בס"ד

michael lemberger
sela goldenberg

final report

Abstract

The need to find a solution to the voice recognition problem, made it a popular topic among the deep learning community. many companies started using voice recognition related technologies like the popular google home, amazon Alexa and Apple Siri. voice recognition has many sub-problems, one of them being gender recognition. The costumer gender is a valuable information most companies will like to attain in order to provide better customer service and better information for Advertisers. This article will suggest one solution to the problem using a deep learning approach; Long Short Term Memory (LSTM) and Recursive Neural Network (RNN).

Introduction

Long Short-Term Memory (LSTM) is widely used in speech recognition. In order to achieve higher prediction accuracy, machine learning scientists have built increasingly larger models. Such large model is both computation intensive and memory intensive. This project Show the use of RNN model with LSTM to classify between male and female voices giving statistic data on the acoustic properties of the voice and speech. The dataset was attained from Kaggle and was tested on machine learning models (Random Forest, SVM, Logistic Regression and more) but not deep learning models. The data contain 3168 voice recording with labels of male or female and was pre-processed in to our feature using acoustic analysis in R and using the seewave and tuneR packages and with analyzed frequency range of 0hz-280hz (which is the human vocal range). in the article we also show The result of logistic regression and multilayer perceptron models on the same dataset and we show That the RNN gave more accurate result and manage to get there much faster than the previous models.

בס"ד

michael lemberger
sela goldenberg

Related work

Context-Sensitive Multimodal Emotion Recognition from Speech[1]:

"A multimodal emotion recognition framework that merges audio-visual information at the feature level and uses a classification technique that allows for the modeling of long-range temporal dependencies."

The model used in this article account for contextual information by applying long short term memory (LSTM).
In this article they show that using LSTM had better results than a regular recurrent neural network (RNN).
The article try to consider facial expressions as well, so they used the bidirectional Long Short Term Memory (BLSTM).

Vowel-Based Voice Activity Detection with LSTM Recurrent Neural Network[2]:

The model suggested in this article is defined "Voice Activity detection (VAD)" that can distinguish between noises and human speech.
The model consider the structure of the noise according to the laws of vowels, and determine if that noise was in fact a human voice.
The reason behind using LSTM and RNN in this project:
"LSTM-RNN is known to the powerful model to capture dynamical context information through time. Moreover, with teaching the LSTM-RNN to only vowel sounds rather than whole speech, LSTM-RNN can learn more effectively because of the reduced manifold of speech."

Dataset & features

The complete dataset can be downloaded from kaggle.com[3].

In order to analyze gender by voice and speech, a training database was required. A database was built using thousands of samples in .wav format of male and female voices, each labeled by their gender of male or female. Voice samples were collected from the following resources:

- **The Harvard-Haskins Database of Regularly-Timed Speech**
- **Telecommunications & Signal Processing Laboratory (TSP) Speech Database at McGill University**
- **VoxForge Speech Corpus**
- **Festvox CMU_ARCTIC Speech Database at Carnegie Mellon University**

[1] https://mediatum.ub.tum.de/doc/1082493/file.pdf
[2] https://dl.acm.org/doi/10.1145/3015166.3015207
[3] https://www.kaggle.com/primaryobjects/voicegender

בס"ד

michael lemberger
sela goldenberg

The pre-processed WAV files were saved into a CSV file, containing 3168 rows and 21 columns (20 columns for each feature and one label column for the classification of male or female).

The features are acoustic properties that where measured in the processing of the wav files.
Some of the acoustic properties are:

Normalized frequency is a unit of measurement of frequency equivalent to cycles/sample. (Wikipedia):

- **meanfreq**: mean frequency (in kHz)
- **sd**: standard deviation of frequency
- **median**: median frequency (in kHz)
- **meanfun**: average of fundamental frequency measured across acoustic signal
- **minfun**: minimum fundamental frequency measured across acoustic signal
- **maxfun**: maximum fundamental frequency measured across acoustic signal
- **meandom**: average of dominant frequency measured across acoustic signal
- **mindom**: minimum of dominant frequency measured across acoustic signal
- **maxdom**: maximum of dominant frequency measured across acoustic signal
- **dfrange**: range of dominant frequency measured across acoustic signal

Features peakf (peak frequency) and duration have no effect on the data, so it's safe to assume that the models described in this article didn't consider them in the training process.

