Speech Emotion Classification using Raw Audio Input and Transcriptions

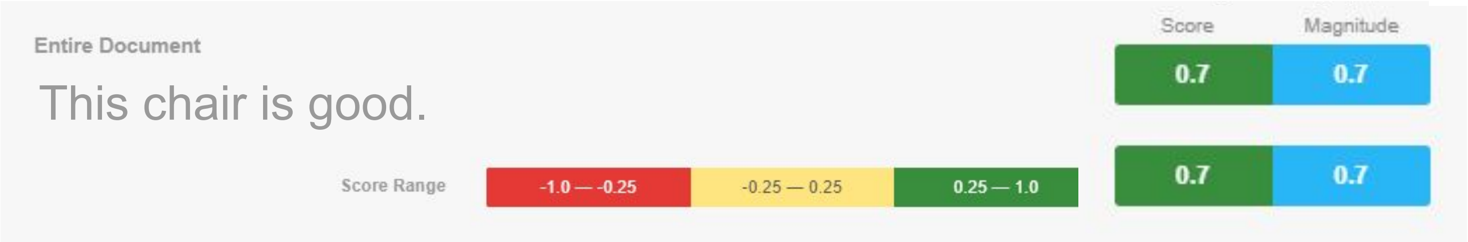
Gabriel Lima and JinYeong Bak
KAIST
{gcamilo, jy.bak}@kaist.ac.kr

Motivation

Systems have increasingly been controlled by voice and they can understand **WHAT** was said or asked.

Motivation

However, systems still lack empathy because they cannot interpret **HOW** the communication was portrayed.



What if I were angry? What if I were sad?

Motivation

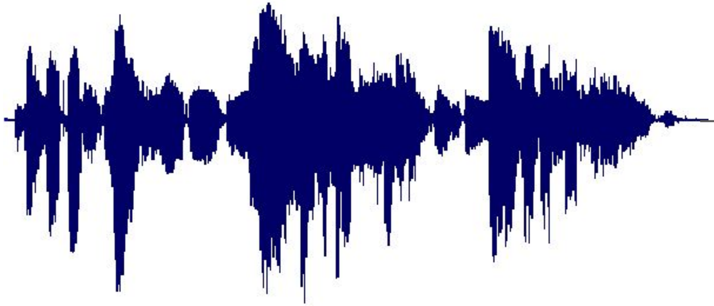
Emotion Classification: Text + Audio + Video. [Rao et al., 2015]

Acoustic features are often extracted using tools not embedded into the classification model. [Poria et al., 2017]

Raw audio waveforms achieved great results for speech generation, modelling and recognition. [Van den Oord et al., 2016; Hoshen et al., 2015]

Attention models focus on the most important features. [Xiao et al., 2015]

Proposal



I’ve been so happy lately.

Feature Fusion
+
Classification

Anger

Neutrality

Happiness

Sadness

Contributions

1) Deep learning model that **classifies emotional speech using raw audio waveforms and transcriptions.**
2) Model capable of **extracting features from raw audio waveforms.**
3) **Interpretability study** in the classification task.
4) Analysis of possible emotional words in the IEMOCAP dataset.

Dataset - IEMOCAP

10 speakers

5000+ utterances

State of the art (acoustic + textual features) → 0.721 accuracy. [Hazarika et al., 2018]

