XXVIII Congresso {virtual} de Iniciação Científica da Unicamp 2020

Rotulação Automática de Músicas Usando Redes Neurais Profundas

Gustavo Da Silva Tafarello Salessi, Marcos Eduardo Ribeiro do Valle Mesquita*

Resumo

Rótulos de músicas são metadados usados, por exemplo, para organização de coleções musicais ou em sistemas de recomendação em serviços de *streaming* como *Last.fm*, *Spotify* e *Deezer*. Nesse projeto de iniciação científica, estuda as aplicações de redes neurais profundas para a rotulação automática de músicas. Redes neurais profundas constituem uma poderosa ferramenta de aprendizado de máquinas e inteligência artificial com aplicações bem sucedidas em diversas áreas, incluindo reconhecimento de padrões e sistemas de recomendação. No contexto da rotulação de músicas, redes neurais profundas aplicadas ao espectrograma na escala mel mostraram-se robustas mesmo na presença de ruídos e dados inconsistentes. Em vista dessa observação, nesse projeto de iniciação científica revisamos os principais conceitos sobre processamento de sinais e redes neurais artificiais necessários para compreender o processo usado atualmente para rotulação de música usando técnicas de aprendizado de máquina e inteligência artificial.

Palavras-chave: Rotulação de música, redes neurais profundas, processamento de sinais.

1 Sinais

O ponto inicial é o conceito de "sinal", que se refere ao processo de transmitir informações em algum formato, a informação sempre está contida em algum tipo de variação [4]. De forma matemática, sinais podem ser representados como funções de uma ou mais variáveis independentes e podem ser separados em dois tipos, do tipo contínuo onde sua variação se dá de forma gradual, e do tipo discreto, onde sua variação se dá de forma abrupta. O som consiste da variação da pressão do ar no tempo, portanto é um sinal, daí a importância de se estudar esse conceito neste projeto. Dentro do computador o sinal de som é representado como um sinal digital que é do tipo discreto. A Figura 1 mostra um sinal representado por uma função de uma única variável, o foco de estudo deste projeto.

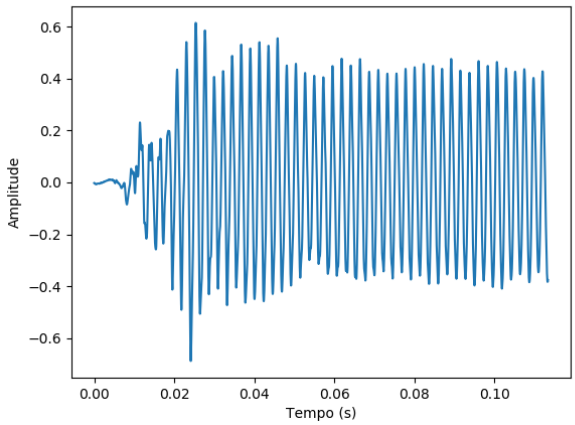

Fig. 1: Alguns segundos da nota Lá de um piano

2 Transformada de Fourier

Uma técnica de se resolver problemas muito complexos é dividi-los em problemas menores. Os problemas menores são resolvidos e combinados para formar a solução do problema original [5]. Outra ferramenta importante é salientar as características fundamentais de um determinado sinal [1]. São nesses pontos que a série de Fourier e a transformada de Fourier se tornam importantes ferramentas na análise de sinais.

2.1 Série de Fourier

Uma função $f(t)$, com $t \in \mathbb{R}$, pode ser representada pela forma complexa da série de fourier em um intervalo arbitrário L [7]:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{in\pi t}{L}}. \tag{1}$$

O coeficiente c_n é determinado pela equação:

$$c_n = \frac{1}{2L}\int_{-L}^{L} f(t) e^{\frac{-in\pi t}{L}} dt. \tag{2}$$

Fazendo a decomposição de um sinal em Série de Fourier ele está sendo expresso em uma soma ponderada infinita de sinais elementares, que são muito mais simples de usar e analisar.

2.2 Transformada de Fourier de Tempo Discreto

A transformada de fourier pode ser aplicada em um sinal contínuo ou em um discreto, no caso do segundo onde se tem apenas um número finito de valores de $f(t)$ para o intervalo finito $[0, T]$, o espaço é dividido em N pontos t_k dados por.

$$t_k = k\frac{T}{N}, \qquad \forall k = 0, 1, ..., N-1. \tag{3}$$

A transformada de Fourier da função $f(t)$ em tempo discreto é dada pela equação [7]:

$$\mathcal{F}(n) = \sum_{k=0}^{N-1} f(t_k) W^{nt_k}, \qquad \forall n = 0, 1, ..., N-1. \tag{4}$$

em que $W = e^{i2\pi/T}$

A Transformada de Fourier de Tempo Discreto transforma a função $f(t)$ do domínio de tempo para o domínio das frequências.

