LATINOWARE 2020
XVII Congresso Latino-americano de Software Livre e Tecnologias Abertas
100% ON-LINE
Latin.Science

Automatic Spoken Language Identification using Convolutional Neural Networks

Lucas Rafael Stefanel Gris
Federal University of Technology - Paraná, Brazil
Medianeira, Brazil
gris at alunos dot utfpr dot edu dot br

Arnaldo Candido Junior
Federal University of Technology - Paraná, Brazil
Medianeira, Brazil
arnaldo at utfpr dot edu dot br

***Abstract*—Automatic Spoken Language Identification systems classify the spoken language automatically and can be used in many tasks, for example, to support Automatic Speech Recognition or Video Recommendation systems. In this work, we propose an automatic language identification model obtained through a Convolutional Neural Network trained over audio spectrograms on Portuguese, English and Spanish languages. The audio for the model training was obtained through audiobooks and different corpora for speech recognition systems. The audios were used to generate instances having five seconds each. We addressed the limitation of having few speakers in our dataset with simple data augmentation techniques such as speed and pitch changing on the original instances to increase the size of the dataset. The proposed model was optimized with a random hyperparameter search which provided a final model able to identify the proposed languages with 83% of accuracy on a new, unseen test data, made with audios from different sources.**

Keywords—Spoken Language Identification;Convolutional Neural Networks;Deep Learning.

I. Introduction

Language Identification (LID) systems are very useful in the Automatic Speech Recognition (ASR) contexts. Examples of applications that benefit from LID are automatic video captioning, automatic generation of minutes, automatic customer routing in call centers, video based content recommendation, among others. The area is actively researched [9], [1], [10], [12], including different approaches for audio base selection, preprocessing techniques and feature extraction.

Deep learning has been widely used in audio processing and it is one of the best approaches available for the LID task. In particular, Convolutional Neural Networks (CNNs), while being originally conceived for image processing, have a high potential to detect patterns in audio, especially when a visual representation like spectrograms are in use.

This work presents a methodology to create datasets and build CNNs for LID and related tasks, as speaker identification. The proposed model automatically detects the language of audio spectrograms in Portuguese, Spanish and English. Several CNN topologies were evaluated in a corpus build specifically for this task composed of five second audios, mostly extracted from audiobooks and freely available resources.

The text is organized as follows. Section 2 presents related work. Section 3 shows the process to build, train and evaluate the models. Section 4 discusses the results and presents the most accurate model. Finally, Section 5 contains the final remarks of the work.

II. Related Work

Among the first techniques for the LID task, the use of I-vectors was popular choice [2] [5]. More recently, due to advances in computational power, machine learning algorithms and with the rise of deep learning based models, other structures for representing audio signals, such as spectrograms, became popular choices, combined with the use of specialized neural network models.

The use of I-vectors can be considered useful for dimensionality reduction. Dehak and Torres-Carrasquillo [2] presented a method using I-vectors for dimensionality reduction on LID tasks, and used other machine learning algo-rithms such as GMMs (Gaussian Mixture Models) and SVMs (Support Vector Machines). They showed that the proposed method works well on the 2009 Language Recognition Evaluation (LRE) challenge.

Montavon [6] proposed the use of deep learning techniques for LID tasks, mainly comparing two neural networks architectures (one deep and one shallow). The data used by the author was obtained from radio sources and somecorpus such as VoxForge. It was used convolutional layers on a time-delay neural network for feature extraction purposes. The model was fed with mel spectrograms from English, German and French audios.

Richardson et al. [10] also rely on the use of I-vectors, but it proposes the use of deep learning for both language recognition and speaker recognition. The model presented a 55% reduction on avg, the main metric used in the LRE benchmark.

Zazo et al. [12] proposes the use of Long Short Term Memory (LSTM) on LRE benchmark. They showed that with short utterances it is possible to recognize well the language being spoken. Audios up to 3 seconds are ideal, however they showed that even audios around 2 seconds can be used for LID tasks.

Bartz et al. [1] created new datasets from different freely available resources such as news from YouTube channels. They developed a deep CNN architecture with a LSTM layer to identify English, Deutch, French and Spanish. The authors solved the problem of LID in the image domain, feeding the network with 500x129 spectrogram images.

III. Methodology

A. Corpus Compilation

A corpus was compiled in order to generate our datasets, their instances and to train the proposed models. The corpus consisted mainly from audiobook record extracted from Librivox[1]. Librivox is a public available audiobook source. About 15 audiobooks were chosen from each target language. Each audiobook sample consists of a selected chapter read by a unique speaker of intercalated genders. Additionally, audios from corpora originally created for Automatic Speech Reconigition (ASR) task were incorpored in our dataset. Besides Librivox, the main corpus used was Ciempiess (Spainish) [4], Common Voice (English)[2], LapsBM[3] [8] (Portuguese), LibriSpeech[4], Spoltech Brazillian Portuguese [11], and VoxVorge[5](English and Portuguese). All corpora used are free for research, except for Spoltech.

