CS230

Speaker Identification with Deep Neural Networks

Alfredo Méndez
amendezp@stanford.edu

Abstract

There is a great research effort in looking for medical solutions for Alzheimer's disease, while significantly less in creating solutions for care-giving post diagnosis. The motivation of this project is to provide patients with a tool to recognize familiar people in common situations. The solution implements a text independent speaker recognition system using deep learning.

This paper compares the performance of a convolutional neural network (CNN) against a fully connected neural network to address the speaker identification problem. The CNN includes 1-dimension convolutional layers, max pooling layers, batch normalization, regularization techniques and a SoftMax output. The models are trained and tested in the freely available VCTK Corpus Data Set for 109 speakers. Our results show that the CNN surpasses the fully connected network with an accuracy of 93.05% compared to 75.88%.

1 Introduction

Being able to identify a person's identity is critical to remember an existing relationship with an individual and achieve a successful conversational dynamic. Alzheimer's patients constantly face the challenge of not being able to recall a person's identity through their physical traits. Some of the existing tools for assisting such patients in the recognition of someone's identity is by their looks, which require them to take pictures, thus generating an awkward and sometimes, emotionally challenging situation.

This project aims to build a deep learning-powered speaker recognizer. The tool could be used to provide useful information about a speaker's identity to an Alzheimer's disease patient. It is a relevant problem from two perspectives: the intended end use of the project and deep learning techniques employed to achieve the desired objective. First of all, the project aims to assist a group that, historically, has not received enough attention: Alzheimer's disease patients. A scaled product that implements this solution could help patients engage and follow group conversations such as the ones in family reunions or phone calls. Secondly, the project aims to leverage progress achieved in the Deep Learning field for the speaker recognition problem in order to perform correctly and effectively assist users.

The input of the algorithm used in the project are recorded utterances spoken by single individuals. It is important to highlight that this is a text-independent model, which means that works regardless of the used words in speech. The utterances must first be pre-processed in order to extract the MFCC features of each recorded speech. Furthermore, the MFCC features vector is used as input for the neural network models that generate predictions.

The neural network model outputs a vector with the probabilities of the utterance belonging to each of the registered speakers. This vector is used to generate a prediction of the speaker's identity, guessing for the one with the highest probability.

CS230: Deep Learning, Winter 2019, Stanford University, CA.

2 Related work

Voice recognition is a well-researched problem. Nevertheless, it is important to differentiate between different tasks within voice recognition problem space, such as speech recognition, speaker validation and speaker recognition. Speech Recognition is the task of recognizing words in speech for them to be interpreted by a machine or converted into written text. Speaker Verification aims to validate if spoken utterances come from of a specific person according to the characteristics of their voice (single individual, binary outputs) Lastly, Speaker Recognition refers to the task of identifying which speaker from a fixed set of speakers recorded has just spoken (Multiple possible outputs). It's also important to notice differences between Text Dependent and Independent Methodologies. Text-Dependent tasks always use the same spoken words, while Text-Independent use any set of spoken words, focusing more in overall frequencies characteristic of voice features.

Regarding text-independent speaker identification, several approaches and research lines have been explored and developed. Since the early 70s, there has been notable efforts in developing speaker recognition systems, which has not been an easy task. even for humans sometimes it represents a big challenge to recognize close friend's voice. One of the main factors that makes speaker and speech recognition a challenge is the need to, first of all, build an algorithm that converts recorded audio into machine-readable features.

Among the early efforts to develop the state of the art, some of the relevant work in the field was accomplished by Fant (1973) who considered speech as a sequence of phonetic commands to model speech in a temporal dimension, Magrin-Chagnolleau et al. (1996) who proposed an AR vector modeling for speech spectral evolution and Campbell (1997) who described a tutorial of automatic speaker recognition systems.

More recently, in the XXI century, there has been some relevant work in developing speaker recognition, primarily through clustering algorithms. Among these clustering approaches we can find the usage of Gaussian Mixture Models (Kaminski et al. (2013)), Principal Component Analyses (Nie and Zeng (2004)), Support Vector Machines (Louradour and Daoudi (2008)) and Neural Networks (Ahmad et al. (2015)), among others.