2.3 Transformada Rápida de Fourier

Na prática o que se utiliza é a transformada rápida de fourier (em inglês *fast Fourier trasform* ou FFT), ela é obtida atra-

vés da Transformada de Fourier de Tempo Discreto. Sem entrar em detalhes do processo ao se utilizar a FFT tem-se um ganho muito alto na quantidade de operações. No total, a FFT realiza $O(N \log_2 N)$ operações, enquanto a tranformada de Fourier de tempo discreto apresenta $O(N^2)$. Na prática essa redução faz uma enorme diferença no tempo de operação para um N grande [5]. A Figura 2 mostra a FFT da nota lá, nela pode-se notar um pico na frequência de 440 Hz que é justamente a frequência fundamental da nota, os outros picos correspondem aos múltiplos de 440.

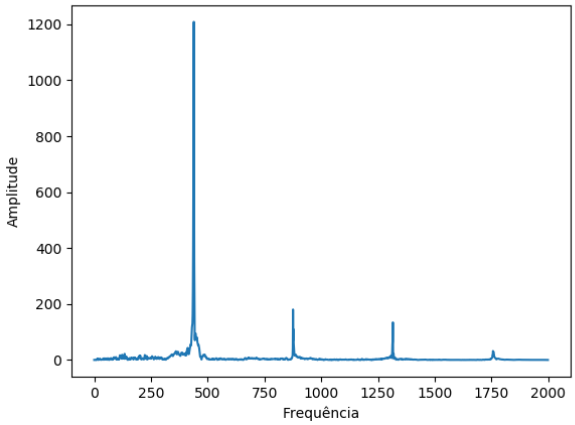

Fig. 2: Tranformada de Fourier da nota Lá do piano

3 Espectrograma

3.1 Definições e Aplicações do Espectrograma

Um espectrograma é usado para saber quais as frequências predominantes do sinal em um determinado tempo, ele é calculado dividindo o sinal em fatias, cada uma delas é caracterizada pela coordenada temporal que se encontra no centro. Multiplica-se cada fatia por uma função de janela (uma função que possui zeros fora de um determinado intervalo, usualmente simétrica), após isso, é computada a FFT desse pedaço. O objetivo da função de janela é focar a visão da transformada de Fourier nas proximidades de um determinado ponto, geralmente o central. O resultado do espectrograma é a magnitude ao quadrado da transformada de Fourier, seu gráfico possui 3 dimensões. Para representá-lo em 2 dimensões a magnitude ao quadrado da transformada de Fourier é caracterizada por um esquema de cores. A Figura 3 mostra o espectrograma da nota lá. É possível notar que existe uma faixa correspondente a frequência de 440Hz indicando que ela está está presente durante todo o sinal.

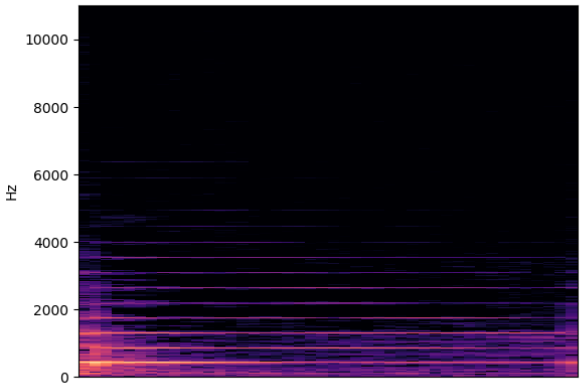

Fig. 3: Espectrograma da nota Lá do piano

4 Redes Neurais

Uma rede neural (RN) é um modelo inspirado no funcionamento do cérebro onde as unidades básicas de processamento são os neurônios. Um neurônio é descrito pela equação:

$$h = \phi\left(\sum_{j=1}^{n}(w_j x_j) + b\right) \tag{5}$$

Do ponto de vista matemático, um neurônio também é uma função $\mathcal{N} : \mathbb{R}^n \to \mathbb{R}$ que transforma o vetor $\mathbf{x} = (x_1, \ldots, x_n)$ com os valores de entrada e utilizando alguns parâmetros (w_j e b), devolve um número real esse número então passa por uma função $\phi()$ (geralmente uma função não linear) [8].

Os neurônios de uma rede neural são geralmente organizados em camadas. A arquitetura de uma rede neural em camadas pode ser entendida pela Figura 4, onde os círculos representam os neurônios.

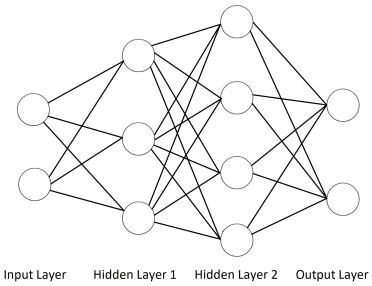

Fig. 4: Representação de Uma Rede Neural

4.1 Treinamento da Rede Neural

Os parâmetros da rede neural podem ser determinados usando um conjunto de dados, chamado dados de treinamento, que são constituídos por pares de entrada e a saída desejada. A rede neural é alimentada pelas entradas dos dados de treinamento e a resposta da rede é comparada com as saídas desejadas. Os parâmetros então são ajustados de modo que a rede reproduza de forma satisfatória os pares dos dados de treinamento e também apresente uma boa capacidade de generalização. Comparando os dados gerados pela rede neural e os dados esperados é possível criar uma função perda não linear.