The selected audios were then preprocessed. Audio fragments having less than 1% of the maximum volume and at least 1 second of during were considered silence and then removed from the audios. We also converted all audio sources to the sample rate of 8 khz, mainly because some audio was originated from radio recordings, which can produce bias. Besides that, the low sample rate can improve the learning speed. The low sample rate is sufficient for the LID model training, considering normal voice range is about 500 Hz to 2 KHz, even a 4 Khz sampler ate suffices to represent voice.

B. Data Augmentation

After preprocessing, data augmentation techniques were applied on the training set aiming to improve the learning and prevent bias. This step is important considering we are using mostly audiobooks as data source, which presents a poor variety of speakers. Three data augmentation techniques were applied: speed changing, pitch changing (eight rates) and background noise insertion (four noise types). The eight speed rates used were 5, 10, 15 and 20% speed for both speed reduction and increasing. Pitch changing used the same rates for both pitch reduction and increasing. Finally, driving, street, crowd and ambiance noises from Free Sound[6] were used to insert background noises into the instances. A previously study was performed on this dataset to check if the data augmentation is a valid technique for this problem and suggested that some techniques improved the learning.

All original and augmented audios were divided in 5-second pieces. During this process, audios with more than 5 seconds were split to produce instances of 5 seconds audio lengths and audios with less than 5 seconds were discarded. Our previous study showed that speed and pitch changing produced a notable effect to the model training and improves the model capability while noise addition does not have a noticeable effect.

An analysis in the dataset was performed in order to verify whether the instances are balanced by gender. There is more audios from males than females, resulting in a slight unballance in the three languagens. Man corresponds to 67, 60 and 54% of the instances in Spanish, English and Portuguese, respectively.

C. Resulting Datasets

Table 1 presents the training dataset created. After data augmentation and balancing, it was possible to obtain a 687 hours dataset from 25 hours of original instances. The dataset contains approximately 165 thousands instances for each language. Therefore 96.3% instances of the training dataset are artificial. The resulting dataset contains above 8 thousands original instances for English, while Portuguese and Spanish have approximately 2 and 7 thousands, respectively. This difference occurs due to the number of training resources available for each language. As a result, more data augmentation was required for Portuguese Language than the other two.

[1] https://librivox.org/

[2] https://voice.mozilla.org/

[3] http://laps.ufpa.br/falabrasil

[4] http://www.openslr.org/index.html

[5] http://www.voxforge.org/

[6] http://www.freesound.org

TABLE I
CREATED DATASET

Language	Original Instances	Aug. Instances	Balanced Original Instances	Balanced Aug. Instances	Total Instances
EN	11,570	230,559	8,019	157,082	165,102
PT	3,644	267,179	2,271	162,831	165,102
ES	7,869	157,233	7,869	157,233	165,102

Additionally, in order to build a validation set, 3,095 instances were extracted from two sources. The first one was unused audios from the original corpora, representing 23.5% of the validation set. The second one was from six audiobooks, corresponding to 76.5%. No data augmentation technique was applied to build this set.

Finally, the test set was build in similar manner than the validation set, but having data extracted from audiobooks, Youtube[7] news channels and freely available podcasts such as JavaScript Air[8]. Different sources for the test set were chosen to prevent biases in the evaluation. In total, 528 instances per language were selected to compose the test set.

D. Model Training

To train the model, we processed in real time each audio, generating log-powered linear spectrograms. In our hardware, this generating process saved resources, while the CPU was performing the necessary computation related to the spectrogram generation, and the GPU was trainning the model. We used the Librosa[9] and Scipy[10] libraries to load the audio and compute the spectrogram, respectively. With the respect to the spectrogram computation, we used the hann window, and a length for each segment equal to 160 (a window size of 20 times the sample rate), and, finally, a number of points to overlap each segment equal to 80 (a step size of 10 times the sample rate).

Our topology is based on the topology proposed by [7]. However, we used a fine-tuning phase in which a random search in hyper-parameter space was performed aiming to improve the model performance in our dataset.

The selected topology is presented in Figure 1 and Table 2. The five hyper-parameters tested and their respective values from our best model are: learning rate (0.0001), filters per layer (details in Table 2), number of neurons in the hidden fully connected layer (128), dropout regularization for convolutional layers (0.2) and optimizer (RMSprop). A unique dropout value for all convolutional layers were adopted. Other hyper-parameters such as number of layers and kernel size were not searched. Details about CNNs and their hyper-parameters can be found on [3]. At least three values were tested for each hyper-parameter, except for the learning rate, in which two values were evaluated. In total, 40 networks were tested in the search and 10 of them did not converge.

