Interspeech 2018
2-6 September 2018, Hyderabad

Attention-based sequence classification for affect detection

Cristina Gorrostieta, Richard Brutti, Kye Taylor, Avi Shapiro, Joseph Moran, Ali Azarbayejani, John Kane

Cogito Corporation

{cgorrostieta,rbrutti,ktaylor,ashapiro,jmoran,ali,jkane}
@cogitocorp.com

Abstract

This paper presents the Cogito submission to the Interspeech Computational Paralinguistics Challenge (ComParE), for the second sub-challenge. The aim of this second sub-challenge is to recognize self-assessed affect from short clips of speech-containing audio data. We adopt a sequence classification-based approach where we use a long-short term memory (LSTM) network for modeling the evolution of low-level spectral coefficients, with added attention mechanism to emphasize salient regions of the audio clip. Additionally to deal with the underrepresentation of the negative valence class we use a combination of mitigation strategies including oversampling and loss function weighting. Our experiments demonstrate improvements in detection accuracy when including the attention mechanism and class balancing strategies in combination, with the best models outperforming the best single challenge baseline model.

1. Introduction

The area of affective computing, and in particular recognition of emotion from voice, has received continually increasing attention in recent years. This has been in large part due to the emergence of functional applications which incorporate emotion sensing, with applications in healthcare, in the call center [1], driver state monitoring, e-learning [2], and music recommendation systems [3], to name a few.

At the same time, there remain significant challenges to speech-based emotion recognition. One major challenge is producing consistently accurate recognition of emotional valence (i.e. positive vs negative emotions). Fusing lexical and acoustic representations has been demonstrated to be a promising approach to addressing this [4], however for many applications automatic speech recognition is not available (or introduces too much latency in the processing). As a result, improving recognition accuracy of emotional valence from acoustic inputs alone remains an important open challenge.

Emotion recognition models are known to suffer from problems with generalizability, with poor accuracy when models are applied to new acoustic environments or to new speakers. There have been some recent advances, with some approaches using knowledge sharing techniques via multi-task learning [5], however this still remains an outstanding problem for the field. A likely cause of these issues regarding generalizability is the lack of availability of large, suitable datasets. This is particularly problematic given the increasing use of complex neural network models, involving extremely high numbers of parameters which generally require large volumes of training data in order to be optimally effective. There have been some recent initiatives to tackle this problem and to create suitably large and varied datasets [6]. Further efforts along these lines will likely lead to improvements in detection accuracy. Solving data availability is arguably the most clear and present opportunity to improving the accuracy of speech emotion recognition.

Additionally, as extreme emotions are typically somewhat rare compared to neutral, on various emotional dimensions (e.g., valence, activation/arousal, dominance), class imbalance tends to be a major problem and some mitigation strategies to class imbalance can often lead to significant overfitting. The effect of class imbalance mitigation strategies has been studied previously in the context of modern neural network models [7], where the authors found moderate oversampling of the low-prevalence class to be the most effective method for dealing with this problem.

Neural network models based on recurrent layers, in particular long-short term memory (LSTM) layers, have been shown to be effective at modeling emotion sequentially [8] (i.e. sequence-to-sequence modeling). However, LSTM-based models can also be extremely effective at classifying static sequences (i.e. sequence-to-one modeling) [9]. Additionally the use of an attention mechanism in recurrent neural network based models is an effective approach for encouraging the model to focus on, and to weight more heavily, certain regions of a sequence [10, 11]. This approach has also been previously applied to the problem of emotion recognition from speech [9, 12, 13].

This paper presents our submission to the Interspeech 2018 Computational Paralinguistics Challenge (ComParE) for the second sub-challenge on recognizing self-assessed affect. In addition to our submission entry, the main contribution of this paper is a systematic assessment of applying attention mechanism to sequence classification, separately and in combination with class imbalance mitigation strategies.

2. Proposed method

We base our experimentation around a neural network architecture, shown schematically in Figure 1. The model takes low level spectral coefficients as input to recurrent neural network layers, which are used to model the temporal sequence of these spectral vectors. The temporal dependencies in the input sequence are modeled using a recurrent layer (or layers) consisting of LSTM cells. In contrast to vanilla recurrent units, LSTM units are capable of learning long-range dependencies without suffering from vanishing (or exploding) gradients [14, 15].

