https://eventos.utfpr.edu.br//sicite/sicite2019

Reconhecimento automático de idioma falado utilizando Redes Neurais Artificiais Convolucionais

Automatic spoken language recognition using Convolutional Artificial Neural Networks

Lucas Rafael Stefanel Gris
gris@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Medianeira, Paraná, Brasil

Arnaldo Candido Junior
arnaldocan@gmail.com
Universidade Tecnológica Federal do Paraná, Medianeira, Paraná, Brasil

RESUMO

Sistemas de reconhecimento automático de idioma classificam automaticamente o idioma falado e podem ser utilizados em muitas tarefas, em especial, podem fornecer o idioma falado para a correta verificação de gramática em sistemas de reconhecimento de voz. Neste trabalho, propõe-se um modelo de reconhecimento automático de idioma obtido com o treinamento de uma Rede Neural Convolucional a partir de espectrogramas de falas nos idiomas português, inglês e espanhol. Os áudios para treinamento do modelo foram obtidos através de audiolivros e diferentes *corpus* destinados a sistemas de reconhecimento de voz. O modelo foi otimizado através de uma busca por hiperparâmetros de forma randômica o que proporcionou a obtenção de um modelo final capaz de reconhecer os idiomas propostos.

PALAVRAS-CHAVE: Redes neurais. Sistemas de reconhecimento de padrões. Inteligência artificial. Reconhecimento automático da voz.

Recebido: 19 ago. 2019.
Aprovado: 01 out. 2019.

ABSTRACT

Automatic spoken language identification systems classifies the spoken language automatically and can be used in many tasks, in special, can provide the language for the correct grammar check on speech recognition systems. In this work, it is proposed an automatic language identification model obtained through the training of a Convolutional Neural Network with audio spectrograms on the Portuguese, English and Spanish languages. The audio for the model training were obtained through audiobooks and different corpus for speech recognition systems. The model was optimized with a random hyperparameter search which provided a final model capable to identify the proposed languages.

KEYWORDS: Neural networks. Pattern recognition systems. Artificial intelligence. Automatic speech recognition.

INTRODUÇÃO

Sistemas de reconhecimento automático de idioma (LID) podem ser extremamente úteis, principalmente no contexto de reconhecimento automático de fala (ASR). As pesquisas na área tiveram início em 1952 a partir de um sistema dependente de locutor que era capaz de reconhecer dígitos falados (Juang and Rabiner, 2005). Pesquisas recentes na área de reconhecimento de idioma incluem (Revay and Teschke, 2019), (Bartz et al., 2017), (Richardson et al., 2015), (Zazo et al., 2016) e (Montavon, 2009). Nestes trabalhos, diferentes abordagens foram utilizadas quanto a seleção de bases de áudios, pré-processamento e extração de características para a tarefa de reconhecimento de idioma.

Uma das técnicas clássicas para reconhecimento de idioma é a utilização de i-vectors (Dehak et al. 2010), apenas recentemente, com o avanço das técnicas em aprendizado profundo, os sistemas classificadores de idioma passaram a utilizar com mais frequência Redes Neurais para a obtenção dos modelos.

As Redes Neurais convolucionais podem ser interessantes na área de reconhecimento de áudio em geral. Apesar de serem redes inicialmente projetadas para a identificação de padrões em imagens, sua capacidade de identificar padrões através do compartilhamento de pesos e receptores locais (LeCun et al. 1995) pode ser extremamente eficaz para reconhecer fonemas em discursos. Este trabalho utiliza a representação do áudio em imagem através da geração de espectrogramas para alimentar um modelo de Rede Neural Convolucional baseado em (Oponowicz, 2018).

Este trabalho mostra como os dados de treinamento podem ser obtidos e aumentados para a obtenção de um dataset adequado para a tarefa de classificação de idiomas. O modelo final foi rapidamente otimizado através de uma busca aleatória em um determinado espaço de hiperparâmetros e demais configurações de quantidade de filtros, quantidade de neurônios e valor de dropout. O melhor modelo foi obtido através da realização de experimentos e da seleção do melhor desempenho em sua melhor época no conjunto de validação, obtendo uma acurácia de 83% no conjunto de teste proposto.

