Interspeech 2022
18-22 September 2022, Incheon, Korea

Deep CNN-based Inductive Transfer Learning for Sarcasm Detection in Speech

Xiyuan Gao, Shekhar Nayak, Matt Coler

University of Groningen, Campus Fryslân, Leeuwarden, the Netherlands
{xiyuan.gao, s.nayak, m.coler}@rug.nl

Abstract

Sarcasm is a frequently used linguistic device which is expressed in a multitude of ways, both with acoustic cues (including pitch, intonation, intensity, etc.) and visual cues (including facial expression, eye gaze, etc.). While cues used in the expression of sarcasm are well-described in the literature, there is a striking paucity of attempts to perform automatic sarcasm detection in speech. To explore this gap, we elaborate a methodology of implementing Inductive Transfer Learning (ITL) based on pre-trained Deep Convolutional Neural Networks (DCNNs) to detect sarcasm in speech. To those ends, the multimodal dataset MUStARD is used as a target dataset in this study. The two selected pre-trained DCNN models used are Xception and VGGish, which we trained on visual and audio datasets. Results show that VGGish, which is applied as a feature extractor in the experiment, performs better than Xception, which has its convolutional layers and pooling layers retrained. Both models achieve a higher F-score compared to the baseline Support Vector Machines (SVM) model by 7% and 5% in unimodal sarcasm detection in speech.

Keywords: sarcasm detection, Deep Convolutional Neural Networks, Inductive Transfer Learning, speech recognition, human-computer interaction

1. Introduction

In order to reliably recognize sarcasm, it is necessary to define what sarcasm is. A common sense definition suggests that sarcasm is a way to convey the opposite of what is said. This is, however, something of a reduction. Consider how a sentence like “the service here is good”, when delivered sarcastically does not only mean that the service is actually not good, but that it is awful. In other words, it is not merely a matter of negation [1]. Sarcasm can also be heightened with certain words which express illocutionary force. In English, this can be achieved with adverbs like “honestly”, as in “honestly, the service here is good.” Another interesting aspect of sarcasm that evades a typical definition is that it is systematic. This means that any sentence can be rendered sarcastically, with a largely predictable interpretation, even in the absence of context. Thus, the interpretation of the example above would be obvious to all eavesdroppers without recourse to context. It is also pertinent to bear in mind that the ways in which sarcasm is conveyed and perceived have been extensively studied across cultures and languages [2, 3, 4]. In linguistics, it has been argued that prosodic cues (e.g., intonation and stress) play a significant role in communicating sarcasm [4, 5, 6]. Critically, these results have not been fully implemented in speech technology. As speech technology is increasingly used in our daily activities, developing a system that is able to recognize and interpret colloquial speech is a necessary step to further enhance human-computer interaction.

This paper addresses two gaps in sarcasm detection research. The first gap stems from the observation that although sarcasm detection in text [7, 8, 9] has been studied for a while now, sarcasm detection in speech receives comparatively much less scholarly attention. The second gap addresses the fact that of the few attempts to perform sarcasm detection from speech, all rely on machine learning algorithms such as tree-based methods and Support Vector Machines (SVM) [10, 11, 12], even though Neural Networks (NNs) based transfer learning could be a very effective tool to these ends. Concretely, Deep Convolutional Neural Networks (DCNNs) advance in several aspects such as the first GPU implementation and the first application of maximum pooling [13]. They stand out in various research fields recently as a result of their ability to tackle challenging tasks, including speech and signal processing [14]. Moreover, transfer learning, which allows models to be re-used in different tasks, domains, and distributions, is gaining increasing attention in speech technology [15, 16], upon which, Inductive Transfer Learning (ITL) is a category of transfer learning, used to tackle target tasks that are different from source tasks. However, to the best of our knowledge, no existing study has applied DCNNs-based ITL to sarcasm detection in speech. Accordingly, we propose the following research questions:

1. Can DCNNs-based ITL enhance sarcasm detection in speech?
2. What DCNN transfer learning approach optimizes sarcasm detection in speech?

To address the first question, we conduct experiments in which we apply pre-trained DCNNs to sarcasm detection by using ITL. It is hypothesized that the selected pre-trained DCNNs models, Xception and VGGish, can enhance the sarcasm detection performance compared to the baseline SVM model. To address the second question, we implement different transfer learning approaches to the selected models, namely, instance-based transfer learning and feature-representation transfer learning. Then, we compare the model performance on the MUStARD [10] dataset. Work by Tsalera et al. [17] used feature-representation transfer learning successfully achieved 100% accuracy in classifying sounds in the Air Compressor dataset. As an improvement upon the previous research, in this paper, we implement both instance-based transfer learning and the feature-representation transfer learning, with an intention to investigate which transfer learning approach generates better performance for sarcasm detection in speech.

