Speech Emotion Classification using Raw Audio Input and Transcriptions

Gabriel Lima and JinYeong Bak
KAIST
Daejeon, Republic of Korea
+82 42 350 7749
{gcamilo, jy.bak}@kaist.ac.kr

ABSTRACT

As new gadgets that interact with the user through voice become accessible, the importance of not only the content of the speech increases, but also the significance of the way the user has spoken. Even though many techniques have been developed to indicate emotion on speech, none of them can fully grasp the real emotion of the speaker. This paper presents a neural network model capable of predicting emotions in conversations by analyzing transcriptions and raw audio waveforms, focusing on feature extraction using convolutional layers and feature combination. The model achieves an accuracy of over 71% across four classes: Anger, Happiness, Neutrality and Sadness. We also analyze the effect of audio and textual features on the classification task, by interpreting attention scores and parts of speech. This paper explores the use of raw audio waveforms, that in the best of our knowledge, have not yet been used deeply in the emotion classification task, achieving close to state of art results.

CCS Concepts

•**Computing methodologies** → **Machine learning; Neural networks;** *Machine learning approaches;*

Keywords

Emotion Classification; Feature Extraction; Signal Processing; Neural Networks; Convolutional Layers.

1. INTRODUCTION

With the introduction of new technologies, human computer interaction has increased immensely. Tasks such as asking your cellphone to automatically set calendar events or alarms, using a home assistant to control your appliances or just ordering food online with your voice have become routine. Everything was developed to improve and facilitate from the simplest to the most complex tasks people complete every day. However, these technologies still lack a basic human ability: empathy. In order to develop empathy, machines must at first be able to understand and analyze emotions from their users, allowing them to change their actions and speech according to the situation.

In this paper, we propose a model that extracts features from raw audio waveforms of speech and their transcriptions and classify them into emotion classes. By utilizing convolutional layers on audio waves and word embeddings, as proposed by [1; 2; 3], we can extract features that when combined together, through different forms, can classify speech's emotion.

In later sections, we analyze the importance of the extracted textual and audio features by interpreting attention scores for both elements. These attention scores represent how much focus the neural network should put on each feature, allowing it to efficiently utilize the most important ones. We also show words and expressions that the model has learned as characteristic for such emotions.

For all results on this paper, we used a multimodal dataset, named IEMOCAP [4], which consists of two-way conversations among 10 speakers. The conversations are then segmented into utterances that are annotated using 4 emotion classes: Anger, Happiness, Neutrality and Sadness. We used 10% of the dataset as testing set and the other 90% for training, resulting in 555 and 4976 utterances, respectively. To the best of our knowledge, the state of the art performance in the emotion classification task with this dataset achieves an accuracy of 0.721 [5].

The main contributions of this paper are 1) a deep learning model that classifies emotional speech into its respective emotion using raw audio waveforms and transcriptions, 2) an audio model capable of extracting audio features from raw audio waveforms, 3) a study of acoustic and textual features importance in the emotion classification task using attention models and 4) an analysis of possible emotional words in the IEMOCAP dataset. We tackle the human computer interaction popularity increase by proposing a model capable of classifying speech emotion in order to allow systems to understand users' emotions and adapt their behavior according to the user.

2. RELATED WORK

Many works in emotion recognition and classification use textual, acoustic and visual features as input for such task [6; 7]. Acoustic features from audio data, such as Mel-Frequency coefficients, are often extracted [8] using tools not embedded inside the classification model. However, raw audio waveforms have achieved great results on speech generation [9], modelling [10] and recognition [11], but have not yet been fully explored in emotion classification. The work with raw waveforms have surpassed and equalized previous results using features extracted outside the model in those

SPML '18, November 28–30, 2018, Shanghai, China

DOI: https://doi.org/10.1145/3297067.3297089

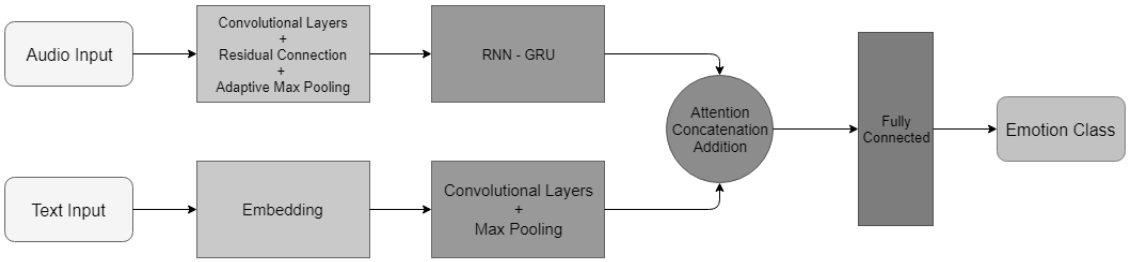

Figure 1. Graphical representation of our model.

areas. By using convolutional layers, we believe that it is possible to also extract features that can be useful for the classification task.