An interesting approach to the speaker recognition task is the described by Li et al. (2017) developed by Baidu Inc. in which a neural network based embedding system is proposed to extract acoustic features and produce utterance level speaker embeddings to be tested for speaker recognition with a residual neural network. It is similar to my work in the aspect that our embeddings generator was created based on an implementation that aims to replicate Deep Speaker's performance, developed by Philippe Remy.

Another interesting approach is the one made by Lukic et al. (2016) or the one made by Torfi et al. (2018) in which convolutional neural networks are used to solve the Speaker identification task. The strength of this approach is that convolutions work well at reducing the high dimensionality needed to work with all of the extracted speech features present in a utterance. Inspiration drawn from this papers regarding the usage of convolutional layers for an efficient architecture of a Neural network was used in the development of my project.

3 Dataset and Features

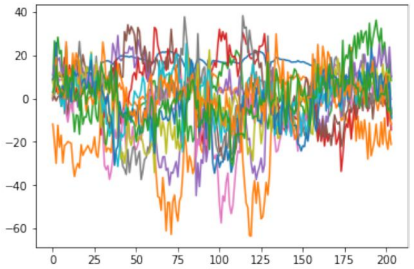

Figure 1: Example of the MFCC embedded frequencies of a 4 second speech audio of the VCTK corpus

The employed dataset is the VCTK Corpus collected by the Centre for Speech Technology Research. Such dataset proved to be adequate for the desired task since it is composed by speech data uttered by 109 english speakers with a diverse set of accents. Additionally, the data set proves to be of an acceptable size and variety, since every speaker reads around 400 sentences that were specifically selected to maximize contextual and phonetic coverage in order to provide the trained model with as much information as possible. Finally, each speaker reads a different set of sentences, which is also adequate to the task of building a text-independent speaker recognizer, which identifies speakers regardless of the spoken words.

It's important to highlight that all speech data in the VCTK Corpus was recorded using the same recording set up, with high performance microphones and no additional noise in the background. This is not optimal for the desired end solution, because in a daily basis scenario Alzheimer's patients need to identify speakers in an environment where the recordings are not of high quality and would include background noise. One possible solution to this issue, solving the train-test data distribution gap, is to implement data augmentation techniques to add background noise to the training set. This issue will be back logged into future work, therefore the current priority is to prove successful results with the provided data.

4 Methods

4.1 Audio reader and embedding generator

To work and manage audio information, it is necessary to, first, implement an algorithm that reads the audio frequencies in order to extract sound features that are machine - readable. The audio reader algorithm performs this task, taking as an input the recorded audio .wav files and providing as an output a 390 x (time frames in the audio) vector with the MFCC preprocessed frequencies as an embedding for each utterance. To successfully perform this task, the audio reader first generates a cache where it stores speech features for each audio file that will later be used to generate the embeddings, this is useful in order to speed up data processing.

4.2 Dense Neural Network model

A two layers densely connected neural network (DNN) was selected as a baseline model. This model is used as a benchmark for the VCTK corpus in the speaker identification problem. Such model uses two fully connected layers with sigmoid activation functions, a normalizing layer and a softmax output layer to provide the probability of the utterance belonging to a particular speaker. This model has 99,305 parameters out of which 21,105 are trainable.

Since it's a model for speaker identification, the loss function has to optimize towards the goal of predicting the correct speaker given a single utterance. We want the model to be penalized when guessing for the wrong category, therefore, we employ categorical cross entropy loss. Since we are using one-hot vectors as labels (only one speaker per embedding), only the positive class Cp is greater than zero in the loss. There is only one element of the Target vector t which is not zero $t_i = t_p$. For this reason, the elements which have a value of zero are discarded from the summation, as a result, we end up with a simplified categorical cross entropy loss function shown in Figure 3.

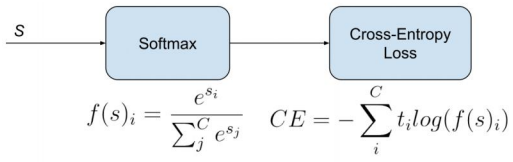

Figure 2: Mathematical expression for the Softmax activation function and the Categorical cross entropy loss function for C classes.