We make the following contributions:

1. We implement ITL based on pre-trained DCNNs to sarcasm detection in speech.
2. We demonstrate that the feature-representation approach enhances DCNNs architecture in the task of transfer learning of detecting sarcasm in speech.

10.21437/Interspeech.2022-11323

3. We demonstrate that data augmentation is a suitable method to deal with data scarcity in sarcasm detection.

This paper is organized as follows: Section 2 is an overview of previous works on sarcasm detection in speech and pre-trained DCNNs models. Section 3 discusses the methodology. Section 4 presents the implementation of the experiments. Section 5 presents the results and Section 6 discusses the results. Finally, Section 7 concludes this paper and outlines directions for future work.

2. Related Work

In speech-related fields, studies dedicated to sarcasm detection can be classified into two categories according to the data modality: those that rely on audio data exclusively and those based upon multimodal data (which may include some combination of text, audio, or visual data). As for the former category, the mainstream is to use machine learning methods, in which feature selection is the most important task. For example, Tepperman et al. [11] conducted experiments involving prosodic, spectral, and contextual cues to investigate the role of prosody in detecting sarcasm in speech. Although the results highlighted the insufficiency of prosody in sarcasm detection, spectral cues were shown to be substantial features. Later, Rakov & Rosenberg [12] extracted sentence-level acoustic features (including pitch, intensity, and speaking rate) and word-level features (including pitch contour and intensity contour) from a dataset composing utterances from a TV show. Then, they conducted k-means clustering to recognize sarcastic utterances. As a result, it was demonstrated that specific acoustic features (including pitch and intensity contours) were indicators of sarcastic speech. As for the latter category, Castro et al. [10] were the first to propose a multimodal approach to sarcasm detection and a benchmark dataset MUStARD was introduced with audio, visual and text modalities. They used SVM to classify audio files from MUStARD dataset into sarcastic and non-sarcastic utterances. Finally, they achieved an F-score of 65% in the audio modality. Based on the MUStARD dataset, a number of studies concerning multimodal sarcasm detection have been presented [18, 19].

DCNNs are developing rapidly as their excellent performance in image classification and recognition. Xception [20] is a DCNN with depthwise separable convolutions. The Xception architecture has 36 convolutional layers serving as feature extractors, followed by a logistic regression layer. All 36 convolutional layers are constructed into 14 modules, each module is connected with linear residual connections. Xception is trained on the JFT dataset which composes of over 350 million images that are classified into 17,000 classes. It has been profoundly used in image identification and classification. VGGish [21] is a sound-based DCNN model modified based on the architecture of VGG which is used for image classification [22]. VGGish contains a sequence of convolution and activation layers, followed by max pooling layers; fully-connected layers at the top generate output feature vectors with a size of 128. VGGish is trained on YouTube-100, which is a dataset consisting of 100 million YouTube videos. Each video is labeled with at least 1 topic (e.g., “Song”, “Cormorant”, “Trumpet”). It is built to classify soundtracks from 70 million videos (5.24 hours) labeled with 30,871 categories. Table 1 presents details of the two models.

Table 1: *Selected DCNNs*

CNNs	Type	Dataset	Layers	Parameters
Xception	Image	JFT	71	21 millions
VGGish	Audio	YouTube-100	24	7 millions

3. Methodology

To address the research questions proposed in Section 1, we elaborate the following replicable approach. Two pre-trained models were selected for the experiment: Xception, which is an image-based model, and VGGish, which is a sound-based model. In the experiments, initially, we preprocessed 690 audio files from MUStARD to reduce background noise and to match the input size of the selected models. Next, features were extracted from the audio files and fed into the models. Then, for both models, data augmentation was conducted to boost the training set. Regarding the transfer learning approaches, the instance-based approach was used for Xception. We adopted VGGish as a feature extractor. Lastly, in the evaluation stage, 5-fold cross-validation was used to evaluate the models. The overview of the process is presented in Figure 1.