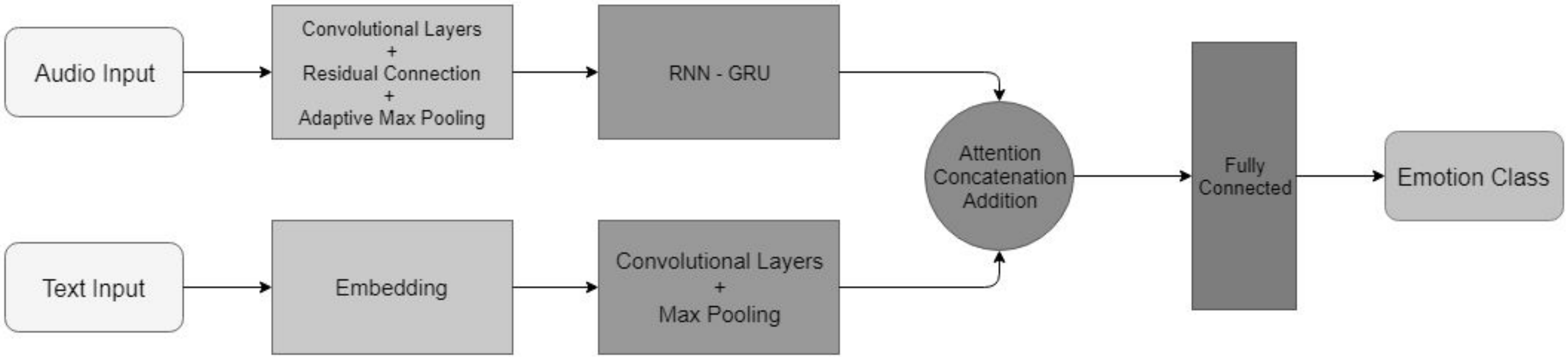
Anger

Happiness

Neutrality

Sadness

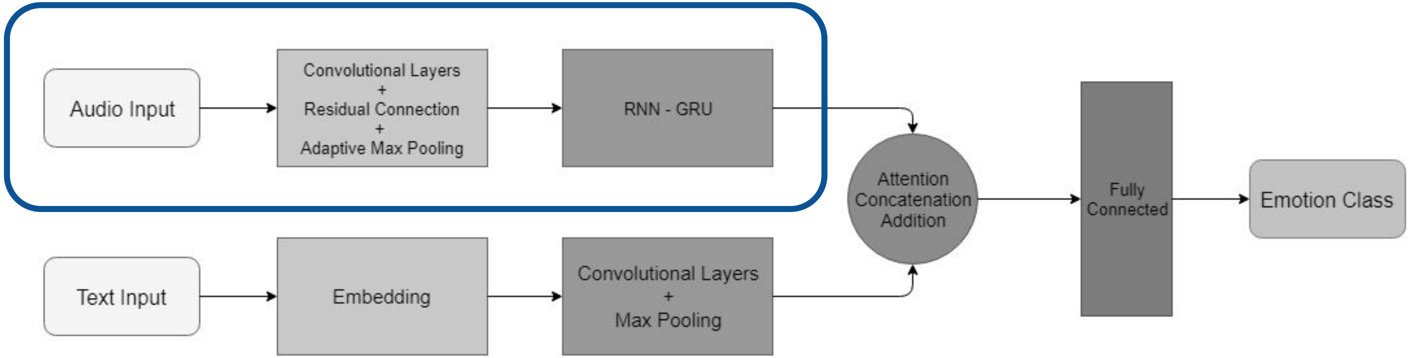
Model


Audio Input
Convolutional Layers
+
Residual Connection
+
Adaptive Max Pooling
RNN - GRU
Text Input
Embedding
Convolutional Layers
+
Max Pooling
Attention
Concatenation
Addition
Fully Connected
Emotion Class

Audio Model

Feature extraction:

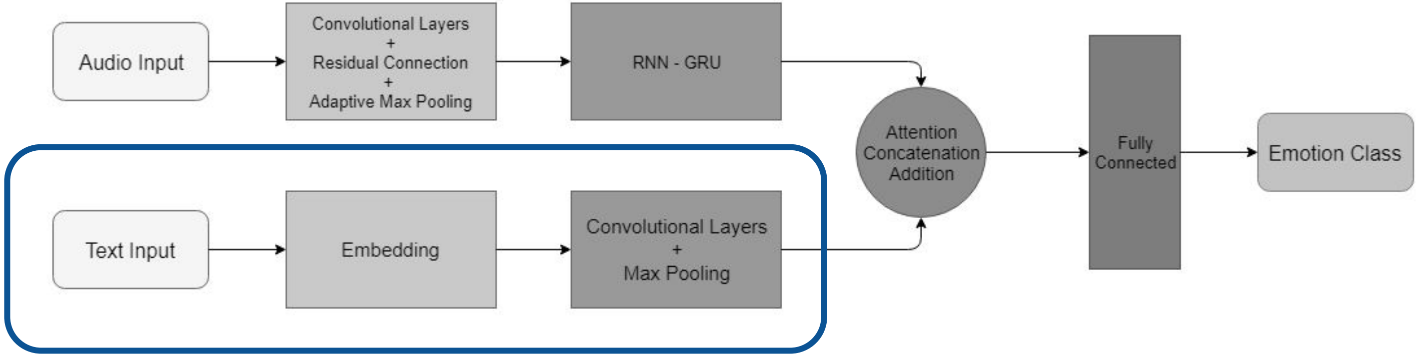
- Convolutional layers.
- Adaptive pooling.
- GRU - RNN.