An attention mechanism is applied to the sequence modeling output at each time step in order to weight regions in the sequence which are more salient for the detection of emotional valence. The key idea behind the attention mechanism is to learn a function, $f(h)$ parametrized by θ_a, which maps from each recurrent layer output at time step i to a weight vector α_i. α is used to determine the size of the effect of each time step on subsequent layers in the network.

10.21437/Interspeech.2018-1610

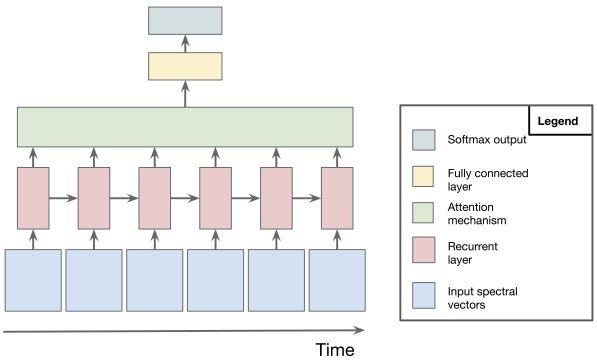

Figure 1: *Schematic illustration of the neural network model architecture used*

Note that whereas the parameters of $f(h)$, i.e. θ_a, are optimized during training, the weight vector α is determined freshly for each new time step input. In the present work we define this function as:

$$f(\mathbf{h}) = \tanh(\mathbf{W}\mathbf{h} + \mathbf{b}) \tag{1}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the parameters of a linear function and in our model $\mathbf{h}$ corresponds to the output of the final LSTM layer, $\mathbf{h} = (h_0, \ldots, h_{T-1})$ (where T is the number of time steps for a given sequence). Note that a more complex function could be used, however we found a linear function with $\tanh$ activation to be sufficient. For a given time step, i, the α_i weight vector is computed as:

$$\alpha_i = \frac{\exp\left(f\left(h_i\right)\mathbf{u}\right)}{\sum_j \exp\left(f\left(h_j\right)\mathbf{u}\right)} \tag{2}$$

In the equations above, $\mathbf{W}$ is a matrix of dimension $S \times D$ (where D is the number of units on final LSTM layer), $\mathbf{u}$ and $\mathbf{b}$ are vectors of size S. These variables comprise the learnable parameter set θ_a. We henceforth refer to S as the attention size.

Application of the attention mechanism is computed as follows:

$$g = \sum_{j=0}^{T-1} \alpha_j h_j \tag{3}$$

where the resulting vector g is input into a fully connected layer with rectified linear activation, before a final fully connected layer with a softmax function applied to the output.

Given the low prevalence of the negative valence class in the dataset we employ two class imbalance mitigation strategies. The first we refer to as "oversampling", and this involves repeating the negative samples some multiple of times for the training set. Following experimentation, we found that simply including instances of each negative valence samples twice in the training set was effective in improving accuracy for the underrepresented class without resulting in overfitting.

The second class imbalance mitigation strategy involved loss function weighting. In our experimentation we optimize model weights, θ, by minimizing either a standard or weighted cross-entropy function. The standard cross-entropy loss can be written as:

$$\mathcal{L}_n(\theta) = -log\ \hat{\mathbf{y}}_{c_n}(\mathbf{X_n}, \theta) \tag{4}$$

where $\hat{\mathbf{y}}_{c_n}(\mathbf{X_n}, \theta)$ is the softmax model output for each observed sequence, n, given the parameters θ, and where X is a $F \times D$ matrix (with F indicating the number of input spectral coefficients per time step). c_n is the class label for observation n, in the present study this is in the range $\{0,1,2\}$ (corresponding to negative, neutral and positive emotional valence classes). We additionally define a class-weighted cross entropy loss function as:

$$\mathcal{L}w_n(\theta) = -w_{c_n} log\ \hat{\mathbf{y}}_{c_n}(\mathbf{X_n}, \theta) \tag{5}$$

where the weight vector $\mathbf{w}^k \in \mathcal{R}^k$, $w_k > 0$ and where $k \in \{0, 1, 2\}$ (i.e. the three class labels). In our experiments when using the weighted version of the loss function we set w to (2.0, 1.0, 1.0).

3. Experimental protocol

3.1. Data

A full description of the challenge data can be found in [16]. To summarize, the data consisted of 2,313 8-second long audio clips with 846 clips in the training set, 742 in development and 724 in test. Each clip has a single categorical target of negative, neutral or positive valence (as derived from self-assessed continuous ratings). There was significant underrepresentation of the negative valence class, with the negative class making up only 11.2 % of the training samples (compared with 45.8 % for neutral and 42.9 % for positive classes).