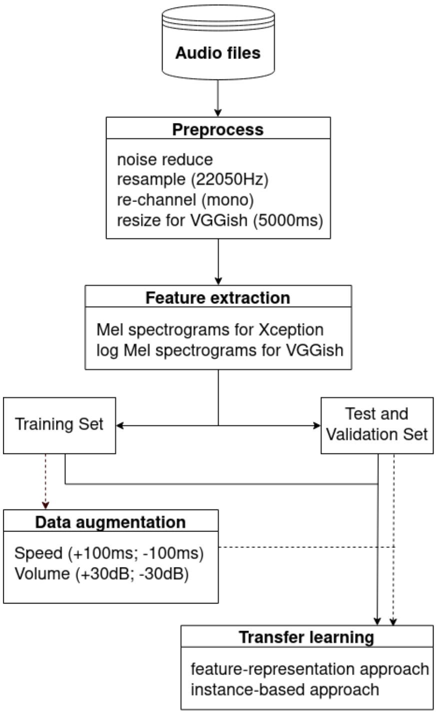

Figure 1: *Overview of audio file processing.*

3.1. Pre-trained models

We implemented instance-based transfer learning and feature-representation transfer learning [23] for Xception and VGGish respectively. In the aspect of Xception, since the similarity between both tasks and domains of the source and the target are low, we considered that the discrepancy may be diminished by updating the weights with the target data. Therefore, we modified the top layer of the original model; layers from 100 onwards were retrained with the target data. The model architecture is presented in Figure 2. As for VGGish, the source and target tasks are different; however, both the source and target domains are related to sound processing. Only the top layer was modified. Figure 3 shows the architecture of VGGish used in this study.

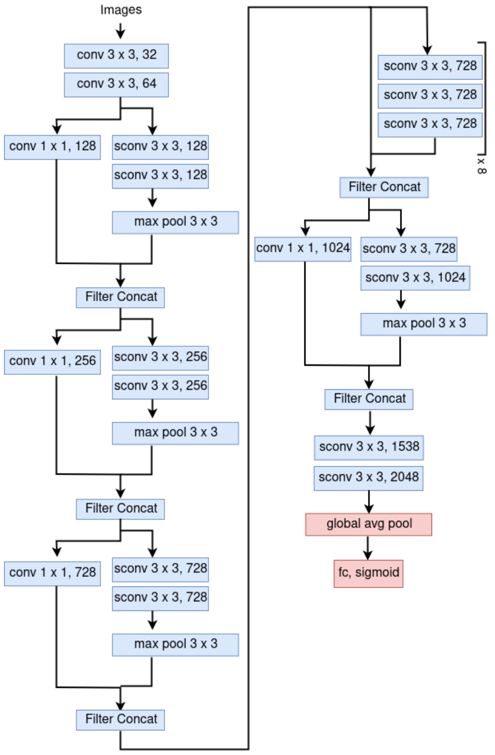

Figure 2: *The Xception architecture. "sconv" is separable convolutional layer. The top layer is changed to a global average pooling followed by a fully connected layer with sigmoid as the activation function. Layers from 100 onward are retrained with the target data.*

3.2. Dataset

We used MUStARD dataset which is a multimodal corpus consisting of 690 audiovisual utterances extracted from TV shows (Friends, The Golden Girls, The Big Bang Theory, etc.). Each utterance is accompanied by a label. In total, 345 sentences are labeled as sarcasm and the other 345 sentences are annotated as non-sarcasm. As we aimed at sarcasm in speech particularly, only the audio files from the dataset were extracted and used in the experiments.

3.2.1. *Audio preprocessing*

All audio files were preprocessed to reduce background noise. We re-sampled each audio file to 22,050 Hz; moreover, all of them were converted to mono channel before extracting features.

3.2.2. *Feature extraction*

Considering Mel spectrograms are spectrograms that reflect the human perception of sounds, and represent both temporal and frequency domain information of sounds, in our experiments, audio files were represented by Mel spectrograms to maximally represent sarcastic patterns in speech. As Xception only accepts images as inputs, pertinent packages were used to transform audio files into images. This included the sound processing package librosa [24] and plot package matplotlib [25]. We used 64 Mel bins, the length of Fast Fourier Transform window was set to be 1024. With respect to VGGish, we used the method provided by the developers of VGGish on Github [1] to generate features with the size 96 ×64.

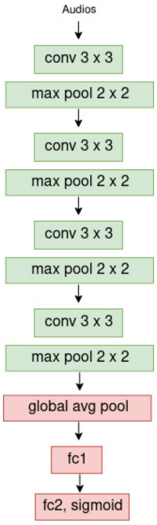

Figure 3: *The VGGish architecture. The top layer is changed to a global average pooling followed by two fully connected layers.*

3.2.3. *Data augmentation*

The MUStARD dataset contains only 690 audio files, which is too small for training DCNNs. Therefore we augmented the data as follows: audio files were accelerated by 100ms and slowed down by 100ms. Additionally, an increase of volume by 30 dB and a decrease by 30 dB collectively contributed to generating more data.