Text Model

Utterance embedding:

- Subsampling: $P_{drop}(w_i) = sqrt(frequency(w_i)/t)$
- Trainable and pre-trained word embeddings.
- Convolutional layers with max pooling. [Kim, 2014]

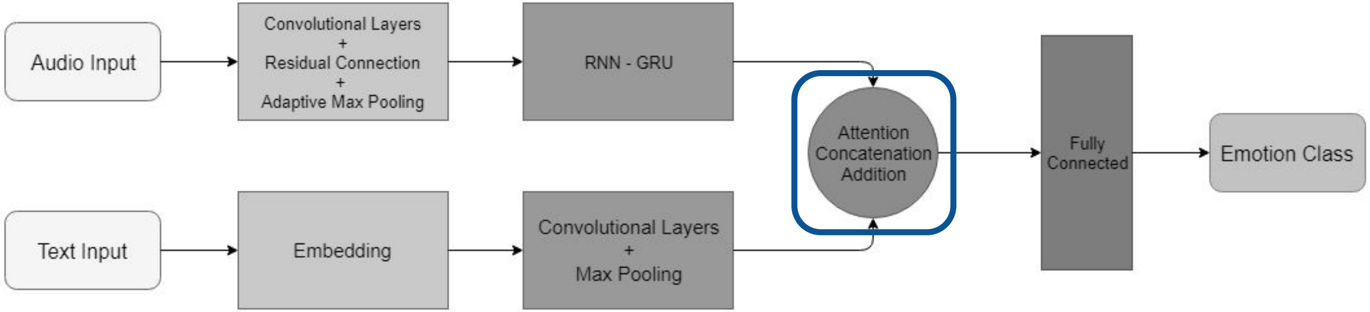


Combining Text and Audio

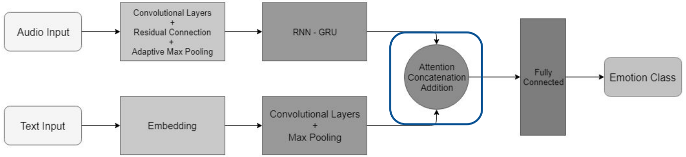
Concatenation and addition:

$$y = x_{text} \oplus x_{audio}$$

$$y = x_{text} + x_{audio}$$



Combining Text and Audio



Attention:

$$Att_s^1: a_i = f(W_1^{1\times n} x_i)$$

$$Att_d^1: a_i = W_2^{1\times n} f(W_1^{n\times n} x_i)$$

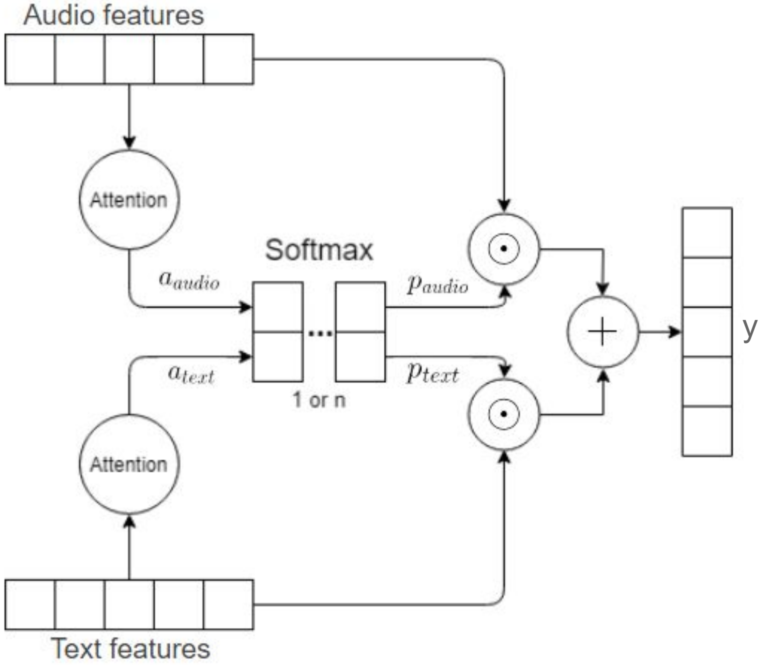
$$Att_s^n: a_i = f(W_1^{n\times n} x_i)$$