Although training targets are derived from ratings by the speakers themselves (and hence are in and of themselves accurate), we wanted to get a sense for how perceivable the emotion labels were from listening to the audio. We had three US English speaking human raters (who had no German language abilities) listen to 830 of the challenge clips. We computed inter-rater agreement for the three US raters, which resulted in a Krippendorff's alpha of 0.243 (indicating moderately low agreement). The US raters individually agreed with the German target data between 25 % and 38 % of the time. Naturally the language difference explains a certain amount of this disagreement. Nevertheless, this data does provide some indications that there is a somewhat tenuous perceptual agreeability between the audio and the target labels. We intend to carry out further experimentation with German-speaking annotators, and with text-only stimuli (via transcripts of the audio recordings), to further assess how perceivable these target labels are and to what extent the lexical mode contributes to this.

Audio was pre-processed by downsampling to an 8 kHz sampling rate with 16-bit precision, prior to feature extraction.

3.2. Features

For all experiments we use Mel coefficients, by computing an FFT magnitude spectrum on 40 ms long Hamming-windowed audio frames (with a 16 ms period) and applying a set of 40 triangular filters linearly spaced on the Mel scale. The magnitude spectrum was clipped to a floor of 10^{-8} prior to application of the filter bank to avoid unstable computations in silence regions.

Individual spectral coefficients are normalized to have 0-centered distributions, with unit variance, by applying z-score normalization. The mean and variance statistics (which are parameters of the z-score normalization) were computed for each coefficient once, using the frame-level observations computed on the entire training set.

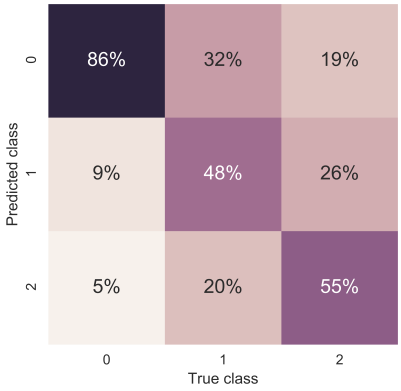

Figure 2: *Confusion matrix (presented as %) on the development set for the best performing model resulting from experiments in this paper, for negative (0), neutral (1) and positive (2) valence classes.*

3.3. Experimental settings and variants

We fixed experimental settings presented at Table 1. Such parameters were defined after sampling combinations of learning rate (in te range 0.05, 0.005, 0.0005, 0.0005), dropout keep probability (in the range 0.5, 0.7, 0.8) and number of fully connected units (in the range 20,50,100) and evaluating the results on the development set. It was considered the Unweighted Average Recall(UAR) as reference validation metric to match Paralinguistics Challenge baseline metric. All training experiments were carried out using TensorFlow code on Tesla V100 hardware, involving a single GPU.

Table 1: *Model and training settings used consistently across experimental variants.*

Settings	Value
Optimization algorithm	Adam [17]
Learning rate	0.0005
# LSTM cells	128
# fully connected units	20
Dropout keep probability	0.7

During training, model parameters are saved and retained separately after each epoch. The parameters associated with the epoch which produces the highest UAR metric on the development, are used as the final model configuration and results are reported based on this best performing epoch.

We investigated the effect of the number of LSTM layers, the attention mechanist and the effect of oversampling and loss function weighting class imbalance mitigation strategies. To that end experiments involve varying the number of LSTM layers, in the range 1 or 2; varying the size of the attention mechanism (i.e. the S parameter), 1 or 2; as well as testing the effect of oversampling and loss function weighting class imbalance mitigation strategies, varying to try with and without class imbalance strategies.

4. Results

The results for the various experiments are summarized in Table 2. At the top of the table are metrics for the challenge baseline models. We include here both the best performing individual model and (on the development set) and the best performing ensemble model.

Figure 3 supplies supplementary information on the evolution of the UAR metric for the development set across training epochs.

