MMODEST - MULTI-MODAL DETECTION OF EMOTIONS USING SPEECH AND TEXT

Kailash Karthik, Jessica Huynh, Amith Ananthram

Columbia University
ks3740@columbia.edu
jyh2127@columbia.edu
aa4461@columbia.edu

ABSTRACT

Emotion Recognition is a challenging task for multiple reason – the abstract nature of human emotions; the context dependent relationship between words and the conveyed emotion; the multi-modal nature in which humans exhibit and understand emotions; and the lack of large annotated datasets to train machine learning models. While most of the historic research on emotions have focused on unimodal data, recent work has focused on combining the modalities of text, speech and vision for this task. In this paper, we present a system that leverages text and speech data for emotion recognition. The issue of limited data availability is tackled using a transfer learning approach, both for text and speech. We fine-tune a pre-trained speaker recognition speech model and text language model for this task and evaluate it on the IEMOCAP dataset. Add one sentence here about the results.

Index Terms— Emotion Recognition, Speech Processing, Natural Language Processing, Transfer Learning, BERT

1. INTRODUCTION

With the increased integration of AI-powered systems into human lives, affective computing has become an important aspect of human computer interaction. Human thoughts and actions are influenced by emotions and thus they play an important role in communication [1]. The ability to leverage context to understand emotions associated with verbal and non-verbal communication is an inherent ability of humans and is currently an important distinction between humans and machines [2]. Emotions depend on human psyche and physiology and is triggered by the perception of situations, people and objects and depends on the mental state of the perceiving individual at that moment [3]. Emotion exhibition and perceptions across demographic factors like age, gender, race, culture and accent [4]. The lack of temporal boundaries between emotions and the variability in manner of expression make this a challenging task [5].

The detection of emotions has social and commercial applications that makes furthers its importance. Emotion detection has social applications in the medical domain - for the identification and diagnosis of depression and stress in individuals [6] [7]. It has been used to assist people with bipolar disorder [8]. Commercial applications like call center customer management [5], advertising through neuro-marketing [1] and social media engagement [2] have benefited from this domain of research. Emotions have become important to the design and building of AI chat systems and the field of paralinguistics [9].

Most of the early research on emotion understanding focused on unimodal analysis, formulating this as yet another classification task on speech [10] [11] or text or images [12]. This led to the development of early techniques for text sentiment analysis which is a simpler task of classifying a sentence as showing positive or negative sentiment [13]. Speech classifiers focused on the vocal dynamics to capture the emotions contained in the utterances. The drawbacks of such techniques are that humans emotions are non-binary in nature and humans often employ multiple modalities while interacting with people and the environment [1]. Thus, the binary granularity is not sufficient for complex applications and a lot of information is lost when communication is viewed with a single modality. The solution thus is to learn the interaction between the multiple modalities through the combination of different modes of information.

While visual information is often important, leveraging speech and text for emotion detection makes the technique suitable for a wider range of applications like on-call therapies and home assistants. Moreover, most vision techniques for emotion depend on facial expressions and this further restricts the scope of applications in which the model can be used. Thus, we propose an emotion understanding system that combines vocal characteristics of the speaker and the semantic content of the utterance. Modern machine learning classifiers are based on neural networks which require a large dataset to generalize well. Though there are multimodal emotion datasets like IEMoCap and MELD, the size of these datasets is restrictive since even though speech data is relatively easy to collect, large-scale annotation of emotions is labor and cost intensive. We thus propose a transfer learning technique to overcome this predicament.

Contemporary research has focused a lot on using transfer learning architectures. State of the art text processing models like ULMFiT [14] and BERT [15] have proved that transfer learning can be successfully applied in language tasks. The cascade-correlation architecture starts with a small neural network, adding and training additional layers to detect different features. The existing network layer weights are frozen before the addition of the next layer [16]. A domain specific model can thus be created, though pre-existing models can't be leveraged in this framework. Progressive networks [17] combat this issue by using the outputs from pre-trained models as embeddings for a new, randomly initialized model. A connection is made between the two to share information and the weights through combined backpropogation. This follows the previously mentioned cascade-correlation architecture in that it freezes the previous output.

In this paper, we propose a novel transfer learning technique to classify emotions using the modalities of text and speech independently and in unison. For the speech classifier, we train a a speaker recognition model and fine-tune it on the emotion detection task. The text classifier is of similar nature and consists of a pre-trained language model that is then used for classifying emotions. A number of experiments have been conducted to demonstrate the relative merits of various unimodal and multimodal classification techniques.

