Toward Emotion Recognition From Early Fused Acoustic and Language Features Using Recursive Neural Networks

Alexander Sutherland

Abstract— **Recognising emotions from language is considered an important aspect of affective computing. However, the application of recognised emotions in an effective manner is often bound to the context where the emotion was detected, without the acquisition of information about the relation between spoken words and the recognised emotion. To apply recognised emotions to a broader context, knowledge about this dynamic must be accrued during the emotion classification process. In this paper, we outline a novel method of extracting these relations, using recursive neural networks to process the syntactic structure of speech in order to better understand how emotions are expressed and what spoken words they relate to.**

I. Introduction

Recognizing and responding to human emotions in HRI scenarios is regarded as important for acceptance of robots [4]. Acceptance and empathy is vital for long-term sustainable Human-Robot relations, as users are liable to reject or ignore robots they feel no connection with. To perform recognition, emotional expressions from multiple modalities, such as vision, audio, and language, are often combined to improve recognition accuracy and system robustness [3].

Current multimodal emotion recognition using speech reduces features of language to an abstract level for more convenient processing. While this simplifies processing, it makes the role of language structure more implicit than explicit and in this simplification, some information is lost. Structureless processing of language also requires the structure to be relearned as an emergent property to facilitate the understanding of relationships between words. While this is possible, the syntactic structure of language is already well defined and should not require relearning and running the risk of potential errors in syntactic understanding.

Through reintroduction of the structure of language into the categorical emotion recognition pipeline, we hope to visualize over the syntax graph how detected emotion expressions relate to acoustics and language, in contrast to what Socher et al. [2] did for sentiment on only language.

II. Recursive Neural Networks

Recursive Neural Networks, RvNNs, [1] are able to process structured data and therein learn feature patterns that occur in the structure of data. This allows RvNNs to use structure as a feature rather than having to relearn structure as an emergent property of the network. The fundamental difference between recursive and recurrent architectures is that recurrent neural networks, RNNs, are a special case of RvNNs that only handle a linear chain of input, whereas a RvNN may take an arbitrary number of inputs from a previous time-step. This is often described in the manner of a bottom-up analysis of a hierarchical tree structure, computing the values of parent nodes based on the nodes of their respective children.

Language has an explicit structure that can be exploited by RvNNs to provide a more nuanced understanding of how humans express sentiment based on the structure of language used [1], [2]. The benefit of using RvNNs for this is the increased granularity of class predictions over the syntax tree. Parent nodes are a product of their children and resulting classifications can be traced back to specific sub-branches within the syntax tree. An example is shown in the work of Socher et al. [2], where the authors show the negating word "not" influences the final sentiment classification. Our novel contribution will be the extension of this approach to incorporate acoustic data in a uni-modal and multi-modal fashion, with text, over categorical labels as opposed to sentiment labels. We expect that this will provide insight on how language and pronunciation influences emotion recognition through visualization over the syntax graph.

III. Proposed Method

In this position paper, a method of using RvNNs to process fused language and acoustic features will be outlined. An overview of how the system will process data can be seen in Figure 1. The composition function, g, will be the one used by Socher et al. [1], as it has shown promise when classifying sentiment. In Figure 1 we see that the network calculates intermediate probabilities in a bottom-up fashion, also allowing for predictions of individual nodes and parent nodes as the network works its way through the graph.

Predictions are attained through a projection layer that learns how to convert intermediate representations to a probability distribution over target emotions. Once the probability of every emotion class for every node is calculated this can be visualised in a tree by choosing the highest probable emotion. An example where this is useful is determining what phrases are responsible for emotional outcomes and motivating why different decisions were made based on occurring phrases and emotion predictions in each node.

Utterances with transcripts will be selected from the IEMOCAP dataset [8], a multimodal emotion recognition dataset. For each utterance, every word will be aligned with it's associated audio segment. Word features will be attained through pretrained word embeddings [6] and for audio

Knowledge Technology, Department of Informatics, University of Hamburg, Germany, sutherland@informatik.uni-hamburg.de

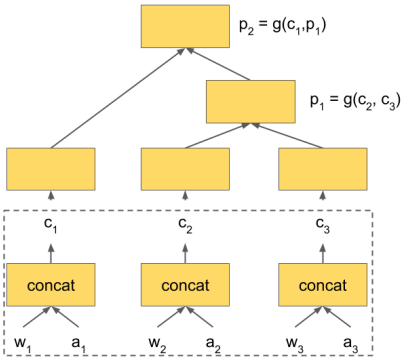

Fig. 1. This figure shows how a recursive neural network would process a three word sentence with paired audio and a specific syntactic structure. Here the word embeddings, w_i, are concatenated with extracted audio features, a_i, corresponding to each word. This results combined feature vector, c_i, that is then fed to the recursive network which calculate target probabilities, p_j, using the composition function g.

