Style transfer of audio effects

Marcos Vinicios Pereira de Morais
Universidade federal de goiás
Goiânia - goiás
marcos.vinicios2@discente.ufg.br

Gustavo Coimbra Cavalcante
Universidade federal de goiás
Goiânia - goiás
gustavocoimbra@discente.ufg.br

Pedro Mundel Biffi
Universidade federal de goiás
Goiânia - goiás
mundel@discente.ufg.br

Enzo
Universidade federal de goiás
Goiânia - goiás
enzodias@discente.ufg.br

Enzo Lemes Marques
Universidade federal de goiás
Goiânia - goiás
enzolemes@discente.ufg.br

I. **Resumo—A pesquisa sobre geração de performance musical expressiva (EMP), é um tema "quente" na criação musical automatizada. Inegavelmente, a musicalidade humana transcende em expressividade quando comparada àquela gerada por máquinas. Para desvelar essa disparidade, é imperativo investigar o papel desempenhado pelos humanos na criação musical.**

Este artigo propõe um modelo que utiliza redes convolucionais que minimizam a perda de conteúdo e estilo, entre as duas faixas de entrada. Apresentamos um método híbrido de avaliação, combinando abordagens qualitativas e quantitativas. Este estudo visa contribuir para uma geração musical mais expressiva e fornecer uma abordagem abrangente e robusta para a avaliação desses modelos.

keywords —Music style transfer, CNN's, processamento de sinal e imagem

II. Introdução

Com a ascensão do aprendizado de máquina (ML) [1], abordagens que exploram criações artísticas estão cada vez mais ganhando popularidade entre os círculos sociais. Um exemplo disso é o DALL-E 2 [2], que possibilitou a geração única de imagens, permitindo que milhares de pessoas pudessem criar arte através de uma sequência de instruções escritas. Essa influência não se restringe apenas a criações visuais, mas estende-se também a ferramentas que manipulam sinais de voz, as quais possuem uma grande relevância na área de estudo.

A geração de música surge como um campo quente na recuperação de informação musical (MIR) [3]. O objetivo é produzir performances musicais extremamente expressivas, assemelhando-se àquelas executadas por músicos profissionais. Isso nos auxilia a compreender a conotação da música computacional e apresenta uma ampla gama de perspectivas de aplicação, incluindo composição inteligente, acompanhamento automático de música, análise de cena auditiva e classificação de gênero musical.

Existem três componentes fundamentais na criação musical: a partitura, o intérprete e o instrumento. O compositor concebe a partitura, o intérprete a aprimora com sua interpretação única, e o instrumento adequado é escolhido para a execução. É a interpretação do músico que confere à música uma autenticidade que ultrapassa a artificialidade de uma reprodução sintetizada por máquina ou renderizada diretamente da partitura.

Este artigo concentra-se na segunda etapa do processo de geração musical: o papel do intérprete através de uma rede convolucional para aprender representações espectrais específicas que minimizem a diferença entre duas faixas de áudio, uma para conteúdo e outra para estilo. O objetivo é compreender a forma como os seres humanos interpretam a música, buscando, assim, reproduzir uma expressividade musical semelhante à humana.

III. Fundamentos Teóricos

A abordagem de transferência de características de timbre focaliza a criação de novos timbres musicais, assemelhando-se à síntese sonora. O Google propôs um autocodificador de timbre com base no modelo Wavenet [4]. Este modelo tem a capacidade de reconstruir a forma de onda original, onde a camada oculta é tratada como uma representação do timbre. O novo timbre é gerado por meio de técnicas de interpolação.

A Universidade de Stanford [5] desenvolveu um sistema de transferência de estilo neural baseado em espectro utilizando a transferência de estilo de imagem para mapear o espectro musical e assim gerar música. Os estudos nesta área concentram-se principalmente na música em si, não sendo capazes de capturar a interpretação humana da expressão emocional implícita na música, resultando na ausência de uma música que se assemelhe à produzida por seres humanos.

Foi apresentado o modelo MuseNet [6], fundamentado em um kernel esparso para realizar a modelagem de padrões rítmicos recorrentes.