As showed by Hazarika et al. [5], it is possible to achieve better results on the emotion recognition task by mixing acoustic and textual features in different forms. The authors also analyzed many models that tackle feature-level fusion in emotion classification, such as attention, and their counterpart without such fusion, inspiring a part of this paper. Also, attention models have achieved good results in some tasks, such as image [12] and document [13] classification, by allowing models to focus on the most important features of the input.

3. MODEL

Our model has two different networks, one for the transcriptions and other for raw audio waveforms, that are chained together with different methods explained in Section 3.3. After combining the features extracted with such methods, we use a fully connected layer to classify the sentence into an emotion class. All networks are trained at the same time with Adam Optimizer, L2 regularization and using the PyTorch framework. The acoustic and textual features extraction models are presented in Section 3.1 and 3.2, respectively.

3.1 Raw Audio Waveforms

As explained in Section 2, raw audio waveforms have showed good results on speech generation [9], modelling [10] and recognition [11] and in this paper, we propose the use of these raw waveforms for audio classification. Even though authors have usually extracted features using outside tools, such as audio coefficients, for such task, we believe that convolutional layers, as shown not only in Computer Vision [14; 15] and Natural Language Processing [2], can learn meaningful and complex features for audio classification as well.

The raw audio waveforms are first padded with zeros in a batch, so they have the same length. Our model has two convolutional layers with different filter sizes (k_1 and k_2) and number of channels (c_1 and c_2), along with Batch Normalization, ReLU and a residual connection. The extracted features are then pooled with Adaptive Max Pooling that outputs a fixed n_{pool}-sized vector for each waveform.

We use a GRU layer to capture the temporal dependency of the waveform with n_{GRU} units. Finally, the extracted audio features are scaled to $x_{audio} \in \mathbb{R}^{2\times \text{emb\_size}}$ in order to be used in our attention model.

3.2 Transcriptions

Each word of the sentence is either embedded using an embedding matrix, trained alongside the model, or embedded using a pre-trained Word2Vec [16] to $x_{raw\_text} \in \mathbb{R}^{\text{emb\_size}}$. We use convolutional layers as proposed by Kim [2] to extract features from the sentences.

A sentence, which is the concatenation of words, of length n is first padded with zeros inside its batch and convoluted with a filter $w \in \mathbb{R}^{\text{emb\_size}\times \text{p}}, p \in \{f_1, f_2\}$. Each sentence is convoluted twice with the same number of channels n_{fms} and max-pooled, resulting in vectors $x_{text,i} \in \mathbb{R}^{\text{n}_{\text{fms}}}, i = 1,2$. The vectors $x_{text,i}$ are then concatenated to x_{text} and used for the attention model explained in Section 3.3.

We also use subsampling of frequent words in order to compensate for the imbalance between frequent and rare words: each word w_i is discarded with probability P_i as in Equation 1, where $f(w_i)$ is the frequency of the word in the dataset. We used the parameter t equal to 8000.

$$P_i(w_i) = \sqrt{f(w_i)/t} \tag{1}$$

3.3 Combining Text and Audio

As methods for combining text and audio, we propose attention models along with trivial concatenation (Equation 2) and addition (Equation 3). The attention models use as starting point the work of Hazarika et al. [5].

$$y = x_{text} \oplus x_{audio} \tag{2}$$
$$y = x_{text} + x_{audio} \tag{3}$$

In Equation 2, $\oplus$ represents trivial concatenation.

As for the attention models, we calculate the attention scores for textual and audio features by using matrix multiplication and projecting the scores on either $\mathbb{R}^{\text{n}}$, where n is the number of features, or $\mathbb{R}^1$. In the former, each feature has its own attention score, while in the latter, the attention score is shared across all dimensions.