The rest of this paper is structured as follows. In the following section, we review the datasets used for this task. We describe the details of our proposed methodology in Section 3. The experimental setup is put forth in Section 4. The empirical results and their analysis are presented in Section 5. Finally, we conclude this paper in Section 6 along with am roadmap for future work.

2. DATASETS

2.1. VoxCeleb2 - Speaker Recognition

VoxCeleb2 is a natural speech dataset containing video and audio clips of 6,112 celebrities which have been uploaded to Youtube [18]. This dataset has more than 1 million utterances for these speakers. It is a fairly well-balanced dataset with a 66 percent male speakers proportion and a variety in the ethnicity of the speakers. It also has an abundance of background noise in the audio recordings. The speaker identification model trained on this dataset will serve as the pre-trained base model for our transfer learning approach. It is observed that a model trained on noisy natural speech does just as well on elicited and acted datasets collected in a noise-free setting.

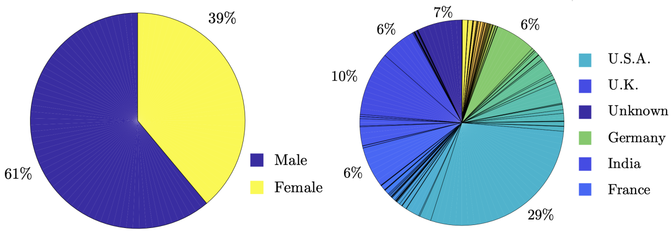
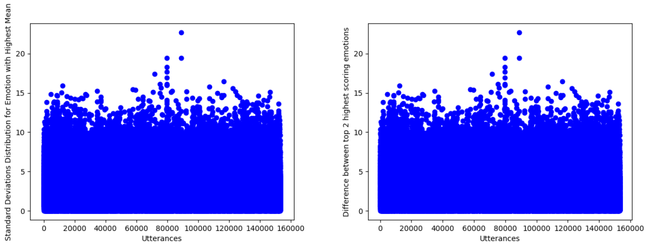


Fig. 1. VoxCeleb2 Data Distribution from [22]

The utterances in the VoxCeleb1, a subset of the VoxCeleb2 dataset, have been classified for emotions using the facial expressions of the speakers. A VGGFace2 based vision model fine-tuned to classify emotions was used to predict the emotions in every frame associated with each utterance. Though the similarity in the domains of the data and its volume make it an enticing dataset to use for our task, empirical evaluations show that its emotion classification has a couple of limitations that make it unusable. From figure 2 (a), it can be observed that the emotion softmax scores are not stable and have a high standard deviation within an utterance. Figure 2 (b) illustrates that in many utterances, the difference between the mean scores of the top 2 emotions is negligible thus making the emotion labels ambiguous to an extent.



(a) (b)

Fig. 2. EmoVoxCeleb Label Analysis

2.2. Wikipedia and Book Corpus - Language Modeling

Book Corpus [19] is a dataset consisting of 11,038 unpublished books from 16 different genres. The original corpus aimed to align books with their movie releases for a richer semantic interpretation of both the text and the video. The Wikipedia dataset contains a 2.5 billion word dump from the website and has been used heavily for training unsupervised text models.

2.3. MELD - Emotion Detection Fine-tuning

Multimodal EmotionLines Dataset (MELD) is an extension of the EmotionLines dataset [20] and contains dialogues from the television show Friends [21]. It is a multi-modal dataset with text, speech and video. MELD dialogues have multiple participants and the size of the dataset is around 1400 dialogues and 13000 utterances. The emotion classification labels used are Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear. MELD also contains a binary sentiment classification label though we don't use it for this task. The data consists of more than 250 speakers and 4000 emotion shifts inside dialogues making it apt for the task at hand. The dataset is well balanced with an even distribution among the emotion labels.

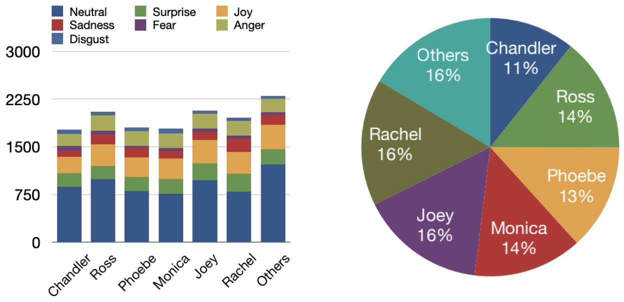


Fig. 3. MELD Data Distribution from [21]

2.4. IEMOCAP - Emotion Detection Testing

The final dataset used is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [22]. The dataset consists of 12 hours of video and audio clips and has a composition of 5 male and 5 female actors. The data consists of the actors participating in improvisations or scripted conversations, both designed to elicit emotions.These clips are then annotated for the dimensional attributes and for categorical attributes: anger, happiness, excitement, sadness, frustration, fear, surprise, other, and neutral. Utterances for which human annotator agreement could not be reached are labelled xxx and the fraction of the dataset with this label is around 25%. Most state-of-the-art model statistics are calculated off this dataset and thus we use it to evaluate our model. Evaluation on acted speech also proves the robustness of the model as it was trained on natural speech.