4. Experiments

Both selected pre-trained models were implemented with Keras [26] and used Tensorflow [27] as the backend. Adaptive Moment Estimation (Adam) was chosen as the optimizer due to its benefits of adapting the learning rate [17]. The Binary Crossentropy was employed to calculate the score of probabilities. As the dataset was small while the number of parameters was huge in the pre-trained models, we set small learning rates to compile the models. For Xception, the top layer was assigned learning rates of 0.001 and 0.01, and the fine-tuning layers (e.g., from 100 layers onward) were retrained with a learning rate of 1e-5. The same batch size (e.g., 8, 16, 32) and epochs (e.g., 6, 8, 10) were appointed during the training for both models. Tabel 2, 3, and 4 show the hyperparameters in detail.

5. Results

Compared with the baseline SVM, both of the proposed pre-trained models gained higher performance. Without data aug-

[1] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

Table 2: *Hyperparameters for Xception's top layer*

Optimizer	Batch size	Epoch	Learning rate
Adam	16,32,64	6,8,10	0.001,0.01

Table 3: *Hyperparameters for fine-tuning Xception*

Optimizer	Batch size	Epoch	Learning rate
Adam	16,32,64	12,16,20	1e-5

mentation, Xception achieves an F-score of 69%, whereas the number is 70% for VGGish. With the augmented data, the F-score grows from 69% to 70% for Xception. Similarly, VGGish experiences an increase by 2%.

The highest performance is gained by VGGish when the learning rate is 0.01, trained with a batch size of 32 and 8 epochs (68% Precision, 77% Recall, 72% F-score). The combination of hyperparameters generates the highest score for Xception is slightly different, which is when the batch size is 16 and epoch is 10, with a learning rate of 0.001 (66% Precision, 76% Recall, 70% F-score). The highest results from training the two models are listed in Table 5.

6. Discussion

We investigated whether DCNNs-based ITL can enhance sarcasm detection in speech by conducting experiments on selected pre-trained models. We then explored the better transfer learning approach by comparing the performance of the two models on the audiovisual sarcasm dataset MUStARD.

The results show that selected DCNNs models, Xception and VGGish improved sarcasm detection by 5% and 7% respectively when compared to the SVM baseline model. Our results are consistent with the results provided by Tsalera et al. [17], in which the performance of the image-based pre-trained models was inferior to sound-based pre-trained models. There may be two causes of this disparity between models. An initial explanation involves the dissimilarity between the source and target data. In terms of the source data, Xception is trained on images while VGGish is trained on audio files. Even though the data domain between VGGish and the current MUStARD dataset is different, the data similarity is still higher than that between Xception and the current dataset. Another explanation involves the different methods of transfer learning. Here, two methods of retraining were applied (See Section 3). VGGish was applied as a feature extractor (i.e., feature-representation approach); however, in the case of Xception, not only the top layer but also the convolutional and pooling layers from 100 onward were retrained (i.e., instance-based approach). The results support that the feature-representation approach works better with the sound-based pre-trained model in the target task.

The results also reveal that data augmentation is an effective tool to deal with data scarcity. In this study, data is augmented by adjusting speed and volume. Xception gains an increase of 1% and VGGish is enhanced by 2%.

7. Conclusions

This paper investigates whether DCNNs-based ITL can enhance sarcasm detection in speech and further explores which transfer learning approach is suitable for the target task. To address the research questions, we applied two pre-trained DCNNs, which are trained on image and audio datasets respectively, with two different transfer learning approaches and further tested them on the benchmark multimodal sarcasm dataset MUStARD. Results indicate that DCNNs-based ITL can enhance sarcasm detection in speech. VGGish gains better performance than Xception. Our experiments extend ITL from image-based and sound-based DCNNs to sarcasm detection in speech, demonstrating that feature-representation transfer learning approach is compatible with DCNNs architecture in sarcasm detection, and further testifying that data augmentation is a suitable method to deal with data scarcity.