Equations 4-8 show the methods for calculating the attention score a_i of audio and text features. Let $W_i^{m\times n}$ be $m \times n$ weights trained alongside the model, $x_i \in \mathbb{R}^{\text{n}}$ be either audio or text features and f the ReLU function.

$$Att_s^1 : a_i = f(W_1^{1\times n} x_i) \tag{4}$$
$$Att_d^1 : a_i = W_2^{1\times n} f(W_1^{n\times n} x_i) \tag{5}$$
$$Att_s^n : a_i = f(W_1^{n\times n} x_i) \tag{6}$$
$$Att_{d1}^n : a_i = W_1^{n\times n} f(W_1^{n\times n} x_i) \tag{7}$$
$$Att_{d2}^n : a_i = W_2^{n\times n} f(W_1^{n\times n} x_i) \tag{8}$$

The feature fusion is done as following:

$$p = softmax([a_{text} \quad a_{audio}]) \tag{9}$$

$$y = p_{text} \odot x_{text} + p_{audio} \odot x_{audio} \tag{10}$$

In equation (8), $\odot$ indicates element-wise multiplication. In the case of projection onto $\mathbb{R}^n$, we use the $softmax$ function on each dimension of a_i.

4. EXPERIMENTS

This section describes the experiments and results of the suggested model as well as other methods for classifying the emotions in IEMOCAP [4] dataset.

4.1 Experiment Setting

Table 1 shows the hyperparameters used during training.

Table 1. Training hyperparameters.

Hyperparameter	Value
Learning Rate	$3e^{-3}$
Learning Rate Decay	0.988 / epoch
Number of Epochs	150
Batch Size	30
L2 Regularization λ	$6e^{-3}$
Dropout	0.55

All the hyperparameters for the acoustic and textual features extraction networks are presented in Tables 2 and 3, respectively.

For the classification task after the proposed attention models, we use a hidden layer with 96 neurons with dropout.

All the parameters were initialized following the previous work of Hazarika et. al [5] with modifications in the acoustic model, due to differences in input between our works. The hyperparameter optimization was performed by Grid Search, using accuracy as the performance metric.

Table 2. Hyperparameters for acoustic features extraction model.

$[k_1, k_2]$	$[c_1, c_2]$	n_{pool}	n_{GRU}
[25, 5]	[4, 8]	200	32

Table 3. Hyperparameters for textual features extraction model.

emb_size	$[f_1, f_2]$	n_{fms}
150[1], 300[2]	[3, 5]	200

Our dataset, IEMOCAP [4], is composed by two-way conversation videos between 10 speakers (5 male and 5 female) in English. The videos are segmented into speech utterances, transcripted and finally annotated into one of four emotions classes: Anger, Happiness, Neutrality, Excitement and Sadness. We merge the Excitement and Happiness classes, since they are close in activation and valence. We separated 10% of the dataset for testing, resulting in 555 utterances, while training with the remaining 90%, composed by 4976 utterances. As measurement, we will be analyzing the accuracy of our model in the emotion classification task. We also present the attention scores of our attention model in order to inspect the relative importance of acoustic and textual features in the classification.

[1]For trainable embedding matrix.

[2]For pre-trained Word2Vec.

4.2 Results

The results achieved by our model with trainable word embeddings are presented in Table 4. In the case of $\mathbb{R}^n$, we present the average values of all attention scores for all dimensions. We also show some results from Hazarika et al. [5], who extracted features from audio using tools not embedded into the classification model. In their work, *uSA* (Uni-dimensional Self-Attention) is the same as our Att_d^1 model, *mSA* (Multi-dimensional Self-Attention) is Att_{d2}^n and *Audio* and *Text* are the unimodal models using only audio and text, respectively. In their work, they utilized pre-trained word embeddings, instead of training it alongside the classification model.

Table 4. Emotion classification accuracy and attention scores of audio and text features using a trainable embedding matrix. Att_d^1 and Att_s^n outperform other proposed attention methods. Using both features increases the performance and their importance is similar by attention score.

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	—

We also ran Att_d^1 and Att_s^n using Word2Vec pre-trained word embeddings. The results are presented in Table 5.

Table 5. Emotion classification accuracy and attention scores of audio and text features. In these models, we use pre-trained word embeddings. Att_d^1 outperforms Att_s^n and previous proposed models, with a higher impact of text features in the classification task. Att_d^1 achieves close to state of the art results.