3. FORMATTING YOUR PAPER

All printed material, including text, illustrations, and charts, must be kept within a print area of 7 inches (178 mm) wide by 9 inches (229 mm) high. Do not write or print anything outside the print area. The top margin must be 1 inch (25 mm), except for the title page, and the left margin must be 0.75 inch (19 mm). All *text* must be in a two-column format. Columns are to be 3.39 inches (86 mm) wide, with a 0.24 inch (6 mm) space between them. Text must be fully justified.

Label Type	Utterances	%
Discrete Emotion	7169	74.56
Unknown (*other*)	3	~0
Unclassified (*xxx*)	2443	25.4

Table 1. IEMOCAP Emotion Label Distribution

4. PAGE TITLE SECTION

The paper title (on the first page) should begin 1.38 inches (35 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information. Please note that papers should not be submitted blind; include the authors' names on the PDF.

5. TYPE-STYLE AND FONTS

To achieve the best rendering both in printed proceedings and electronic proceedings, we strongly encourage you to use Times-Roman font. In addition, this will give the proceedings a more uniform look. Use a font that is no smaller than nine point type throughout the paper, including figure captions.

In nine point type font, capital letters are 2 mm high. **If you use the smallest point size, there should be no more than 3.2 lines/cm (8 lines/inch) vertically.** This is a minimum spacing; 2.75 lines/cm (7 lines/inch) will make the paper much more readable. Larger type sizes require correspondingly larger vertical spacing. Please do not double-space your paper. TrueType or Postscript Type 1 fonts are preferred.

The first paragraph in each section should not be indented, but all the following paragraphs within the section should be indented as these paragraphs demonstrate.

6. MAJOR HEADINGS

Major headings, for example, "1. Introduction", should appear in all capital letters, bold face if possible, centered in the column, with one blank line before, and one blank line after. Use a period (".") after the heading number, not a colon.

6.1. Subheadings

Subheadings should appear in lower case (initial word capitalized) in boldface. They should start at the left margin on a separate line.

6.1.1. Sub-subheadings

Sub-subheadings, as in this paragraph, are discouraged. However, if you must use them, they should appear in lower case (initial word capitalized) and start at the left margin on a separate line, with paragraph text beginning on the following line. They should be in italics.

7. PRINTING YOUR PAPER

Print your properly formatted text on high-quality, 8.5 x 11-inch white printer paper. A4 paper is also acceptable, but please leave the extra 0.5 inch (12 mm) empty at the BOTTOM of the page and follow the top and left margins as specified. If the last page of your paper is only partially filled, arrange the columns so that they are evenly balanced if possible, rather than having one long column.

In LaTeX, to start a new column (but not a new page) and help balance the last-page column lengths, you can use the command "\pagebreak" as demonstrated on this page (see the LaTeX source below).

8. PAGE NUMBERING

Please do **not** paginate your paper. Page numbers, session numbers, and conference identification will be inserted when the paper is included in the proceedings.

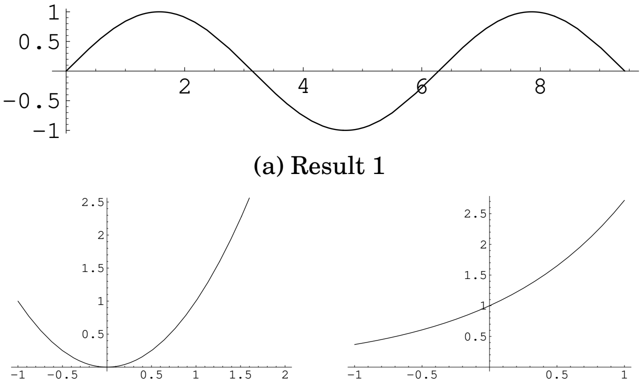
9. ILLUSTRATIONS, GRAPHS, AND PHOTOGRAPHS

Illustrations must appear within the designated margins. They may span the two columns. If possible, position illustrations at the top of columns, rather than in the middle or at the bottom. Caption and number every illustration. All halftone illustrations must be clear black and white prints. Colors may be used, but they should be selected so as to be readable when printed on a black-only printer.