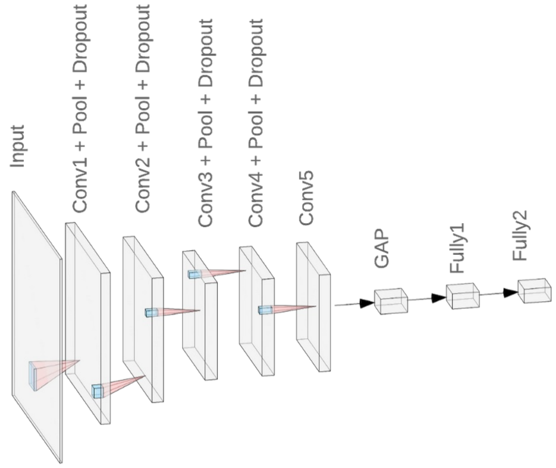

Fig. 1. Network Topology.

Regarding other hyper-parameters, all convolutional layers uses stride set to 1. Besides convolutional and fully conected layers, the model also uses pooling (Max Pooling) and Global Average Pooling (GAP). The later allows the network to produce confidence maps and improves the overall performance [7]. Most of the activation is done by the function Rectified Linear Unit (ReLU), while the output layer uses the Softmax function. The training was performed during 15 epochs for all models, which resulted in 100 steps per epoch. The chosen batch size was 10.

TABLE *2*
FINAL MODEL HYPERPARAMETERS

Layer	Size	Depth	Kernel	Activation
Input	499x81	1		
Conv1	497x79	400	3x3	ReLU
Pool1	248x39	400	2x2	
Conv2	246x37	200	3x3	ReLU
Pool2	123x18	200	2x2	
Conv3	121x16	400	3x3	ReLU
Pool3	60x8	400	2x2	
Conv4	58x6	256	3x3	ReLU
Pool4	29x3	256	2x2	
Conv5	27x1	64	3x3	ReLU
GAP	1x1	64		
Fully1	127			ReLU
Fully2	3			Softmax

[7] https://www.youtube.com
[8] https://javascriptair.com/
[9] https://librosa.github.io/
[10] https://docs.scipy.org/

IV. RESULTS

After the blind search, the best model was selected, retrained during 500 epochs, a batch size of 6 and a step per epoch equal to 400, and tested against the test set. Figures 2 and 3 presents the accuracy and the loss of the resulting model. Similar values for the training and validation sets presented suggest that there is no overfitting taking place. Furthermore, it can be observed that the model stabilizes around the 150th epoch. Figure 4 presents the confusion matrix for the three target languages.

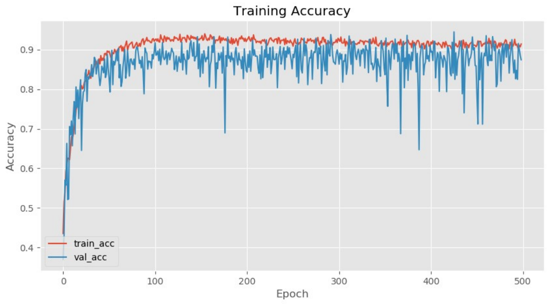

Fig. 2. Training and validation accuracy of the best topology.

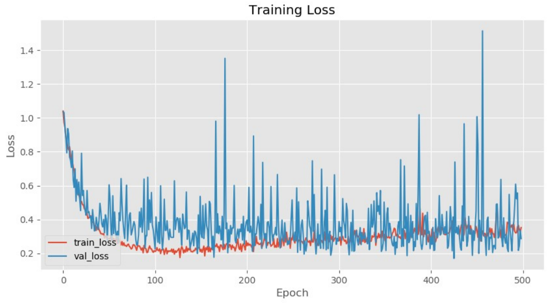

Fig. 3. Training and validation loss of the best topology.

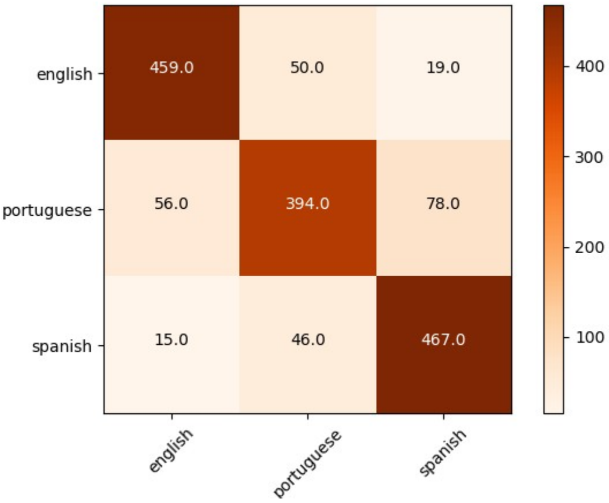

Fig. 4. Confusion matrix of the best model.

