

Analyzing the Influence of Dataset Composition for Emotion Recognition

Alexander Sutherland¹, Sven Magg¹, Cornelius Weber¹, Stefan Wermter¹

Abstract— Recognizing emotions from text in multimodal architectures has yielded promising results, surpassing video and audio modalities under certain circumstances. However, the method by which multimodal data is collected can be significant for recognizing emotional features in language. In this paper, we address the influence data collection methodology has on two multimodal emotion recognition datasets, the IEMOCAP dataset and the OMG-Emotion Behavior dataset, by analyzing textual dataset compositions and emotion recognition accuracy. Experiments with the full IEMOCAP dataset indicate that the composition negatively influences generalization performance when compared to the OMG-Emotion Behavior dataset. We conclude by discussing the impact this may have on HRI experiments.

I. INTRODUCTION

One of the key components of affective computing is the ability to recognize emotion features in humans. Numerous approaches focus on combining features from multiple input modalities [1]. The most common modalities are centred on features collected from video and audio stimuli since these features are the easiest to collect and label. Recent experiments on multimodal emotion recognition have made prolific use of features of language and text. Poria et al. [4] in particular have shown impressive results using CNNs when combining text with other modalities on experimental data, however, little attention has been spent on how and why textual features are able to perform so well in a multimodal context when compared to visual and acoustic features.

We address this issue by analyzing how the composition of the IEMOCAP dataset influences neuron activation during classification and show how current usage of the IEMOCAP manifests in visible over-fitting. Experimental results indicate that a network pretrained on the IEMOCAP barely performs as well as a network pretrained on the OMG-Emotion Behavior dataset [7] when attempting to classify the other dataset, in spite of being a larger dataset.

II. DATASET DESCRIPTIONS

We make use of the IEMOCAP dataset [2] and the OMG Emotion-Behavior dataset [7]. The IEMOCAP dataset consists of videos containing either scripted or improvised utterances. Videos are recorded in different sessions, with different actors in each session. We merge the “Happy” and “Excited” data samples in the IEMOCAP in order to be able to emulate previous results from Poria et al. [4] and

TABLE I
MEAN 10-FOLD ACCURACY AND STANDARD DEVIATION FOR EMOTION RECOGNITION ON SELECTED LABELS FROM THE IEMOCAP, ONLY IEMOCAP IMPROVISED SAMPLES, AND OMG TEXTUAL DATA USING 1D-CNNs FOR WORD EMBEDDINGS, SEMANTICS, AND A FUSION.

	IEMOCAP	IEMOCAP Improv.	OMG
Words	67.504 (1.436)	62.738 (2.490)	44.200 (1.377)
Semantics	52.672 (2.306)	47.330 (3.512)	39.371 (1.337)
Fusion	67.718 (2.235)	63.075 (1.876)	43.557 (2.165)

to even out class balance. We use 5616 transcribed labelled utterances from the dataset with 1117 *Angry* samples, 1644 *Happy* and *Excited* samples, 1753 *Neutral* samples, and 1102 *Sad* samples. The set of emotions chosen for this task is based on the data used in previous experiments on the same task [4].

The OMG Emotion-Behavior dataset [7] consists of videos labelled through crowd-sourcing. Each video is annotated within the context of a longer clip, with both categorical and continuous emotion labels. Categorical labels include *Anger*, *Neutral*, *Happy*, *Sad*, *Surprise*, *Fear*, and *Disgust*. We use 4656 samples consisting of 639 *Sad*, 665 *Anger*, 1794 *Neutral*, and 1558 *Happy* samples to compare a network’s performance on the IEMOCAP for the same emotions. Videos were selected using a web-crawler that searched for YouTube videos, with keywords such as “monologue” and “acting”, with each video being split into utterances.

III. EXPERIMENTAL DATASET ANALYSES

This experiment determines the influence of the IEMOCAP dataset construction method on emotion recognition from text. Previous works have found emotion features at both word [4] and semantic levels [3]. We use the pre-trained Google News Word Embedding [6] and a semantic embedding of size 50 trained on frames, from SEMAFOR [5], unimodally and multimodally as input to a 1D-CNN architecture. The architecture has channels for each input with a dropout of 0.2, a temporal convolution layer with a kernel size of 3, a stride of 1, and a filter size of 150, a global max pooling layer, a penultimate fully connected layer of size 32 and a final softmax layer of size 4. Late fusion between inputs is performed via concatenating max pooling outputs before feeding them to the fully connected layer. Classification accuracies from 10-fold cross-validations with 8-1-1 splits can be seen in Table I for both datasets, and Table II presents the results of dataset generalization accuracy.

¹Knowledge Technology, Department of Informatics, University of Hamburg, Germany, {sutherland, magg, weber, wermter}@informatik.uni-hamburg.de

TABLE II

MEAN 10-FOLD ACCURACY AND STANDARD DEVIATION WHEN PRETRAINING THE CNN ON THE DATASET TO THE LEFT OF THE ARROW AND TESTING ON THE RIGHT FOR TEXT EMOTION RECOGNITION. DATASETS USED ARE THE IEMOCAP, **I**, THE IMPROVISED SAMPLES OF THE IEMOCAP, **I IMPROV.**, AND THE OMG DATASET, **O**.