previous attempt

For The First Attempt for The Gender classification we tried to use Logistic regression. for the logistic regression The best result came with a learning rate of 0.00002 and 200,00 epochs. with these parameters and with the use of The Default Gradient Descent Optimizer of TensorFlow The model achieved an accuracy of 94.53%. In order to try to improve the model for the second attempt two new models of multilayer perceptron were Tested. The first one with 2 hidden layers and the second one with only 1 hidden layer. Surprisingly the model with only 1 hidden layer achieved better result than the model with 2 hidden layer with very similar parameter. The 2-layer model got pretty good result and were able to achieved 93% accuracy after only 70,000 epochs. but attempt to change number of epochs and the learning rate drooped the accuracy to the ~70%. The 1-layer model managed to get to an accuracy of 81% in only 20,000 epochs (and a learning rate of 0.00001) and with some tinkering with the parameters was able to get to 95% after 100,000.

בס"ד

michael lemberger
sela goldenberg

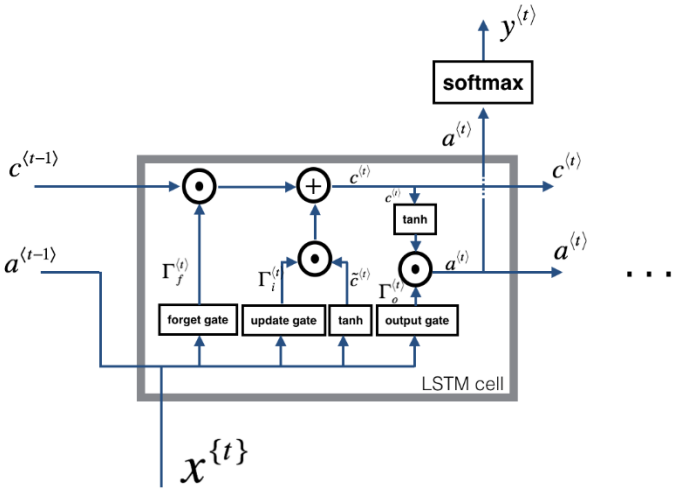
Project description

The main idea of this project is to try and distinguish between male and female voices, as a tool that can be used by interested parties, like advertising agencies that use web posts, search engines and social networks to promote their product. Those advertising agencies would want to know that there is a match between the client's interests and the offers that he is exposed to. In order to do so, companies like Facebook and Google collect user data from their platforms, and sometimes devices, and analyze it accordingly.
The Data can be various kind of formats. One of them is the voice of the costumer using its phone device.
The main goal of the project is to offer a model that can distinguish between the genders by voice that can be used as a tool in the process of data analysis.

When approaching the voice recognition problem and particularly the gender recognition, defining the problem as a classification problem of sequences is an important step towards the solution.

The Recurrent Neural network (RNN) using Long Short Term Memories is the most popular approach when dealing with those kind of problems.
The reason for it is that it is so popular is because it is an easy and sufficient model for sequential problems and classification problems that can output great results.

Voice recognition model make predictions based on time series data, because the voice samples and their features are a representation of important events in the frequencies throughout the time of speaking. And the gender recognition is a matter of classification to male or female. Using LSTM cells with a forget gate is crucial in order that the different frequencies would make sense. The model compare past inquiries by their output.



$$\Gamma_f^{\langle t \rangle} = \sigma(W_f[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_f)$$
$$\Gamma_u^{\langle t \rangle} = \sigma(W_u[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_u)$$
$$\tilde{c}^{\langle t \rangle} = \tanh(W_C[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_C)$$
$$c^{\langle t \rangle} = \Gamma_f^{\langle t \rangle} \circ c^{\langle t-1 \rangle} + \Gamma_u^{\langle t \rangle} \circ \tilde{c}^{\langle t \rangle}$$
$$\Gamma_o^{\langle t \rangle} = \sigma(W_o[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_o)$$
$$a^{\langle t \rangle} = \Gamma_o^{\langle t \rangle} \circ \tanh(c^{\langle t \rangle})$$

LSTM is basically considered to avoid the problem of vanishing gradient in RNN.
Theoretically, the information in RNN is supposed to follow for arbitrary large sequence but in practice this doesn't hold up.

