Disciplinas: Algoritmos de Reconhecimento de Padrões e Redes Neurais Profundas

Reconstrução de arquivos de áudio com rede convolucional

Lucas Roque

INSTITUTO DE INFORMÁTICA UFG

Introdução

Diversas técnicas de *Machine Learning* vem sendo aplicadas para resolver o problema da modelagem de áudio. Tais trabalhos permitiram avanços em áreas como a geração de áudio [4], reconhecimento [5] e classificação de voz[1]. A maioria desses trabalhos manipulam arquivos de áudio em formato *raw*, que apesar de possibilitar uma maior flexibilidade de manipulação, trás consigo um alto custo computacional.

Com base nesse cenário [3] propôs uma abordagem que minimizasse este problema, por meio de uma técnica de reconstrução de áudio de alta qualidade a partir de uma amostra contendo apenas uma fração das informações do sinal original (entre 15 e 50%). Esta técnica pode ser aplicada em telefonia, compressão e geração de texto, além de sugerir novas arquiteturas para geração de áudio. A rede proposta é conceitualmente simples, pois opera diretamente no arquivo de áudio bruto, escalável, já que utiliza redes convolucionais e *feed-foward*, além de já ter sido testada em arquivos de áudio sem fala.

O objetivo deste trabalho é aplicar esta técnica em um ambiente diferente do anteriormente exposto, utilizando um *dataset* na língua portuguesa, com áudios reais, medindo assim medir sua eficácia em outros cenários.

Materiais e métodos

Arquitetura de Rede

Dado um sinal de baixa resolução, o objetivo da abordagem proposta é reconstruir a versão de alta resolução, para isso é utilizada a rede descrita na Figura 1. Inicialmente a entrada é submetida técnica de *upsampling* cúbico proposta por [2], para projeta-la para o espaço de alta dimensão. Após este tratamento inicial, a entrada passa por uma série de camadas de *downsampling*, composta por uma camada de convolução, normalização *batch*, e aplicação de ReLU, como pode ser visto na Figura 2. A redução da dimensionalidade é realizada com a aplicação de um *stride* de tamanho dois, e dobrando a quantidade de filtros em cada camada.

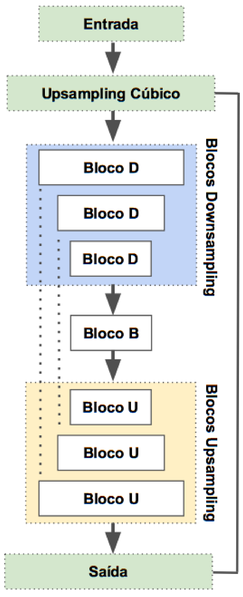

Figura 1: Arquitetura de rede utilizado no trabalho.

A amostra é reconstruída a partir dos recursos aprendidos através de uma série simétrica de camadas *upsampling*, detalhadas na Figura 2. Para que seja possível utilizar a características de baixa resolução durante o *upsampling*, foi criado uma conexão com a camada de *downsampling*. Por semelhante modo, foi estabelecida uma conexão entre a camada de *upsampling* cúbico com a saída, assim, o modelo precisa apenas melhorar a aproximação cúbica.

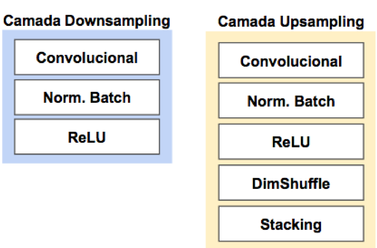

Figura 2: Estrutura interna das camadas de *downsampling* e *upsampling*.

Dataset

O treinamento e avaliação da arquitetura foi realizado com duzentas amostras de áudio com duração variada entre dois e oitos segundos. As amostras foram retiradas de quatro discursos do presidente do Brasil, Michel Temer. Foram selecionados discursos com diferentes condições de áudio, variando entre ambientes abertos e fechados, com e sem ruídos.

O modelo foi instânciado com oito camadas, sendo quatro de *downsampling* e quatro de *upsampling*. O treinamento foi realizado durante 400 épocas usando o otimizador ADAM, como taxa de aprendizado de 10^{-4}, decaindo linearmente após a métade das épocas.

Resultados

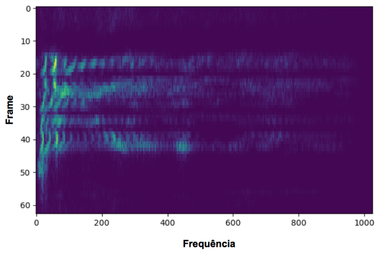

Figura 3: Espectrograma dos sínais de uma amostra original.

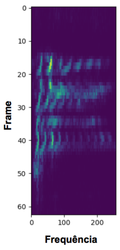

Figura 4: Espectrograma dos sínais de uma amostra em baixa relosução.

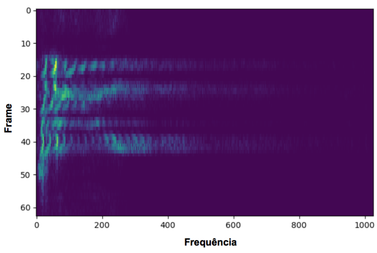

Figura 5: Espectrograma dos sínais de uma amostra reconstruida pela abordagem.

A Figura 5 mostra o resultado de uma avaliação da rede experimentada. Como pode ser visto, a configuração implementada na rede foi capaz de reconstruir os principais sinais da amostra, porém alguns aspectos foram perdidos. Tal fato pode ser explicado pelo fato de que, ao contrário do *dataset* utilizado pelo autor da proposta, onde as amostras continham apenas voz, o *dataset* atual possui ruído em alguma das amostras. Um aumento na quantidade de épocas pode também melhorar a qualidade da reconstrução.

Considerações finais

Os testes realizados demonstraram que a rede foi capaz de reconstruir o áudio original com relativa eficiência. Porém é necessário realizar mais experimentos para conseguir determinar se o motivo da perda está no cenário do *dataset*, ou nos parâmetros da rede.

Agradecimento

Aos colegas pelo aprendizado em conjunto, ao meu orientador Professor Celso Camilo pela oportunidade e incentivo, e ao Professor Anderson Soares pelos ricos ensinamentos e ajuda.

Referências

[1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[2] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[3] V. Kuleshov, S. Z. Enam, and S. Ermon. Audio super-resolution using neural networks. 2017.

[4] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[5] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville. Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*, 2017.