Balancing, i.e. a combination of oversampling of the negative class, and loss function weighting, individually provide a moderate improvement in accuracy with UAR of 51.5 % and 57.7 % respectively, compared to 48.9 %. Considerable complementarity across these two additions is observed, as when used in combination a UAR of 63.1 % is achieved on the development set. Note that this improvement is not translated on test set. One of the reasons of the observed gap between development and test set is the amount of parameter tuning performed on the development dataset. The more evaluations performed on parameter selection on the development set, the greater the chance of finding a parameter combination leading to an apparent significant improvement.

From Figure 3, we can observe the change of fluctuations in the development UAR metrics. We also observed comparable behavior in the development loss values (not reported here). These large fluctuations are likely due to a combination of a fairly small overall dataset as well as “fuzziness” of the target, possibly owing to the low perceptual agreeability between the audio and the labels.

The confusion matrix for the best performing model in terms of development set from our experiments is shown in Figure 2. This demonstrates the high recall produced by the class imbalance mitigation strategies although this was at the expense of precision. These strategies are also likely to be the reason why the neutral class is the one with the highest degree of confusion. Note that high recall obtained on development set is not consistent at test set, indicating how class imbalance mitigation approach could also contribute to the observed overfitting.

5. Discussion and conclusion

The Interspeech 2018 Computational Paralinguistics Challenge (ComParE) sub-challenge on self-assessed affect presented a difficult speech emotion recognition problem. The combination of a relatively small dataset with the underrepresentation of the negative valence class as well as the ”fuzziness” of the target labels, compared with the perception of the audio, contributed to making this an extremely challenging machine learning task. Nevertheless, we demonstrated that by applying oversampling of the underrepresented class, weighting the loss function, and using an attention mechanism applied to the LSTM outputs, an improved detection accuracy across all three classes (compared to the best individual challenge baseline model) could be achieved. The attention mechanism is particularly effective at ignoring regions in the audio that do not contribute to the inference of emotional valence, and at heightening the contribution of salient regions. Hence, our findings corroborate previous applications of this approach to emotion recognition from speech [9, 12, 13]. The present research builds on these findings by demonstrating that when used in combination with oversampling and loss function weighting, that the benefits of the attention mechanism can be enhanced when there are unbalanced classes (as is often the case with these emotion recog-

Table 2: *Summary of evaluation metrics across the various experiments. Precision and recall metrics are displayed separately per class (negative valence class \ neutral valence class \ positive valence class). The "Settings" column indicates the number of LSTM layers and the size of the attention mechanism (i.e. the S parameter). In the "Case" column, balancing refers to a combination of oversampling and loss function weighting.*

Case	**Settings**	**Development set**			**Test set**		
		UAR (%)	Precision	Recall	UAR (%)	Precision	Recall
Baseline (best individual)	–	**56.5**	–	–	61.7	–	–
Baseline (best ensemble)	–	–	–	–	66.0	–	–
Case #1 LSTM-based model	1 LSTM	44.6	0/53/67	0/66/68	–	–	–
	2 LSTM	46.3	18/52/66	28/48/63	–	–	–
Case #2 +balancing	1 LSTM	48.6	17/53/68	63/30/53	–	–	–
	2 LSTM	51.5	21/56/61	72/26/57	–	–	–
Case #3 +attention	1 LSTM, attention 2	47.6	18/54/69	37/46/61	–	–	–
	2 LSTM, attention 1	57.7	21/60/86	76/53/42	–	–	–
	2 LSTM, attention 2	49.4	22/55/80	18/77/54	–	–	–
Case #4 +attention+balancing	1 LSTM, attention 2	**63.1**	29/60/74	86/48/55	47.28	16/65/58	21/39/81
	2 LSTM, attention 1	59.3	26/58/72	80/44/54	47.49	16/64/58	24/37/80
	2 LSTM, attention 2	62.6	26/66/76	85/46/57	48.90	15/63/63	30/36/80