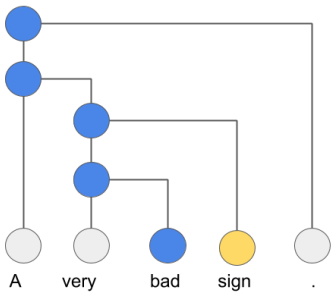

Fig. 2. RvNN output for categorical emotion recognition over the syntactic graph of a simple sentence. Different colours represent the highest predicted emotion in a particular node in the sentences syntax tree. Blue is sadness, yellow is happiness, and grey is emotionally neutral.

MFCC features [7] will be fed in sequence to a pre-trained LSTM to extract audio features for every word. Thereafter we will perform early fusion through concatenation of word and audio representations and feed the subsequent vectors to the RvNN based on syntactic structure to perform emotion recognition. Syntactic structure will be extracted through the use of standard NLP libraries available in Python.

IV. Expected Results

Expected results of applying a RvNN for this task will allow us to attain emotion predictions, using acoustic and language features, for every node in the syntactic graph of an input utterance. This will allow us to see which syntactic sub-trees contribute to emotion classifications. A possible additional outcome would be a higher emotion classification accuracy from acoustics alone and combined multimodally than with standard recurrent approaches.

Currently, we are able to show preliminary examples of using the text modality alone. In Figures 2 and 3 we see the results of applying a RvNN to the task of categorical emotion recognition, as opposed to sentiment classification. To do this we translate the labels of the Stanford Sentiment Treebank (SST) [2] to categorical labels, wherein positive labels are translated to happy, neutral to neutral, and negative labels to either angry or sad based on the classification of

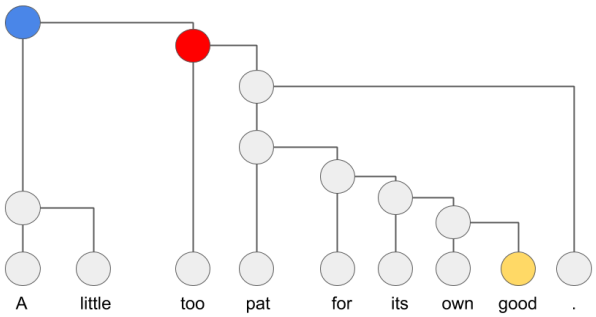

Fig. 3. An example of RvNN output showing that it is able to capture how certain sequences of words are able to shift the classification between negative categories over a syntax graph. Here the words “A little” shift the prediction from an overall angry classification (red) to a more sombre sad classification (blue).

an LSTM pretrained on the IEMOCAP dataset [8] for categorical emotion recognition. The SST is used for exemplary purposes and will be replaced by the IEMOCAP when the required syntax trees have been generated.

V. Conclusion

In this paper, we show preliminary work toward a novel method of processing language and audio features for improving visualisation and understanding of emotion recognition using recursive neural networks. We also visualised how categorical emotion predictions distribute themselves over the syntax graphs of two simple sentences.

Acknowledgements

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

References

[1] Socher, R., Lin, C.C., Manning, C. and Ng, A.Y., 2011. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 129-136).

[2] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A. and Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

[3] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A. and Morency, L.P., 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) (Vol. 1, pp. 873-883).

[4] Lim, A. and Okuno, H.G., 2015. A recipe for empathy. International Journal of Social Robotics, 7(1), pp.35-49.

[5] Tai, K.S., Socher, R. and Manning, C.D., 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

[6] Pennington, J., Socher, R. and Manning, C., 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[7] Logan, B., 2000, October. Mel frequency cepstral coefficients for music modeling. In ISMIR (Vol. 270, pp. 1-11).

[8] 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.