Classificação de gênero musical por meio de redes neurais profundas e diferentes representações de áudio

Micael Valterlânio da Silva, Diego Furtado Silva

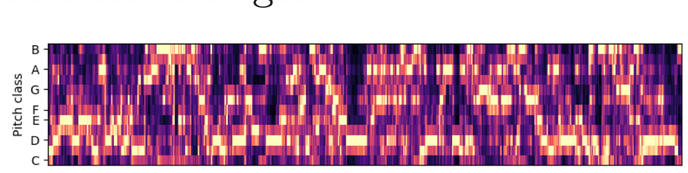
micael.sax@gmail.com | diegofs@ufscar.br

Introdução

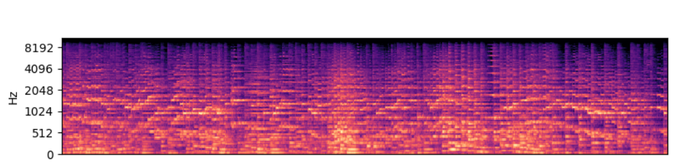
- Com o crescimento cada vez maior no cenário da música digital, plataformas como YouTube, Spotify e Deezer vêm armazenando e manipulando uma quantidade gigantesca de dados.
- Faz-se necessário pensar na melhor forma de organizar e recuperar toda essa informação musical através de métodos computacionais.
- Uma das abordagens é rotular cada música com uma informação que a represente resumidamente. Essa informação usualmente é o seu gênero musical.
- A partir daí, é possível que as plataformas de reprodução online possam organizar o seu acervo associando artistas e músicas com perfis parecidos. Abre-se então a possibilidade de criar um sistema de recomendação mais eficiente, criação automática de playlists, e até uma maneira mais assertiva de traçar um perfil de usuário.
- Existe uma subjetividade muito grande atrelada ao gênero em que uma música está associada.
- Os trabalhos que definem o estado-da-arte se baseiam na técnica de aprendizado por meio de redes neurais profundas (RNPs). [1];
- Esses trabalhos propõem métodos que aprendem a partir de uma representação visual que é obtida a partir do áudio.
- Então foi investigado se é possível obter uma melhoria na tarefa de classificação de gêneros musicais utilizando a combinação de variadas representações visuais, que evidenciam diferentes características musicais de uma gravação.

Metodologia

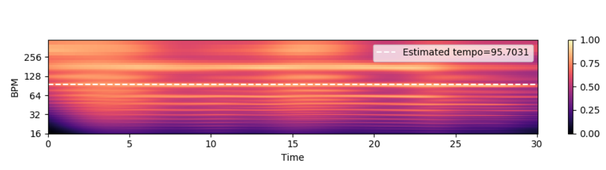
- Inicialmente, com o propósito de entender as representações de áudio mais utilizadas pela comunidade científica, e quais as características musicais elas evidenciavam, foi realizado um processamento em diversos áudios com músicas de diferentes gêneros.



Cromagrama

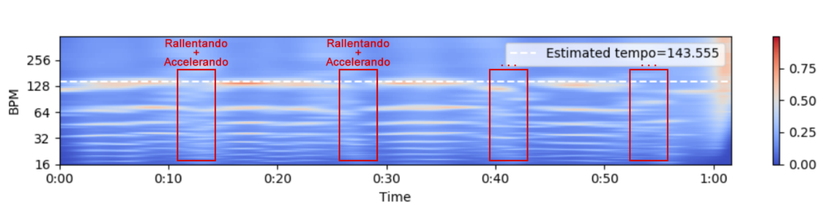


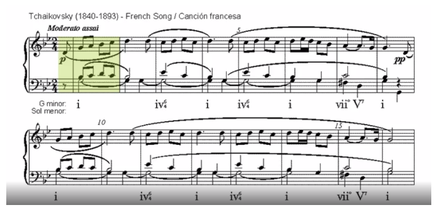
Mel-espectrograma

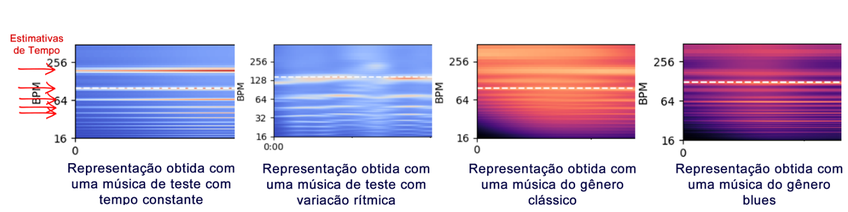


Tempograma

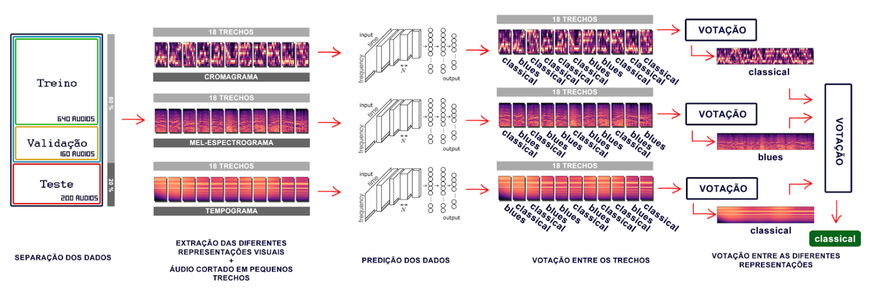
- Foram processados áudios mais simples para que fosse possível entender quais eram as informações musicais que estavam disponíveis em cada tipo de representação, e como cada característica musical (músicas com fórmula de compasso composta, figuras rítmicas de tempo maior, alterações de andamento como *rallentando* ou *accelerando*, entre outras características) se refletia nessas mesmas representações visuais.







- Após obter as predições fornecidas pelas melhores redes neurais em cada representação visual, foi feita a combinação dessas predições, como mostra a figura ao lado.



Resultados

- A figura abaixo mostra os resultados obtidos com o experimento. Observando a melhoria alcançada com a combinação de diferentes representações de áudio, a hipótese proposta pelo projeto se confirma.

Representação	Acurácia (%)	
	p/ cada trecho (1.5s)	p/ áudio inteiro (voting)
espectrograma	75,35	83,50
log frequency spectrum	70,13	83,50
mel-espectrograma	73,55	79,00
mfcc	60,29	68,50
cromagrama	48,05	63,00
cromagrama24	52,41	60,00
espectrograma harmonico	38,23	52,00
tempograma	38,08	49,00
tonnetz	39,02	55,00

COMBINAÇÃO	
Tipo de Combinação	Resultado
Voting	86
Scores Average	87
Voting by Scores	85
Scores Multiply	85,5

Referências Bibliográficas

[1] DIXON S. SIGTIA, S. Improved music feature learning with deep neural networks. In *2014 IEEE international conference (ICASSP), p. 6959-6963, 2014.*

Apoio

11º UFSCar CIDTI CIC 26º UFSCar CNPq ufscar