As show in Figure 2, the validation accuracy is above 90%. The next step was to analyze how this model performs in the test set. This evaluation showed an 83% accuracy on a new test data formed by audio from different sources. This rate allows several uses of the model. For example, one can filter large unlabeled audio bases to extract audios from Portuguese only. As there may occur errors in this process, it is possible to minimize them by selecting only the highest confidence predictions for the target language.

We also performed an evaluation dividing the test set according to source types. The best test accuracy is 96.8%, obtained using audiobooks. As expected, this result is possible due to the better audio quality of this type of audios. Corpora, podcasts, and Youtube sources resulted in accuracies of 91.5%, 73.6% and 87.8%, respectively.

The trained model presented recall for the Portuguese language. This can be explained by the small number of resources available for this language when compared to the other two. It also can be observed that the confusion between Portuguese and the other two languages is higher than the confusion between English and Spanish, possibly due to the same reason, while it would be expected more errors between Portuguese and Spanish considering both are romance languages.

V. CONCLUSIONS

This work presented the process used to build a corpus and to create CNN for LID. We have focused on the three most spoken languages on North and South Americas: English, Spanish and Portuguese. In order to train our model, we have adapted an existing CNN, optimizing its hyper-parameters through a blind search. Our model reached 83% accuracy and can be used in tasks that benefit from LID, such as a preprocessing step in unlabeled audio for extracting a specific language, specially if only high confidence values are used during this step.

Our model perform well on new data that is similar to the data used on the training process. Besides that, we achieved a good accuracy within YouTube audios. In this case, we show that our model is perfectly capable to label new instances on these scenarios.

Unfortunately, our model performed poorly on the podcasts/radio set. It might be possible that the selected audio source was not formed only by speeches, but also by music and noises that confused our model.

It is perfectly possible to build a CNN model for LID tasks, even with freely available data. We show that a 5-second length audio contains the necessary information for LID models, and the data augmentation process can increase the dataset size considerably. We also show that with a low sample rate we can train a model even with low quality audios. We suggest that it is possible to improve the model accuracy with new audio sources, preferably with more quality, and a model with more convolutional layers.

ACKNOWLEDGMENTS

We would like to thank Itaipu Technological Park (PTI) for the financial support to this paper, specially the members of Latin American Center of Open Technologies (CELTAB): Gustavo Rafael Valiati, Jonios Costa Maximo, Miguel Diogenes Matrakas and Rolf Massao Satake Gugisch.

We also gratefully acknowledge the support of Nvidia Corporation with the donation of the GPU used for this research.

REFERENCES

[1] Bartz, C., Herold, T., Yang, H., Meinel, C.: *Language identification using deep convolutional recurrent neural networks.* In: International Conference on Neural Information Processing. pp. 880–889. Springer (2017).

[2] Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D., Dehak, R.: *Language recognition via i-vectors and dimensionality reduction*. In: Twelfth annual conference of the international speech communication association (2011).

[3] Goodfellow, I., Bengio, Y., Courville, A.: *Deep learning.* MIT press (2016).

[4] Hernández-Mena, C.D., Herrera-Camacho, J.: *Ciempiess: A new open-sourced mexican spanish radio corpus*. In: Proceedings of the ninth international conferenceon language resources and evaluation (LREC'14). pp. 371–375. European LanguageResources Association (ELRA) Reykjavik, Iceland (2014).

[5] Li, H., Ma, B., Lee, C.H.: *A vector space modeling approach to spoken language identification*. IEEE Transactions on Audio, Speech, and Language Processing15(1), 271–284 (2006).

[6] Montavon, G.: *Deep learning for spoken language identification* (01 2009)

[7] Oponowicz, T.: *Spoken language identification* (2018), https://github.com/tomasz-oponowicz/spoken_language_identification.

[8] Quintanilha, I.M., Biscainho, L.W.P., Netto, S.L.: *Towards an end-to-end speech recognizer for portuguese using deep neural networks*. XXXV Simpósio Brasileirode Telecomunicações e Processamento de Sinais pp. 709–714 (2017).

[9] Revay, S., Teschke, M.: *Multiclass language identification using deep learning onspectral images of audio signals* (2019).

[10] Richardson, F., Reynolds, D., Dehak, N.: *Deep neural network approaches to speaker and language recognition.* IEEE signal processing letters22(10), 1671–1675 (2015).

[11] Schramm, M., Freitas, L., Zanuz, A., Barone, D.: Cslu: *Spoltech brazilian portuguese version 1.0* ldc2006s16 (2006).

[12] Zazo, R., Lozano-Diez, A., Gonzalez-Dominguez, J., Toledano, D.T., Gonzalez-Rodriguez, J.: *Language identification in short utterances using long short-term memory (lstm) recurrent neural networks*. PloS one11(1), e0146917 (2016).