4.3 Convolutional Neural Network model

$$CE = -log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j}}\right)$$

Figure 3: Simplified Categorical cross entropy loss function when using one-hot vectors as labels, where S_p is the score for the positive class

To compare against the baseline model (DNN), a 1D convolutional neural network model (CNN) was implemented. The model consists of 3 convolutional layers, each one of them with its respective max pooling operation and batch normalization process. Afterwards, two fully connected layers each with a dropout mask with a dropout rate of 0.25 to reduce variance. Finally, an output softmax layer to predict the utterance associated speaker . The output layer has a softmax activation function, while the rest of the layers implement a ReLu activation function. The idea behind using a ReLu activation function instead of sigmoid is to avoid gradients close to zero that can come along when using certain values of the sigmoid function. The model has 676,613 parameters out of which 675,969 are trainable.

Just as the DNN model, the CNN model aims to solve the speaker identification task, therefore the same categorical cross entropy loss function is used. The rationale behind using a convolutional neural net is related to the problem of dealing with high dimensionalities that come along with extracting multiple speaker features.

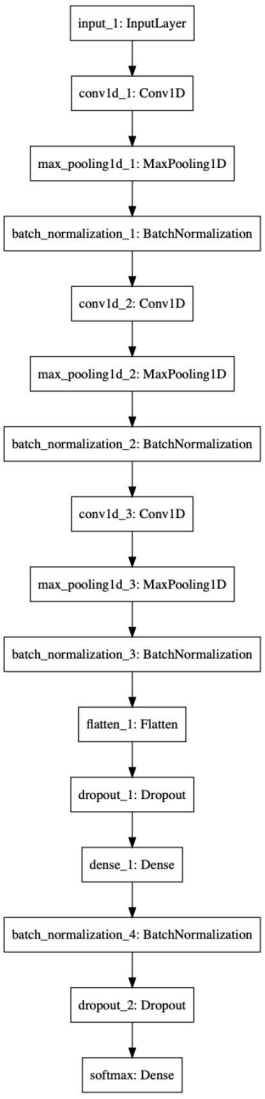

Figure 4: CNN model architecture

5 Experiments/Results/Discussion

For the CNN the mini-batch size chosen was 512. Drawing insight from previous papers and implementations made with this same data set, good results were achieved with a batch size of 900, however this batch size was costly to learn over with the available resources, so it was reduced to the closest 2^n multiple. For the learning rate, a value of 0.001 was used, which is reduced to half if the model's accuracy does not increase after 10 epochs. The reason to use such values us that several papers using similar structures for this task such as Li et al. (2017) share that common learning rate.

To evaluate both models, the performance metric chosen was accuracy since the end goal is to build a speaker identification model that accurately guesses the right speaker most of the time, there is no special scenario where a mistake should be highly penalized.

The DNN model achieved an accuracy of 78.05% with the training set and 75.88% with the validation set. This model appears to have a bias problem due to the relatively low accuracy attained, yet, there is no sign of overfitting or variance issues, as we can see in Figure 5.

Evaluating the performance of the CNN model, we assume that there might be an overfitting problem, as there is clear variance and a significant discrepancy between the performance with the test set and the train set. Analyzing Figure 6, it is relevant to highlight the apparently "random" variances in the test set accuracy and the fact that this, sometimes, exceeds the training set accuracy, this suggest that further training might be needed. The model appears to be heading towards convergence, but it is not clear in the first 60 epochs elapsed. This high variance in the test set could be expected due to stochastic behavior that comes along with mini batch training. If the model manages to converge trained longer with additional computational resources, while keeping high accuracy, then the deep CNN would prove to be better than the baseline DNN.