MATERIAL E MÉTODOS

A primeira tarefa realizada foi a obtenção dos dados para a realização dos treinamentos dos modelos. Os áudios foram obtidos principalmente através de gravações de audiolivros Librivox[1] nos idiomas português, inglês e espanhol e também de diversos corpus nestes idiomas. Os principais *corpus* utilizados foram a Ciempiess (Mena and Camacho, 2014), Spoltech Brazillian Portuguese (Schramm et al., 2006), VoxVorge[2] e a Common Voice[3].

Os áudios foram separados em três conjuntos distintos de treinamento, validação e teste. Cada conjunto preparado possui as mesmas características de

[1] https://librivox.org
[2] http://www.voxforge.org/
[3] https://voice.mozilla.org

pré-processamento, mas divergem entre técnicas de aumento de dados empregadas, origem dos dados, e locutores.

O conjunto de validação utilizado possui características similares ao conjunto de treinamento, já que sua origem foi a mesma, ou seja, ambos foram processados a partir de áudios dos corpus e dos audiolivros obtidos. Entretanto, foram tomadas ações para que o resultado da validação pudesse ser próximo do resultado a ser obtido no teste final, já que o modelo final seria adquirido na melhor época do treinamento, isto é, quando a acurácia de validação tivesse seu maior pico. Para garantir que o conjunto de validação fosse o mais fiel possível, os locutores dos audiolivros foram cuidadosamente selecionados de forma que dois locutores de cada sexo de cada idioma e não utilizados no conjunto de treinamento fossem destinados a validação, e os áudios dos corpus foram randomizados e selecionados.

O conjunto de teste também foi preparado de forma similar ao conjunto de validação, selecionando locutores de audiolivros que não foram utilizados no treinamento e também na validação. Os áudios de corpus não foram usados neste caso, ao invés destes, foram processados algumas instâncias a partir de podcasts e notícias disponíveis no YouTube. O objetivo foi a obtenção de uma base de teste adequada para a avaliação final do modelo, simulando seu uso em um ambiente real.

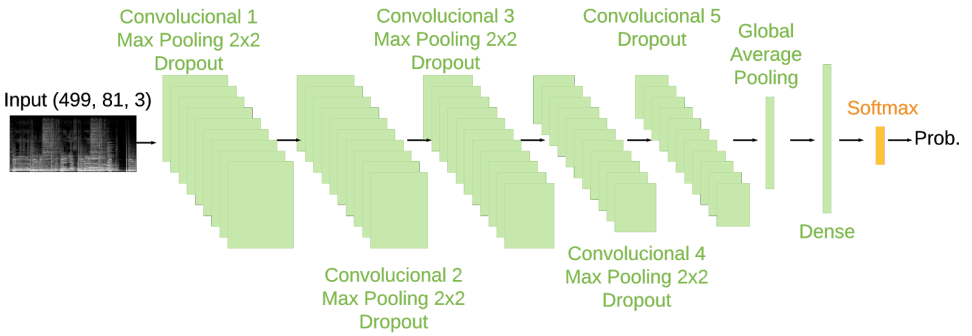
Por fim, a maior parte dos áudios de audiolivros e corpus foi destinada ao conjunto de treinamento. Além do pré-processamento padrão realizado nos demais conjuntos, o conjunto de treinamento foi aumentado utilizando técnicas de aumento de dados como adição de ruído de fundo, mudança de entonação e mudança de velocidade. Cerca de 95% dos áudios neste conjunto são artificiais e foram obtidos através do aumento de dados empregado.

O aumento de dados foi utilizado principalmente porque existe pouca variedade de locutores nas instâncias obtidas a partir dos audiolivros. Essa técnica ajuda a evitar o sobreajuste do modelo e também a geração de vícios, isto é, o aprendizado de conceitos que não são interessantes para o problema, como a qualidade do áudio ou a memorização de um locutor em específico.