$$Att_{d1}^n: a_i = W_1^{n\times n} f(W_1^{n\times n} x_i)$$

$$Att_{d2}^n: a_i = W_2^{n\times n} f(W_1^{n\times n} x_i)$$

$$p = softmax([a_{text} \quad a_{audio}])$$

$$y = p_{text} \odot x_{text} + p_{audio} \odot x_{audio}$$



Results

Trainable word embeddings

Model	Accuracy	Score (Audio / Text)
Att_s^1	0.679	(0.516 / 0.484)
Att_d^1	**0.703**	(0.509 / 0.491)
Att_s^n	**0.703**	(0.500 / 0.500)
Att_{d1}^n	0.694	(0.499 / 0.501)
Att_{d2}^n	0.672	(0.500 / 0.500)
Concat	0.695	—
Addition	0.676	—
uSA [5]	0.721	Not available
mSA [5]	0.714	Not available
Audio [5]	0.541	—
Text [5]	0.625	—

Pre-trained word embeddings

Model	Accuracy	Score (Audio / Text)
Att_d^1	**0.715**	(0.423 / 0.577)
Att_s^n	0.688	(0.500 / 0.500)

Results

Mutual information and ratio between predictions and appearances in testing set:

Emotion	Top Words	Ratio
Anger	not	0.462
	she's	0.800
	business	0.667
Happiness	laughter	0.867
	oh	0.704
	so	0.592
Neutral	um	0.800
	can	0.667
	uh	0.650
Sadness	they	0.023
	else	0.857
	the	0.121

Discussion

Overfitting → small size of dataset

We believe our model can learn more meaningful features with more data → higher ceiling for improvement

SQuAD2.0

The Stanford Question Answering Dataset

Discussion

Non emotional words mispredictions → lack of data and context

Speech fillers → do not support emotional speech

Laughter tag → happy part of speech

Emotion	Top Words	Ratio
Anger	not	0.462
	she's	0.800
	business	0.667
Happiness	laughter	0.867
	oh	0.704
	so	0.592
Neutral	um	0.800
	can	0.667
	uh	0.650
Sadness	they	0.023
	else	0.857
	the	0.121

Discussion

Non emotional words mispredictions → lack of data and context

Speech fillers → do not support emotional speech

Laughter tag → happy part of speech

Emotion	Top Words	Ratio
Anger	not	0.462
	she's	0.800
	business	0.667
Happiness	laughter	0.867
	oh	0.704
	so	0.592
Neutral	um	0.800
	can	0.667
	uh	0.650
Sadness	they	0.023
	else	0.857
	the	0.121

Discussion

	Emotion	Predictions				Accuracy	Recall
Labels	Anger	**80**	9	14	8	0.721	0.721
	Happiness	16	**117**	25	6	0.713	**0.801**
	Neutrality	13	18	**109**	31	0.637	0.673
	Sadness	2	2	14	**91**	**0.835**	0.669

Neutrality is hard to classify → central part of speech.

Happiness and Sadness are the best performers.

Discussion

	Emotion	Predictions				Accuracy	Recall
Labels	Anger	**80**	9	14	8	0.721	0.721
	Happiness	16	**117**	25	6	0.713	**0.801**
	Neutrality	13	18	**109**	31	0.637	0.673
	Sadness	2	2	14	**91**	**0.835**	0.669

Neutrality is hard to classify → central part of speech.

Happiness and Sadness are the best performers.

Discussion

But no words classified as sad actually have an emotional connotation.

Emotion	*WaveAvg*	*WaveAvgStd*	*WaveStdAvg*
Anger	0.034	0.039	0.062
Happiness	0.019	0.023	0.034
Neutral	0.009	0.007	0.017
Sadness	**0.005**	**0.004**	**0.009**

Discussion

But no words classified as sad actually have an emotional connotation.