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

It is known that textual information is highly important for the understanding of emotion. Many words have an sentimental connotation and can express emotion. Take as example the word "love". Surely, one would not relate such word to a negative emotion, but the opposite. In order to examine our model's ability to associate words to an emotion, we inspected the prediction ratios and mutual information of each word in our dataset.

We propose, as a word analysis, scoring each word with the ratio between how many times that word was in a sentence predicted to a specific emotion and the total number of times it was present in the testing set. For this analysis, we will be using Att_d^1 with pre-trained word embeddings, since it achieved the highest accuracy.

For the Happiness class, the laughter tag, which comprehends when the actors laughed, has a ratio of 0.867 and a big relationship with the classification. The word "excited" also has a high ratio (0.750)

for the Happiness class. As for the Neutrality class, the expressions "um" and "uh", which are speech fillers, are understood by the network as neutral expressions, with ratios of 0.8 and 0.65, respectively. As for the other classes, no word that achieved high ratios for Anger and Sadness reflect the emotion, for example, "she's" (0.800) and "girl" (0.714) for the former and "else" (0.857) for the latter.

In Table 6, we show the words with the highest mutual information feature for each class on the proposed Att_d^1 model with pre-trained embeddings and their ratios, as proposed.

Table 6. Words which have the highest mutual information score on each class and their ratios in predictions. The Happiness and Neutral classes have characteristic parts of speech, while the other classes do not.

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
Neutrality	um	0.800
	can	0.667
	uh	0.650
Sadness	they	0.023
	else	0.857
	the	0.121

5. DISCUSSION

The audio feature extraction model presented in this paper is a basic variation of the ResNet [17]. After analyzing the increase in accuracy after adding residual connections, we tried to increase the network; however, it was not possible, especially due to the small size of the dataset, which caused quick convergence and overfitting.

As demonstrated in other areas of deep learning, such as computer vision, convolutional layers have surpassed manual feature extraction when trained with large sets of data [15; 18]. Thus, we believe that with a larger dataset, we can increase and improve the feature extraction model, allowing it to learn more meaningful features. We suspect that a model similar to what we propose has a higher ceiling for improvement than models that use features extracted outside the model.

Even though we were not able to achieve state of art results, we strongly believe that with more training data, we can exceed them and also apply this acoustic feature extraction model to other audio classification tasks.

By analyzing the mutual information value of the words presented in Table 6, the small size of the dataset can again be seen as a drawback. Due to the lack of training and testing data, many words that normally would not be considered related to an emotion class, such as "business" and "they" have a high mutual information, resulting in a biased classification. This bias is caused by words not related to a single emotion being annotated almost and even completely to a single class. However, it is important to emphasize the model's ability to learn some happiness and neutrality related words and expressions, such as the laughter tag and speech fillers.

Speech fillers are parts of speech that usually do not contain formal meaning and are frequently composed by pauses and repairs. Even though speech fillers are used in the most diverse situations and it can be argued that they do not carry emotional information, we claim that speech fillers can be classified as neutral. Neutrality can be defined as not containing marked characteristics or features. We believe that speech fillers, as neutrality, do not compose or support either emotional spectrum and can be classified as a central and neutral part of speech.

Our proposed ratios for word analysis reflect similar results from the mutual information analysis, especially for the neutral and happy emotions. Word fillers have high mutual information and ratios for neutral speech, as the laughter tag in the Happiness class.

As expected and shown by previous works [5], mixing audio and textual features result in higher classification accuracy. As shown by our model, words and expressions carry emotional content and can be used to more precisely classify speech into emotion. Even though most words can be considered emotionally neutral since they are used under most conditions and environments, the existence of expressions that demonstrate sentiment is an important feature that must be tackled in emotion classification.

Alongside the use of textual information, acoustic features must be used for a better emotion classification. Deep text classification can be easily fooled [19]. As an example of the importance of acoustic features, we will also use the Google Cloud Natural Language API[3] to show that one can effortlessly classify most sentences to a neutral sentiment, since the API does not consider how the speech was portrayed. Take as example the simple sentence: "This chair is good". Google's API attributes a positive score of 0.7/1.0 to the sentence. However, what if this sentence was spoken with an angry voice, disagreeing with a previous idea? Surely, an emotion classification model cannot always classify such sentence to a neutral class. Context and acoustic features play a big role in emotion classification and must be used for better results. Not only what is said, but also how is extremely important for a more complete speech understanding.