In this project the LSTM accuracy was most accurate when using the "sparse categorical cross-entropy" loss function, "Softmax" and "relu" activation functions.
The prediction accuracy was around 98-99 percent.

michael lemberger
sela goldenberg

In an afterthought "Binary Cross-Entropy" might had achieved the same results as "sparse categorical cross-entropy" because the number of classes in the dataset is binary. In spite of that, the data was better suited to the "sparse categorical cross-entropy" loss function because of technical issues. "Binary Cross-Entropy" divides data into 2 classes, while "sparse categorical cross-entropy" can divide them into multiple number of classes.

About the loss function:
Cross entropy is a loss function, used to measure the dissimilarity between the distribution of observed class labels and the predicted probabilities of class membership. Categorical refers to the possibility of having more than two classes (instead of binary, which refers to two classes).

About the optimizer:
"Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data." [4]

The Adam optimizer is easy to implement, efficient and has little memory requirements among other advantages.
It is best suited for "non-stationary objectives" and it fit well to the data used in this project, because the data is time based data.
Also, it is best suited for "Appropriate for problems with very noisy/or sparse gradients". The data in this project use frequencies that has no clear meaning on their own.

The Adam function use different kind of gradient descent rules than the simpler stochastic gradient descent:

- **Adaptive Gradient Algorithm** (AdaGrad) that maintains a per-parameter learning rate that improves performance on problems with sparse gradients (e.g. natural language and computer vision problems).
- **Root Mean Square Propagation** (RMSProp) that also maintains per-parameter learning rates that are adapted based on the average of recent magnitudes of the gradients for the weight (e.g. how quickly it is changing). This means the algorithm does well on online and non-stationary problems (e.g. noisy).[5]

https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/ [4]
Same source as 4 [5]

בס"ד

michael lemberger
sela goldenberg

experiment result

Using RNN with LSTM clearly gave the best result with an accuracy of 99% after a relatively small learning epochs. The best result of The Logistic Regression model only gave an accuracy of 94% and The Best Result we received with A multilayer perceptron was 95% (using A MLP with 1 hidden layer, that was the Best result we received for MLP).

```
Accuracy: 94.53%
[[483  17]
 [ 35 415]]
```

The accuracy and confusion matrix of the Logistic Regression

```
Accuracy: 95.05%
[[464  20]
 [ 27 439]]
```

The accuracy and confusion matrix of the Multilayer perceptron

Further More in order the get a 99% from the RNN Model, the model had to Trained for only about 200 epochs. in order to get The Result mention above for the Logistic regression and a multilayer perceptron The model had to Trained for 100,000 to 200,000 epochs. That a very significant Amount. using a standard dell Laptop (core i5, 16 GB Ram) The Logistic Regression Train for about 8 hours while the RNN only needed to train for about 20 minutes. After some trial and error of The Rnn model we decided to use the Adam optimizer and the sparse categorical cross entropy loss function.

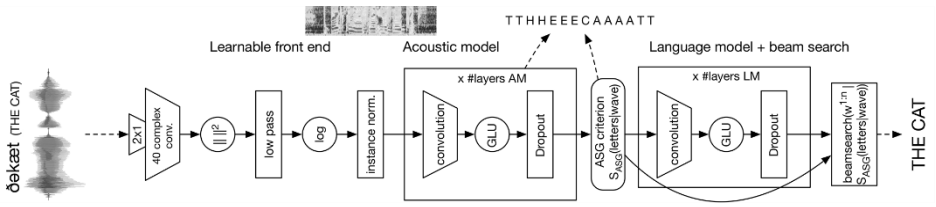
conclusion

In this paper, we developed RNN model for voice classification between male and female voices according to statistical data on the acoustic properties of the voice and speech. we presented 3 models: logistic regression, multilayer perceptron and Recursive Neural Network and showed that The RNN gave the best result.

In first glance it looks like our model suited our needs accordingly and it was shown in the result. despite of that we found that there are a number of different approaches to this problem, one of them was suggested by Facebook and its seem like the results are better than this project suggest.

michael lemberger
sela goldenberg

It was suggested by Facebook[6] the use of convolutional neural networks (CNNs) for speech and voice recognition.



Facebook CNN for speech recognition

[6] https://engineering.fb.com/ai-research/wav2letter/