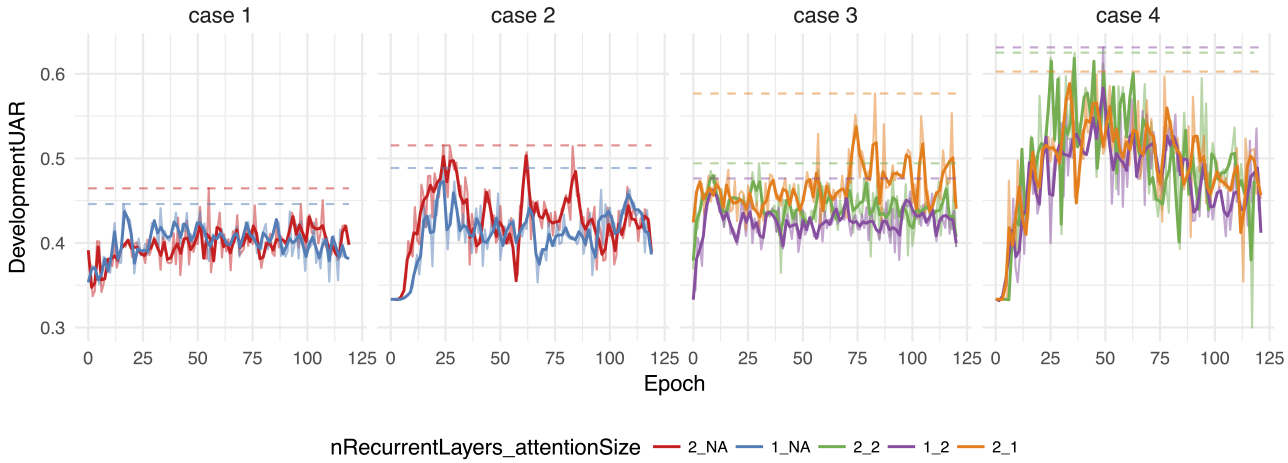

Figure 3: *UAR development set curves across training epochs for the various experiments.*

nition problems).

At the same time, we observed considerable volatility in training these models due to the fairly small dataset (only 5 hours in total). However, it can often happen that there is a small amount of training data for a new language. As a result and given applications which seek to apply emotion sensing to variety of languages and contexts, our future work will explore multi-task learning and other techniques to share knowledge and feature representations learnt from languages where there are larger sources of suitable data (e.g., US English [6]), which we hope can result in improved accuracy for tasks such as that presented in this challenge. We also intend to investigate more efficient ways of applying these types of models to processing of conversational speech. In such interactions, one or both parties may be silent from periods of time, and as a result applying recurrent neural network based models to extended periods of silence is computationally wasteful. By applying a hierarchical approach where we only process regions of contiguous speaking but still exploit the power of sequential modeling techniques, we hope to find more appropriate architectures for efficiently recognizing emotion in two-party interactions.

6. References

[1] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, Y Auno, "Hierarchical LSTMs with Joint Learning for Estimating Customer Satisfaction from Contact Center Calls," *Proceedings of Interspeech*, 2017.

[2] N. Banda, P. Robinson, "Multimodal Affect Recognition in Intelligent Tutoring Systems" *D'Mello S., Graesser A., Schuller B., Martin JC. (eds) Affective Computing and Intelligent Interaction. Lecture Notes in Computer Science*, vol 6975, 200-207, 2011.

[3] B. Ferwerda, M. Schedl, "Enhancing Music Recommender Systems with Personality Information and Emotional States: A Proposal." *UMAP Workshop proceedings*, 2014.

[4] Z. Aldeneh, S. Khorram, D. Dimitriadis, E. Mower-Provost, "Pooling acoustic and lexical features for the prediction of valence", *Proceedings of ICMI*, 2017.

[5] B. Zhang, E. Mower-Provost, G. Essl, "Cross-corpus Acoustic Emotion Recognition with Multi-task Learning: Seeking Common

Ground while Preserving Differences", *IEEE Transactions on Affective Computing*, 2017.

[6] R. Lotfian, C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings", *IEEE Transactions on Affective Computing*, 2017.

[7] M. Buda, A. Maki, M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks", *arXiv:1710.05381*, 2017.

[8] J. Lee, I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition", *Proceedings of Interspeech*, 2015.

[9] M, Seyedmahdad, E. Barsoum, C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention.," *Proceedings of ICASSP*, 2017.

[10] L., Minh-Thang, H. Pham, C. D. Manning., "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[11] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, "Hierarchical attention networks for document classification," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[12] M. Neumann, V. Ngoc Thang, "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.

[13] H. Che-Wei, S. S. Narayanan, "Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition," *Proceedings of Interspeech*, 2016.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol 9 number 8, 1735-1780, 1997.

[15] Y. Bengio and P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult", *IEEE Transactions on Neural Networks*, vol 5 number 2, 157-166, 2011.

[16] B. Schuller et al., "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats" *Proceedings of Interspeech*, 2018.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization" *arXiv preprint arXiv:1412.6980*, 2014.