O problema então torna-se achar os valores de parâmetros que minimizam essa função perda. Uma das técnicas para achar o mínimo de uma função não linear é o método do gradiente, onde dado um ponto calcula-se o gradiente da função nele e toma-se a direção oposta, o próximo ponto da iteração é nessa direção. Para calcular o gradiente da função de perda é utilizada a técnica do backpropagation, onde se utiliza a regra da cadeia.

4.2 Redes Convolucionais

Redes Neurais Convolucionais são geralmente utilizadas para análise de imagens, seu diferencial é a chamada camada de convolução. Inspirada no funcionamento do córtex visual os neurônios de uma camada convolucional são conectados apenas com um pequeno pedaço da camada anterior como ilustra a Figura 5. Como nas RN's tradicionais cada neurônio da camada anterior é multiplicado por um peso.

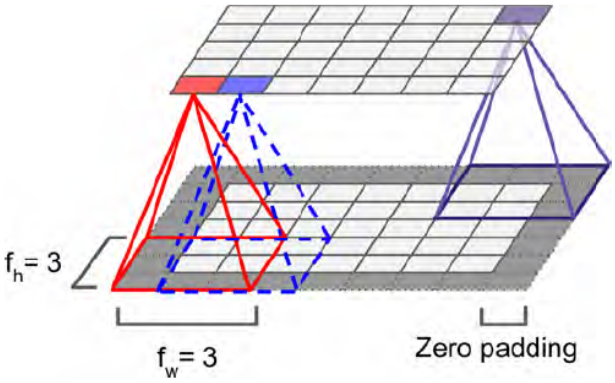

Fig. 5: Representação de Uma Rede Neural Convolucional. Imagem retirada da referência [3].

5 Base de Dados, Modelo e Testes

5.1 Base de Dados

Durante a construção do modelo foi utilizada uma base de dados para as músicas chamada GTZAN. Essa base possui 1000 pedaços de 30 segundos de músicas com seus respectivos rótulos identificando para qual estilo musical ela pertence entre 10 opções: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae e rock. Essa base de dados é muito utilizada em *machine learning*, porém possui alguns problemas que podem impactar a performance dos modelos [6]. Cada estilo musical possui 100 músicas, que foram divididas entre dados de treinamento e dados de teste para validar o modelo, 70% destinados a treinamento e 30% para testes.

Para alimentar a rede neural é preciso antes preparar os dados. Como a arquitetura da RN foi escolhida para analisar imagens, pois são algoritmos bem eficientes, é necessário transformar as músicas em uma imagem. Para isso foram gerados os espectrogramas, e cada um foi definido como uma imagem de dimensões 128 x 1290. Conforme a referência [2] indica, cortou-se as imagens em 86 partes aumentando a quantidade de dados de teste e treinamento, cada uma com dimensão de 128 x 15.

5.2 Modelo

Para os experimentos computacionais foi utilizada a biblioteca Keras, muito indicada, pois é de fácil implementação e consegue trabalhar com complexos modelos. Keras utiliza uma biblioteca chamada Tensorflow, que foi desenvolvida pelo Google, é *open sorce* e possui uma ótima documentação . A arquitetura da Rede Neural utilizada é composta por 4 camadas convolucionais, uma camada tradicional de rede neural e por fim a saída da RN.

5.3 Resultados

No final se obteve uma acurácia de 78% para os dados de treinamento e 64% para os dados de teste; um resultado interessante para o GTZAN. Como mencionado anteriormente a base de dados GTZAN possui alguns problemas, como repetições de músicas, distorções, entre outros [6].

6 Agradecimentos

Agradecimentos para Rodolfo Aníbal Lobo Carrasco pela ajuda para compreender os tópicos estudados e ao meu orientador Prof. Dr. Marcos Eduardo Ribeiro do Valle Mesquita. Por fim, agradecimentos ao CNPq por financiar este projeto.

Referências

[1] CADZOW, J. A., AND LANDINGHAM, H. F. V. *Signals, Systems, and Transforms*. Prentice-Hall, 1985.

[2] COSTA, Y. M. G., OLIVEIRA, L. S., AND JR., C. N. S. An evaluation of Convolutional Neural Networks for music classification using spectrograms.

[3] GÉRON, A. *Hands-On Machine Learning with Scitkit-Learn & TensorFlows*. O'Reilly, 2017.

[4] OPPENHEIM, A. V., AND WILLSKY, A. S. *Sinais e Sistemas*, 2 ed. Pearson Education do Brasil, 2010.

[5] PHILLIPS, C. L., AND PARR, J. M. *Signals, Systems, and Transforms*. Prentice-Hall, 1995.

[6] STURM, B. L. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use, 2013.

[7] VAZ JR., J., AND DE OLIVEIRA, E. C. *Métodos Matemáticos*, vol. 2. Editora da Unicamp, 2016.

[8] WIKISTAT. Neural Networks and Introduction to Deep Learning.