	I → O	I Improv. → O	O → I
Words	34.487 (0.862)	35.195 (1.174)	35.589 (0.710)
Semantics	30.483 (1.152)	33.251 (0.766)	31.193 (1.059)
Fusion	35.217 (1.636)	35.612 (1.023)	35.185 (0.547)

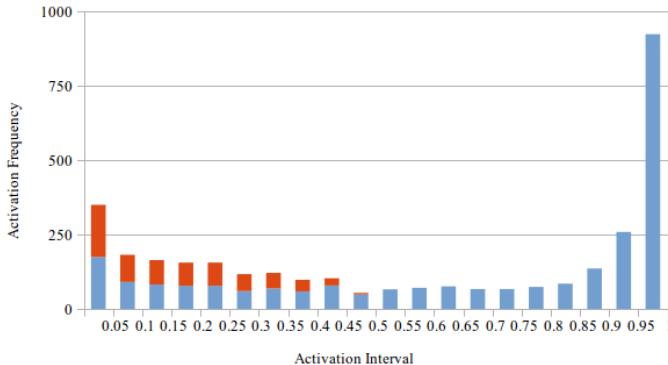


Fig. 1. Scripted utterance softmax activation frequencies predicted by the utterance-level unimodal word model for the 2645 IEMOCAP scripted utterances. The blue bars indicate the number of target label activation levels occurring within particular activation brackets. Red bars indicate the number of incorrect classifications when a target label neuron activation is in a particular bracket. There is no red bar for activations above 0.45, as this entails that the classification was correct, i.e. above 0.5.

IV. RESULTS & DISCUSSION

In Table I we see the results of emotion classification accuracy for input feature combinations. Clearly word embeddings are the primary contributor to classification results and that the network attains higher accuracy on the IEMOCAP as opposed to the OMG. Table II shows results of pretraining the network on one dataset and classifying samples from the other. We see that the full IEMOCAP performs equal to or worse than the OMG and IEMOCAP improvised samples when generalizing, despite the full IEMOCAP's high performance on its own test data.

We believe this over-performance during testing but under-performance during generalization was due to the subset of scripted values in the IEMOCAP. To support this, in Figure 1, we visualize the activation values attained when classifying scripted IEMOCAP utterances during the process of 10-fold cross-validation. We see abnormally high neuron activations in the 0.95 to 1 bracket, indicating a high certainty of correctness. The reason for this over-confidence is visualized in Figure 2, where we show that scripted IEMOCAP utterances have a high number of data-points that partially overlap with others due to sentences being scripted and nearly identical. This is likely the reason for the textual modalities high performance in previously reported multimodal architectures that used the IEMOCAP dataset.

In summary, we have shown a reason for textual modalities

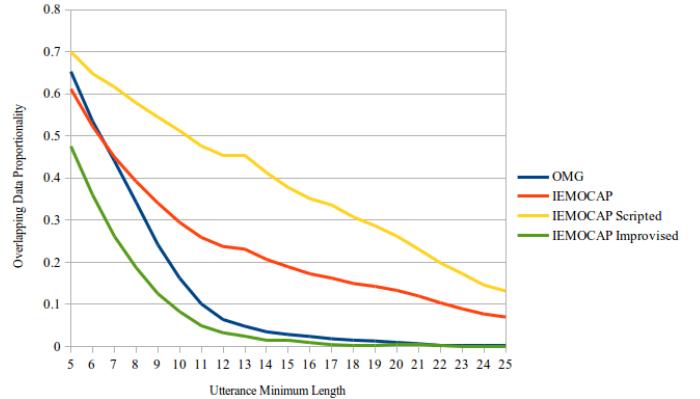


Fig. 2. The proportion of data that overlaps with some other data point with regard to a certain number of shared words. The X-axis describes the number of words required for a sentence to be considered and the Y-axis describes what proportion of considered sentences overlap with some other sentence. The IEMOCAP scripted sessions suffer from the highest level of overlap while the IEMOCAP improvised sessions suffer from the least.

high performance in previous works using the IEMOCAP in this manner. This over-fitting will also influence the understanding of natural language in applications, including HRI scenarios. Robot agents could place far greater salience in specific words than is desirable, such as the word "beast" in the IEMOCAP only occurring in angry data-points, leading to angry prone classifications. In future work, we suggest that researchers proceed with caution when selecting and applying datasets, as incorrect training procedures can lead to illogical and unreliable behaviour from models.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 (SOCRATES).

REFERENCES

- [1] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, (pp. 98-125).
- [2] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4), p.335.
- [3] Wu, C. H., Chuang, Z. J., & Lin, Y. C. (2006). Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2), (pp. 165-183).
- [4] Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016, December). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on* (pp. 439-448). IEEE.
- [5] Das, D., Schneider, N., Chen, D., & Smith, N. A. (2010). SEMAFOR 1.0: A probabilistic frame-semantic parser. *Language Technologies Institute, School of Computer Science, Carnegie Mellon University*.
- [6] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. Chicago.
- [7] Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A. and Wermter, S., 2018. The omg-emotion behavior dataset. *arXiv preprint arXiv:1803.05434*.