

Projeto Demonstrativo 6

Reconhecimento de Fala

Hevelyn Sthefany Lima de Carvalho
170059031

hevelyn.sthefany@gmail.com

Igor Bispo de Moraes
170050432

igor.rabbit99@gmail.com

Departamento de Ciência da
Computação

Universidade de Brasília

Campus Darcy Ribeiro, Asa Norte
Brasília-DF, CEP 70910-900, Brazil,

Abstract

Ruído, pitch e variabilidade de velocidade sempre foi um desafio para identificar comandos de voz. Apresentamos um modelo eficiente para reconhecimento de fala. A arquitetura da rede constiue-se de convoluções separáveis em profundidade para construir redes neurais profundas leves. Apresentamos testes obtidos pela rede treinada com forte desempenho e comparamos com os resultados obtidos pela rede com pesos pré-treinados com o ImageNet.

1 Introdução

O Reconhecimento de Fala, também conhecido como reconhecimento automático de fala (automatic speech recognition - ASR), reconhecimento de fala por computador ou fala para texto (computer speech recognition or speech to text - STT), é um subcampo de Processamento de Linguagem Natural que se concentra na capacidade e nas limitações de uma máquina em entender a linguagem dos seres humanos. O processo consiste em mapear uma entrada de áudio para alguma palavra existente em um certo vocabulário.

Neste trabalho, transportamos o problema de reconhecimento de áudio para o âmbito de classificação de imagens, uma área muito explorada em Visão Computacional. A análise acústica tem a espectrografia do som como uma de suas principais ferramentas. O espectrograma pode ser definido como uma mostragem dinâmica da densidade de energia por meio do escurecimento ou coloração do traçado, as faixas de frequência no eixo vertical e o tempo no eixo horizontal. Sua representação mostra estrias horizontais, denominadas harmônicos. A classificação das imagens foram obtidas através de Rede Neural Convolutacional (CNN), mais precisamente, a arquitetura MobileNet [1].

O conjunto de dados trabalhados, Speech Commands Data Set - versão 0.02, tem um vocabulário relativamente pequeno de 30 [1]. O conjunto compõe-se de falas de 20 palavras de comando gravadas por uma variedade de falantes diferentes: "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go", "Zero", "One", "Two", "Three", "Four", "Five", "Six", "Seven", "Eight", e "Nine". Para ajudar a distinguir palavras não reconhecidas, há também dez palavras auxiliares: "Bed", "Bird", "Cat", "Dog", "Happy", "House", "Marvin", "Sheila", "Tree", e "Wow".

Algumas das ferramentas auxiliares principais foram OpenCV [8] para tratamento de imagens e bibliotecas do Python: Numpy [9] para manipulação de tensores (imagens), a API Keras para desenvolvimento da rede [10] e o módulo signal do SciPy [11] para geração dos espectrogramas.

2 Metodologia

Pré-processamento

Para cada audio da base da dados, foram gerados espectrogramas

Arquitetura da Rede

MobileNet [12], proposta pelo Google, foi a arquitetura para a CNN usada neste trabalho a qual usa convoluções separáveis em profundidade.

Uma convolução separável em profundidade é composta por duas operações: uma convolução em profundidade e uma convolução pontual. A convolução em profundidade mapeia uma única convolução em cada canal de entrada separadamente, ou seja, aplica um único filtro a cada canal de entrada (camada separada para filtragem). Portanto, seu número de canais de saída é o mesmo do número de canais de entrada. Seu custo computacional é $Df^2 \cdot M \cdot Dk^2$, sendo Df a dimensão do recurso de entrada, M e N o número de canais de entrada e saída e Dk o tamanho do kernel. A Figura 2 (a) e (b) mostra a diferença entre a Convolução normal e a separável. Já a pontual, Figura 2 (b) é uma convolução com um tamanho de kernel de 1x1 que combina os recursos criados pela convolução em profundidade (camada separada para combinação), com custo computacional de $M \cdot N \cdot Df^2$.

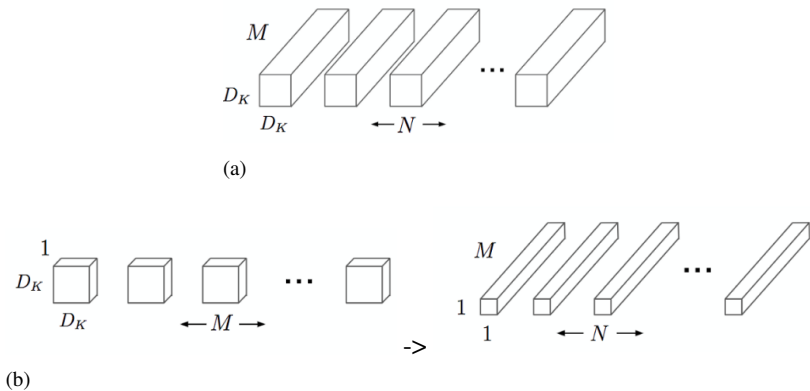


Figure 1: Convolução padrão (a) e Convoluções separáveis em Profundidade (b) e (c)

Todas as camadas são seguidas por uma normalização em lote, batchnorm [13], especificada na equação abaixo, onde z^k é a camada, $E[x]$ e $V[x]$ são o primeiro e segundo epoch de x respectivamente, e a função de ativação ReLU [14], com exceção da camada final totalmente conectada que não possui não-linearidade e alimenta uma camada softmax para

classificação.

$$\tilde{z}^k = \frac{z^k - E[z^k]}{\sqrt{V[z^k]}} \tag{1}$$

Os detalhes da arquitetura da rede são mostrados no arquivo Anexo.