Table 7. Confusion matrix for our pre-trained Att_d^1 model. Neutrality class is the worst performer, having the lowest accuracy and the most mispredictions. Sadness has the highest accuracy and Happiness has the highest recall.

	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

As shown by Table 7, we also analyzed that the Neutrality class achieves the lowest accuracy and recall, with values up to 0.1 lower than the other classes. The mispredictions of this class are spread out all the other emotions, while for the other classes, most of these errors usually predict the Neutrality class. We can interpret this result as the neutrality on speech being hard to classify since it does not contain any characteristic words or expressions, as it embodies most of the language used.

Classifying happy speech, as shown by the characteristic expressions learned by our model and the highest recall, is learned well by our model. Even though our model was not able to learn characteristic parts of speech for the Sadness class, our model is able to predict sad speech with high accuracy. This can be explained by the smaller amplitude of the waveforms; sad speech is usually

[3]https://cloud.google.com/natural-language/

portrayed in a smaller volume and without large variances. Table 8 shows the average of the absolute waveforms (*WaveAvg*), the standard deviation of the averages (*WaveAvgStd*) and the average of the standard deviations (*WaveStdAvg*) of the signals for each emotion class.

Table 8. Averages and standard deviations of the raw audio waveforms for each emotion. Sad speech is softer and has smaller variances, while angry speech is louder and highly variable.

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

Black box models are hard to interpret, and many studies focus on the interpretability of such models. With the use of attention, we were able to analyze the attention scores and interpret the importance of audio and textual features in the classification. As the results of Table 4 and Table 5 indicate, the average of the attention scores are quite similar throughout the testing set. When analyzing each testing input separately, the values vary with a standard deviation of approximately 0.09 in the case of Att_d^1 and 0.002 for Att_s^n without pre-trained word embeddings. As for the pre-trained Att_d^1 model, the values have a standard deviation of 0.08. The higher attention score for textual features in this model can be explained by the fact that the word embeddings were trained with a large dataset, allowing them to learn better features. These results confirm the hypothesis that both audio and textual features play a big role in the emotion classification task, and both should be used for better results.

6. CONCLUSION

In this paper, we introduced results of speech emotion classification using transcriptions and raw audio waveforms, instead of the usual use of audio features extracted using outside tools, achieving over 71% accuracy and close to state of the art results.

We also analyzed the importance of acoustic and textual features in the emotion classification task by observing attention scores for both inputs, showing that both are highly important.

Lastly, we examined which words and reactions have a high impact in the emotion classification, especially in the Happiness and Neutrality classes, which contain characteristic words and expressions learned by our model.

7. Future Work

As future work, a new dataset must be developed in order to increase the number of emotion classes and data and, thus, improve the model's ability to effectively classify complex human speech. With the new dataset, we intend to work deeply on the development of the audio feature extraction model, since small changes, even with our present dataset, resulted in increased performance. Another important aspect of speech that we intend to work with is the temporal dimensionality of conversations. Context, speech duration, pauses, among other textual and acoustic features, play a big role in emotional speech.

Also, we intend to use other machine learning techniques alongside neural networks with the objective of improving the interpretability of our model, in order to interpret at a lower level how the classification process works [20].

8. ACKNOWLEDGEMENTS

This research was supported by the Korean MSIT (Ministry of Science and ICT), under the National Program for Excellence in SW (2016-0-00018), supervised by the IITP (Institute for Information & communications Technology Promotion).

9. REFERENCES

[1] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2392–2396, IEEE, 2017.

[2] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[3] J. Lee, T. Kim, J. Park, and J. Nam, "Raw waveform-based audio classification using sample-level cnn architectures," *arXiv preprint arXiv:1712.00866*, 2017.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[5] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 196–201, April 2018.

[6] K. S. Rao and S. G. Koolagudi, "Recognition of emotions from video using acoustic and facial features," *Signal, Image and Video Processing*, vol. 9, no. 5, pp. 1029–1045, 2015.

[7] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing*, vol. 261, pp. 217–230, 2017.

[8] R. Xia and Y. Liu, "Using i-vector space model for emotion recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[9] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio.," in *SSW*, p. 125, 2016.

[10] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4624–4628, IEEE, 2015.

[11] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4295–4299, IEEE, 2015.

[12] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[13] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[18] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, pp. 379–387, 2016.

[19] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," *arXiv preprint arXiv:1704.08006*, 2017.

[20] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv preprint arXiv:1711.09784*, 2017.