Model	Train Accuracy	Test Accuracy	Train Loss	Test Loss
DNN	78.05%	75.88%	0.6951	0.7421
CDD	90.84%	93.08%	0.2956	0.2420

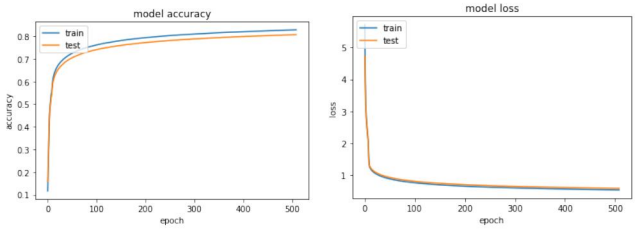

Figure 5: DNN accuracy and loss over 500 epochs

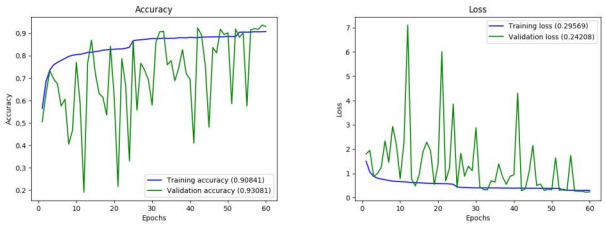

Figure 6: CNN accuracy and loss over 60 epochs

6 Conclusion/Future Work

With this project, it is clear that neural networks are effective at the task of performing speaker recognition. However, the achieved results are not sufficient to guarantee that a "complex" convolutional neural network that implements regularization and batch normalization techniques is necessarily better than a more simple densely connected neural network. The results obtained suggest that the CNN achieves higher accuracy on the training set, nonetheless it presents a high variance which reflects in a lower accuracy for the test set. This indicates that this model has an overfitting problem and further tuning is needed to achieve optimal performance.

A possible explanation for the superior robustness of the DNN over the CNN might be the audio preprocessing pipeline. The inputs are reshaped to be a vertical vector for each time frame, which might not be the optimal way to feed data into a CNN which, thanks to the properties of a mathematical convolution, can identify and extract features information on certain parts of the input.

For future work, it would be interesting to try different embedding systems to see how they affect the performance of the model. Additionally, it would also be interesting to compare the performance to Sequential Neural Networks, which are commonly used in speech recognition. Lastly, prior to building a product based on the aforementioned models in order to assist Alzheimer's patients, it would be critical to perform data augmentation in order to add common background noise before training the data. This way one can guarantee consistency in the distributions of the training and test sets.

The most relevant future work, taking into consideration the initial motivation for this project, would be to work on a product road map to ship an Alzheimer's disease post diagnosis assistant. The main functionality of the solution would be identifying someone by their voice features through deep learning. It is of vital importance to take into consideration other aspects of building a machine learning project, such as the user interface, the likelihood of the distribution of the data to change over time and other potential problem-specific aspects. A good next step for understanding the product specific adoption challenges would be to build a user-friendly interface for the speaker recognizer and test it with real patients.

Project Code

References

Ahmad, K. S., Thosar, A. S., Nirmal, J. H., and Pande, V. S. (2015). A unique approach in text independent speaker recognition using mfcc feature sets and probabilistic neural network. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, pages 1–6. IEEE.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462.

Fant, G. (1973). Speech sounds and features. *The MIT Press*.

Kaminski, K., Majda, E., and Dobrowolski, A. P. (2013). Automatic speaker recognition using a unique personal feature vector and gaussian mixture models. In *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 220–225. IEEE.

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., and Zhu, Z. (2017). Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*.

Louradour, J. and Daoudi, K. (2008). State-of-the-art sequence kernels for svm speaker verification. In *2008 IEEE Workshop on Machine Learning for Signal Processing*, pages 498–503. IEEE.

Lukic, Y., Vogt, C., Dürr, O., and Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE.

Magrin-Chagnolleau, I., Wilke, J., and Bimbot, F. (1996). A further investigation on ar-vector models for text-independent speaker identification. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 101–104. IEEE.

Nie, K. and Zeng, F.-G. (2004). Using neural network and principal component analysis to study vowel recognition with temporal envelope cues. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pages 4592–4595. IEEE.

Torfi, A., Dawson, J., and Nasrabadi, N. M. (2018). Text-independent speaker verification using 3d convolutional neural networks. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.