Table 4: *Hyperparameters for VGGish*

Optimizer	Batch size	Epoch	Learning rate
Adam	16,32,64	6,8,10	0.001,0.01

Table 5: *Compared to the baseline SVM, the highest results gained before and after data augmentation in Xception and VGGish. Note: DA=data augmentation*

Models	DA	Precision	Recall	F-score
Xception	No	65%	73%	69%
Xception	Yes	66%	76%	70%
VGGish	No	67%	76%	70%
VGGish	Yes	68%	77%	72%
SVM(baseline)	No	66%	65%	65%

There are multiple avenues for future research. Below are five. First, more data could generate higher performance as it adds up the trainable patterns for models. Alternatively, various data argumentation methods (e.g., pitch change, time-shifting, noise injection, etc.) can be applied for further improvement. Second, a diversity of transfer approaches (e.g., parameter-transfer and relational-knowledge-transfer) could be involved to further investigate the optimal transfer learning approach. As the next step, we can apply the instance-based-transfer to VGGish to further improve our study. Third, future studies can focus on extracting fine-grained features from audio files. In this study, the Mel spectrogram is extracted from each audio file to represent acoustic features. However, sarcasm-related features (e.g., pitch, intensity, speaking rate, etc.) are not fully represented by Mel spectrograms. Selected features associated with sarcasm could improve the performance further. Fourth, fine-tuning pre-trained models and hyperparameters adjustment could potentially improve the model performance. Fifth, as previous research indicates that multimodal application increases the accuracy of sarcasm detection [10], it may be of great value to combine textual and audiovisual modalities in the future.

In sum, this paper enhances sarcasm detection in speech by elaborating a methodology of implementing DCNNs-based ITL to detect sarcasm in speech, the two selected pre-trained models (Xception and VGGish) achieved an increase of 5% and 7% in the benchmark sarcasm dataset MUStARD.

8. Acknowledgement

We would like to express our gratitude to Dr. Henk van den Heuvel (Director, Centre for Language and Speech Technology, Radboud University Nijmegen) for his helpful input. Any errors and omissions are our own.

9. References

[1] R. J. Fogelin, *Figuratively speaking*. Yale university press, 1988.

[2] S. Peters, K. Wilson, T. W. Boiteau, C. Gelormini-Lezama, and A. Almor, "Do you hear it now? a native advantage for sarcasm processing," *Bilingualism: Language and Cognition*, vol. 19, no. 2, pp. 400–414, 2016.

[3] C. Techentin, D. R. Cann, M. Lupton, and D. Phung, "Sarcasm detection in native english and english as a second language speakers," *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 75, no. 2, pp. 133–138, 2021.

[4] H. S. Cheang and M. D. Pell, "Acoustic markers of sarcasm in cantonese and english," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1394–1405, 2009.

[5] P. Rockwell, "Lower, slower, louder: Vocal cues of sarcasm," *Journal of Psycholinguistic research*, vol. 29, no. 5, pp. 483–495, 2000.

[6] G. A. Bryant and J. E. Fox Tree, "Is there an ironic tone of voice?" *Language and speech*, vol. 48, no. 3, pp. 257–277, 2005.

[7] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-based sarcasm sentiment recognition in twitter data," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2015, pp. 1373–1380.

[8] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 704–714.

[9] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PloS one*, vol. 13, no. 10, 2018.

[10] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an obviously perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2019.

[11] J. Tepperman, D. Traum, and S. Narayanan, "Yeah right: Sarcasm recognition for spoken dialogue systems," in *Ninth international conference on spoken language processing*, 2006, pp. 17–21.

[12] R. Rakov and A. Rosenberg, "sure, i did the right thing: a system for sarcasm detection in speech," in *Interspeech*, 2013, pp. 842–846.

[13] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[14] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," 2016. [Online]. Available: https://arxiv.org/abs/1606.06565

[15] S. Ntalampiras, "Bird species identification via transfer learning from music genres," *Ecological Informatics*, vol. 44, p. 76–81, 2018.

[16] A. Triantafyllopoulos and B. W. Schuller, "The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7268–7272.

[17] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of pretrained cnns for audio classification using transfer learning," *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, p. 72, 2021.

[18] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, and J. Li, "Modeling incongruity between modalities for multimodal sarcasm detection," *IEEE MultiMedia*, vol. 28, no. 2, pp. 86–95, 2021.

[19] X. Zhang, Y. Chen, and G. Li, "Multi-modal sarcasm detection based on contrastive attention mechanism," 2021. [Online]. Available: https://arxiv.org/abs/2109.15153

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[21] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[22] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[23] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[24] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[25] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[26] F. Chollet *et al.* (2015) Keras. [Online]. Available: https://github.com/fchollet/keras

[27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/