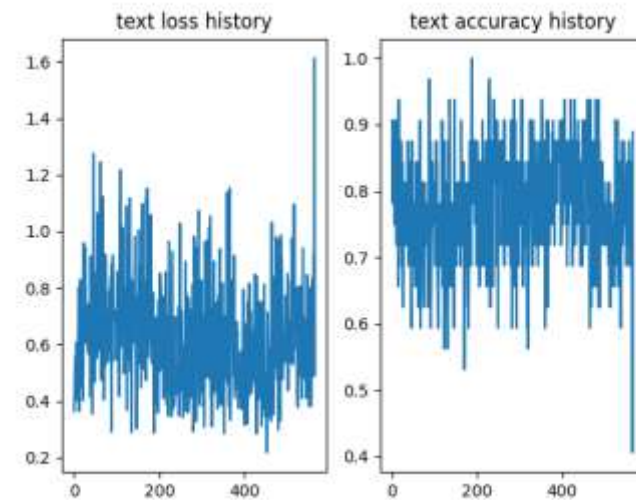
Results



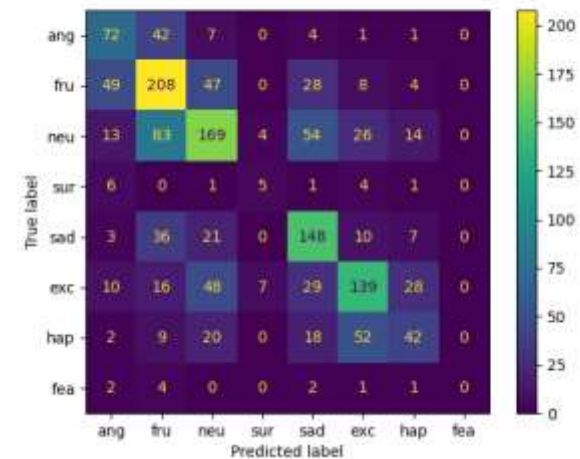
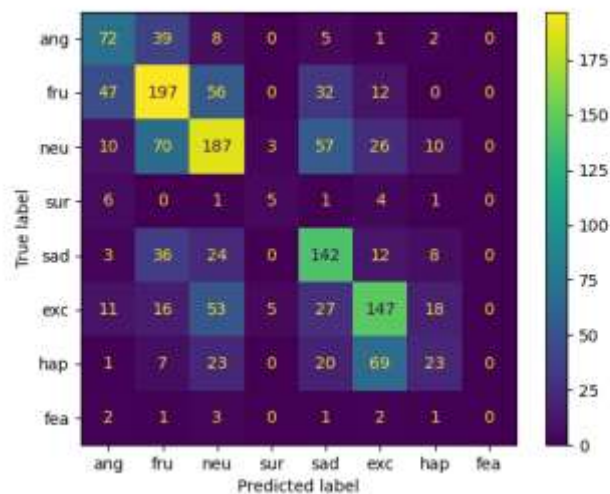
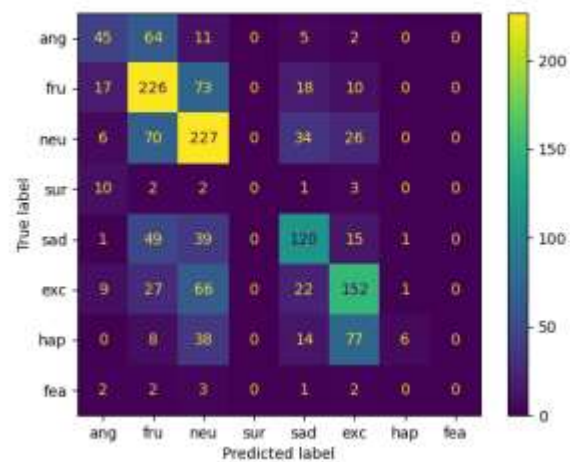
Epochs 1 -> 3



Epochs 4 -> 6



Epochs 7 -> 9



Results interpretation

- The model was trained for 9 epochs, which took 450 minutes on my machine. Different metrics were taken 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.
- We can see that the loss function converges very well for the first 2 epochs, 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 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.

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

<i>Model</i>	<i>Accuracy</i>
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%

Conclusions & Future Works

- 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.



Thank you

Bibliography

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)