Resultados

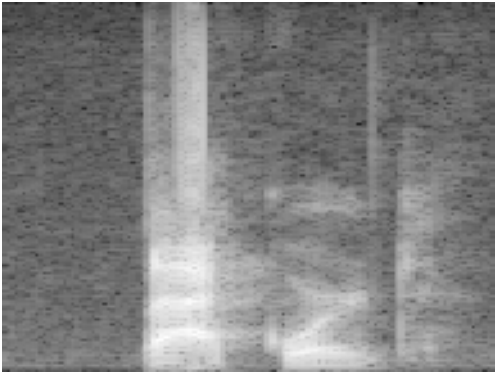


Figure 2: Espectrogram (fala de “backward”)

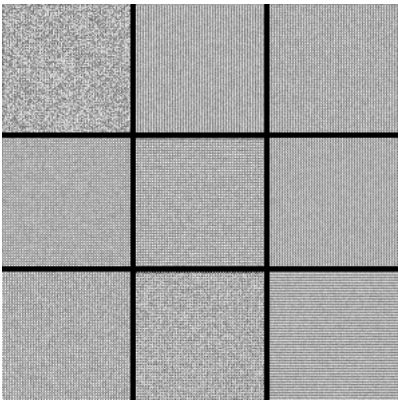


Figure 3: Filtros aplicados na Convolução em Profundidade

	precision	recall	f1-score	support
bed	0.95	0.98	0.97	500
bird	0.96	0.99	0.98	500
cat	0.96	0.99	0.97	500
dog	0.97	0.96	0.97	500
down	0.95	0.93	0.94	500
eight	0.95	0.97	0.96	500
five	0.98	0.91	0.94	500
four	0.97	0.93	0.95	500
go	0.88	0.95	0.92	500
happy	1.00	0.99	0.99	500
house	0.99	0.99	0.99	500
left	0.99	0.95	0.97	500
marvin	0.99	0.98	0.99	500
nine	0.96	0.95	0.96	500
no	0.99	0.86	0.92	500
off	0.98	0.93	0.96	500
on	0.95	0.94	0.95	500
one	0.97	0.94	0.96	500
right	0.99	0.95	0.97	500
seven	0.95	0.98	0.96	500
sheila	0.99	0.98	0.98	500
six	0.94	0.98	0.96	500
stop	0.94	0.95	0.94	500
three	0.99	0.97	0.98	500
two	0.97	0.96	0.96	500
up	0.95	0.96	0.95	500
wow	0.98	0.98	0.98	500
yes	0.98	0.98	0.98	500
zero	0.97	0.95	0.96	500
micro avg	0.	0.96	0.	14500
macro avg	0.97	0.79	0.96	14500
weighted avg	0.97	0.96	0.96	14500

Table 1: Relatório de Classificação

Comparação com pesos treinados com ImageNet

Usamos pesos pré-treinados com o ImageNet [2] para comparação. Apesar da base de dados do ImageNet não ter características semelhantes às imagens de espectrograma, pois contém mais de 3 milhões de imagens aleatórias sendo nenhuma espectrograma, o resultado foi indiferente. Além disso, a convergência é mais rápida, atingindo o resultado com 2 epochs a menos que o método anterior.

3 Conclusão

Neste projeto foi abordado o problema de reconhecimento de fala aplicando o modelo MobileNet nos dados de imagens formados pelos espectrogramas do banco de dados Speech Commands Data Set. O método foi implementado com sucesso, com 0.91 % de acurácia de precisão. Além disso, destacamos a indiferença entre o método treinando os pesos e o método com os pesos pré treinados com ImageNet, apesar da aleatoriedade deste banco de imagens, no que se trata de acurácia. No entanto, neste segundo método, a convergência é mais rápida.

References

- [1] François Chollet et al. Keras. <https://keras.io>, 2015.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [3] NumPy developers. Numpy, Aug 2018. URL <http://www.numpy.org/>.
- [4] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [6] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL <http://www.scipy.org/>. [Online; accessed <today>].
- [7] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, pages 807–814, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. URL <http://dl.acm.org/citation.cfm?id=3104322.3104425>.
- [8] OpenCV team. Opencv, Aug 2018. URL <https://opencv.org/>.
- [9] P. Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv e-prints*, April 2018. URL <https://arxiv.org/abs/1804.03209>.