Seguindo a proposta deste trabalho, foram elaboradas algumas arquiteturas de Redes Neurais para o treinamento do modelo. A arquitetura que melhor respondeu inicialmente aos dados foi a arquitetura baseada no modelo proposto por (Oponowicz, 2019). Esta arquitetura utiliza o poder de reconhecimento de padrões em imagens promovido pelas CNNs e a representação gráfica do áudio em espectrogramas como dado de entrada para a rede. A arquitetura utilizada nos experimentos pode ser vista na figura 1.

Utilizando a arquitetura inicial proposta, foram realizadas buscas por hiperparâmetros de forma randômica quanto ao número de filtros em cada camada convolucional, o valor de dropout e o número de neurônios na camada densa anterior a camada de ativação softmax. Também foram testadas possibilidades quanto ao otimizador utilizado (Adam, RMSprop e Adagrad) e taxa de aprendizado de 0.001 e 0.0001.

Figura 1 – Arquitetura utilizada



Fonte: Autoria própria (2019).

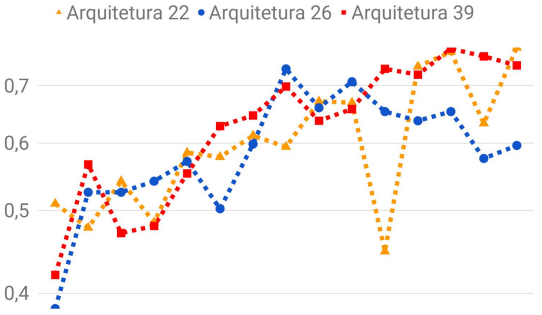
Foram observados os resultados dos primeiros 40 testes realizados, e as três arquiteturas que produziram a maior acurácia de validação em sua melhor época foram selecionados para a realização do treinamento do modelo final. Todos os treinamentos foram realizados extraindo as características (espectrogramas) no momento do treinamento para alimentar a rede.

Foram propostos alguns experimentos a partir das arquiteturas obtidas na busca por hiperparâmetros. Os experimentos foram executados utilizando todo o dataset e o melhor modelo de cada experimento foi analisado. Alguns experimentos foram realizados variando o tamanho do batch e a quantidade de passos utilizados por iteração. Procurou-se a realização do treinamento de forma que todos os dados fossem passados a rede pelo menos 2 vezes.

RESULTADOS E DISCUSSÃO

Quanto a otimização dos hiperparâmetros, das 40 configurações testadas, 10 configurações não convergiram e poucas obtiveram um resultado promissor. De uma forma geral, no entanto, a taxa de aprendizado de 0,0001 e o otimizador RMSprop pareceram ser os mais adequados se comparado às demais configurações testadas. As arquiteturas com a acurácia de validação mais alta em sua melhor época foram selecionadas e o seu desempenho ao longo das épocas pode ser visualizado na figura 2, enquanto suas respectivas configurações podem ser vistas na tabela 1.

Figura 2 – Desempenho das três melhores arquiteturas testadas



Fonte: Autoria própria (2019).

Tabela 1 – Melhores arquiteturas e seus hiperparâmetros

Hiperparâmetro	Arq. 22	Arq. 26	Arq. 39
Filtros. Conv. 1	200	400	100
Filtros. Conv. 2	400	200	400
Filtros. Conv. 3	200	400	500
Filtros. Conv. 4	128	256	32
Filtros. Conv. 5	32	64	64
Densa 1	128	128	256
Dropout	0,4	0,2	0,3
Otimizador	RMSprop	RMSprop	RMSprop
Taxa de Apre.	0,0001	0,0001	0,0001

Fonte: Autoria própria (2019).

A arquitetura 22 apresenta configurações menos robustas (menor quantidade de parâmetros para treinamento), e uma taxa de dropout consideravelmente alta. A arquitetura 26 é talvez, a mais robusta das três, e possui a menor taxa de dropout. Já a configuração 39 possui configurações medianas, se comparada com as arquiteturas 22 e 26.