Emotion	*WaveAvg*	*WaveAvgStd*	*WaveStdAvg*
Anger	0.034	0.039	0.062
Happiness	0.019	0.023	0.034
Neutral	0.009	0.007	0.017
Sadness	**0.005**	**0.004**	**0.009**

Discussion

But no words classified as sad actually have an emotional connotation.

Emotion	*WaveAvg*	*WaveAvgStd*	*WaveStdAvg*
Anger	0.034	0.039	0.062
Happiness	0.019	0.023	0.034
Neutral	0.009	0.007	0.017
Sadness	**0.005**	**0.004**	**0.009**

Discussion

Black box models are hard to interpret → Attention.

Model	Accuracy	Score (Audio / Text)
Att_s^1	0.679	(0.516 / 0.484)
Att_d^1	**0.703**	(0.509 / 0.491)
Att_s^n	**0.703**	(0.500 / 0.500)
Att_{d1}^n	0.694	(0.499 / 0.501)
Att_{d2}^n	0.672	(0.500 / 0.500)
Concat	0.695	—
Addition	0.676	—
uSA [5]	0.721	Not available
mSA [5]	0.714	Not available
Audio [5]	0.541	—
Text [5]	0.625	—

Model	Accuracy	Score (Audio / Text)
Att_d^1	**0.715**	(0.423 / 0.577)
Att_s^n	0.688	(0.500 / 0.500)

Small standard deviation.

Pre-trained word embeddings were trained with extensive data.

Discussion

Black box models are hard to interpret → Attention.

Model	Accuracy	Score (Audio / Text)
Att_s^1	0.679	(0.516 / 0.484)
Att_d^1	**0.703**	(0.509 / 0.491)
Att_s^n	**0.703**	(0.500 / 0.500)
Att_{d1}^n	0.694	(0.499 / 0.501)
Att_{d2}^n	0.672	(0.500 / 0.500)
Concat	0.695	—
Addition	0.676	—
uSA [5]	0.721	Not available
mSA [5]	0.714	Not available
Audio [5]	0.541	—
Text [5]	0.625	—

Model	Accuracy	Score (Audio / Text)
Att_d^1	**0.715**	(0.423 / 0.577)
Att_s^n	0.688	(0.500 / 0.500)

Small standard deviation.

Pre-trained word embeddings were trained with extensive data.

Conclusion

We believe our deep learning model can learn more meaningful features with more data.

Neutrality vs. Happiness and Sadness.

Using embeddings trained with extensive data improved the model and increased their importance.

Future Work

Develop new dataset.

Explore **temporal dimensionality of speech** such as **context**.

Emotion	Top Words	Ratio
Anger	not	0.462
	she's	0.800
	business	0.667
Happiness	laughter	0.867
	oh	0.704
	so	0.592
Neutral	um	0.800
	can	0.667
	uh	0.650
Sadness	they	0.023
	else	0.857
	the	0.121

Future Work

Other machine learning methods for **better interpretability**.

Distilling a Neural Network Into a Soft Decision Tree

Nicholas Frosst, Geoffrey Hinton

Google Brain Team

Abstract. Deep neural networks have proved to be a very effective way to perform classification tasks. They excel when the input data is high dimensional, the relationship between the input and the output is complicated, and the number of labeled training examples is large [Szegedy et al., 2015, Wu et al., 2016, Jozefowicz et al., 2016, Graves et al., 2013]. But it is hard to explain why a learned network makes a particular classification decision on a particular test case. This is due to their reliance on distributed hierarchical representations. If we could take the knowledge acquired by the neural net and express the same knowledge in a model that relies on hierarchical decisions instead, explaining a particular decision would be much easier. We describe a way of using a trained neural net to create a type of soft decision tree that generalizes better than one learned directly from the training data.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

References

- Rao, K. S., & Koolagudi, S. G. (2015). Recognition of emotions from video using acoustic and facial features. Signal, Image and Video Processing, 9(5), 1029-1045.
- Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, *261*, 217-230.
- Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016, September). WaveNet: A generative model for raw audio. In *SSW* (p. 125).
- Hoshen, Y., Weiss, R. J., & Wilson, K. W. (2015, April). Speech acoustic modeling from raw multichannel waveforms. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4624-4628). IEEE.
- Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., & Zhang, Z. (2015). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 842-850).
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Frosst, N., & Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

Thank you!

Gabriel Lima
School of Computing - KAIST
gcamilo@kaist.ac.kr