Since there are many ways, often incompatible, of including images (e.g., with experimental results) in a LaTeX document, below is an example of how to do this [**?**].

10. FOOTNOTES

Use footnotes sparingly (or not at all!) and place them at the bottom of the column on the page on which they are referenced. Use Times 9-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).



(a) Result 1

(b) Results 3

(c) Result 4

Fig. 4. Example of placing a figure with experimental results.

11. COPYRIGHT FORMS

You must submit your fully completed, signed IEEE electronic copyright release form when you submit your paper. We **must** have this form before your paper can be published in the proceedings.

12. RELATION TO PRIOR WORK

The text of the paper should contain discussions on how the paper's contributions are related to prior work in the field. It is important to put new work in context, to give credit to foundational work, and to provide details associated with the previous work that have appeared in the literature. This discussion may be a separate, numbered section or it may appear elsewhere in the body of the manuscript, but it must be present.

You should differentiate what is new and how your work expands on or takes a different path from the prior studies. An example might read something to the effect: "The work presented here has focused on the formulation of the ABC algorithm, which takes advantage of non-uniform time-frequency domain analysis of data. The work by Smith and Cohen [**?**] considers only fixed time-domain analysis and the work by Jones et al [**?**] takes a different approach based on fixed frequency partitioning. While the present study is related to recent approaches in time-frequency analysis [3-5], it capitalizes on a new feature space, which was not considered in these earlier studies."

13. REFERENCES

[1] Chan Woo Lee, Kyu Ye Song, Jihoon Jeong and Woo Yong Choi. Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data, 2018; arXiv:1805.06606.

[2] Feiyang Chen, Ziqian Luo and Yanyan Xu. Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis, 2019; arXiv:1904.08138.

[3] Samarth Tripathi, Sarthak Tripathi and Homayoon Beigi. Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning, 2018; arXiv:1804.05788.

[4] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak and Julien Epps. Direct Modelling of Speech Emotion from Raw Speech, 2019; arXiv:1904.03833.

[5] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Björn Schuller and Stefanos Zafeiriou. End-to-End Multimodal Emotion Recognition using Deep Neural Networks, 2017; arXiv:1704.08619. DOI: 10.1109/JSTSP.2017.2764438.

[6] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, Automated depression diagnosis based on deep networks to encode facial appearance and dynamics, IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 578–584, 2018.

[7] R. Rana, S. Latif, R. Gururajan, A. Gray, G. Mackenzie, G. Humphris, and J. Dunn, Automated screening for distress: A perspective for the future, European Journal of Cancer Care, p. e13033, 2019.

[8] R. Rana, Context-driven mood mining, MobiSys 2016 Companion-Companion Publication of the 14th Annual International Conference on Mobile Systems, Applications, and Services. Association for Computing Machinery (ACM), 2016, p. 143.

[9] Seunghyun Yoon, Seokhyun Byun and Kyomin Jung. Multimodal Speech Emotion Recognition Using Audio and Text, 2018; arXiv:1810.04635.

[10] Vladimir Chernykh, Grigoriy Sterling, and Pavel Prihodko, Emotion recognition from speech with recurrent neural networks, arXiv preprint arXiv:1701.08071, 2017.

[11] Michael Neumann and Ngoc Thang Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” arXiv preprint arXiv:1706.00612, 2017.

[12] A. Dhall, R. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In Proceedings of the 17th International Conference on Multimodal Interaction, ICMI '15. ACM, 2015.

[13] Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836.

[14] Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification, 2018; arXiv:1801.06146.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018; arXiv:1810.04805.

[16] S. E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," Advances in neural information processing systems, 1990, pp.524-532

[17] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," arXiv preprint arXiv:1706.03256, Jun. 2017

[18] S. Chung, A. Nagrani, and A. Zisserman, Vox-celeb2: deep speaker recognition, arXiv preprint arXiv:1806.05622, Jun. 2018.

[19] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, 2015; arXiv:1506.06724.

[20] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang and Lun-Wei Ku. EmotionLines: An Emotion Corpus of Multi-Party Conversations, 2018; arXiv:1802.08379.

[21] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations, 2018; arXiv:1810.02508.

[22] C. Busso, et. al, IEMOCAP: Interactive emotional dyadic motion capture database, Language resources and evaluation, 42(4), pp. 335-359.