Os experimentos finais foram realizados utilizando todo o dataset e as arquiteturas 22, 26 e 39. Os resultados obtidos na validação e no teste na melhor época podem ser vistos na tabela 2. O experimento 5 produziu os melhores resultados, com uma acurácia no conjunto de teste igual a 83%. A matriz de confusão desse teste pode ser vista na tabela 3.

Tabela 2 – Resultados finais

Exp.	Arquitetura	Batch	Épocas	Passos	Ac. Teste
1	39	100	500	20	0,65
3	39	6	500	400	0,75
5	**26**	**6**	**500**	**400**	**0,83**
8	22	6	500	400	0,79

Fonte: Autoria própria (2019).

Tabela 3 – Matriz de confusão do melhor modelo

	En	Pt	Es
En	459	50	19
Pt	56	394	78
Es	15	46	467

Fonte: Autoria própria (2019).

CONCLUSÃO

A identificação de idioma pode ser considerada uma tarefa desafiadora, principalmente porque deve-se identificar os fonemas dos idiomas em intervalos curtos de áudio, e muitos idiomas compartilham uma série desses fonemas. Neste trabalho, foram exploradas algumas estratégias para a elaboração de um modelo de Rede Neural Convolucional capaz de identificar o idioma falado a partir da geração de espectrogramas de áudios com duração de cinco segundos, desde as bases selecionadas e técnicas de aumento de dados até estratégias de otimização da arquitetura da rede e busca por hiperparâmetros. O modelo final foi capaz de reconhecer 83% de um dataset de teste composto por áudios de fontes diferentes das utilizadas para os conjuntos de treinamento e validação.

Apesar da eficácia em reconhecer os idiomas propostos, nota-se que o resultado poderia ser ainda melhor, especialmente se o modelo fosse mais profundo, ou se os hiperparâmetros fossem melhor ajustados. Para trabalhos futuros, pode ser interessante a exploração dessas estratégias e o treinamento de modelos para o reconhecimento de outros idiomas.

AGRADECIMENTOS

Agradeço ao PTI (Parque Tecnológico Itaipu) pelo apoio ao projeto e pela bolsa de iniciação científica. Também agradeço a UTFPR pelo apoio financeiro dado ao projeto através do edital 03/2018 de inovação da UTFPR-MD. Agradecimentos também ao meu orientador por todo o apoio desde o início.

REFERÊNCIAS

Bartz, C., Herold, T., Yang, H., and Meinel, C. (2017). Language identification using deep convolutional recurrent neural networks. In International Conference on Neural Information Processing, pages 880–889. Springer.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification.IEEE Transactions on Audio, Speech, and Language Processing, 19(4):788–798.

JUANG, B.-H. AND RABINER, L. R. (2005). Automatic speech recognition–a brief history of the technology development.Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1:67.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and timeseries.The handbook of brain theory and neural networks, 3361(10):1995.

Mena, C. D. H. and Camacho, A. H. (2014). Ciempiess: A new open-sourced mexican spanish radio corpus. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H.,Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors,Proceedings of the Ninth International Conference on Language

Resources and Evaluation(LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

Montavon, G. (2009).Deep learning for spoken language identification. In NIPS Workshop on deep learning for speech recognition and related applications, pages1–4.

Oponowicz, T. (2018). spoken_language_identification. https://github.com/tomasz-oponowicz/spoken_language_identification.

Revay, S. and Teschke, M. (2019). Multiclass language identification using deep learning on spectral images of audio signals.arXiv preprint arXiv:1905.04348.

Richardson, F., Reynolds, D., and Dehak, N. (2015). Deep neural network approaches to speaker and language recognition.IEEE signal processing letters, 22(10):1671–1675.

Schramm, M., Freitas, L., Zanuz, A., and Barone, D. (2006). Spoltech brazilian portuguese version 1.0 idc2006s16.Philadelphia: Linguistic Data Consortium.

Zazo, R., Lozano-Diez, A., Gonzalez-Dominguez, J., Toledano, D. T., and Gonzalez-Rodriguez, J. (2016). Language identification in short utterances using long short-term-memory (lstm) recurrent neural networks.PloS one, 11(1):e0146917.