IV. Metodologia

A. O que é uma CNNs

Uma Rede Neural Convolucional (CNN), também conhecida como ConvNet, é uma classe de redes neurais profundas especialmente concebida para a tarefa de processamento e análise de dados de padrões espaciais, comumente utilizada em visão computacional. O design das CNNs é altamente inspirado pelo modo como o cérebro humano percebe e interpreta informações visuais.

A principal distinção das CNNs em relação a outras arquiteturas de redes neurais reside na aplicação efetiva de camadas convolucionais. Essas camadas são fundamentais para extrair características e padrões relevantes de uma imagem. Os filtros ou kernels convolucionais são utilizados para realizar operações de convolução sobre a entrada da imagem, o que permite a detecção de características locais, como bordas, texturas e formas.

A estrutura típica de uma CNN consiste em camadas convolucionais, camadas de pooling para redução de dimensionalidade e camadas totalmente conectadas para tomada de decisões finais. A aplicação sequencial dessas camadas permite que a rede aprenda hierarquias de representações cada vez mais complexas, começando por características simples e evoluindo para características mais abstratas.

Durante o treinamento, a CNN ajusta automaticamente os pesos dos filtros convolucionais por meio de algoritmos de otimização, como o gradiente descendente, para minimizar a discrepância entre as previsões da rede e as verdadeiras etiquetas dos dados de treinamento. Essa capacidade de aprendizado hierárquico e adaptativo faz com que as CNNs sejam particularmente eficazes em tarefas de classificação, detecção de objetos, segmentação de imagens e reconhecimento facial, entre outras aplicações em visão computacional.

Além de sua preeminência em visão computacional, as CNNs também foram aplicadas com sucesso em outros domínios, como processamento de linguagem natural, onde foram adaptadas para lidar com sequências de dados, destacando a versatilidade e a capacidade de generalização dessas redes neurais convolucionais.

Aqui iremos abordar as estruturas do modelo de Redes Neurais Convolucionais (CNNs) utilizado para a transferência de estilos musicais.

A arquitetura geral para a transferência de estilo musical recebe como entrada dois sinais distintos, nos quais é realizado o processo de combinar o conteúdo de uma música com o estilo de outra, utilizando uma função de perda que minimiza a diferença entre as características extraídas das duas músicas.

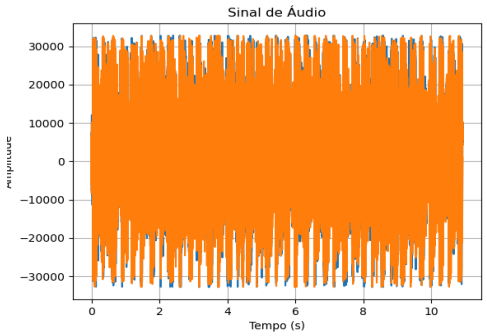

fig.1 visualização do sinal de entrada no domínio do tempo.

B. Transformação no sinal

Na fase de pré-processamento, realizamos a aplicação da Transformada de Fourier de Curto Período (STFT) nos arquivos de áudio, o que resultou na transformação do sinal, anteriormente no domínio do tempo, para o domínio da frequência. Essa etapa tem como objetivo converter o sinal de áudio bruto em um espectrograma.

Em seguida, ocorre um processo conhecido como Normalização de Magnitude. Esse método é frequentemente empregado na análise de dados para ajustar valores a uma escala uniforme. A normalização visa proporcionar uma compreensão mais consistente dos dados independentemente das diferenças nas escalas originais.

No início, são identificados os valores mínimo e máximo presentes na magnitude. Esses extremos são utilizados para normalizar os valores na magnitude. Cada valor é ajustado subtraindo-se o valor mínimo, e o resultado é dividido pela diferença entre o valor máximo e mínimo da magnitude. Isso é crucial para uma representação visual consistente e facilita a interpretação subsequente.

Por fim é gerado um espectrograma do sinal normalizado que será a entrada para o nosso modelo convolucional. O espectrograma é uma representação gráfica das frequências presentes no espectro de um sinal, proporcionando uma representação visual das variações temporais dessas frequências ao longo do tempo.

E com o espectro do sinal que será possível a obtenção de padrão de estilo durante o treinamento do modelo de Redes Neurais Convolucionais (CNNs)

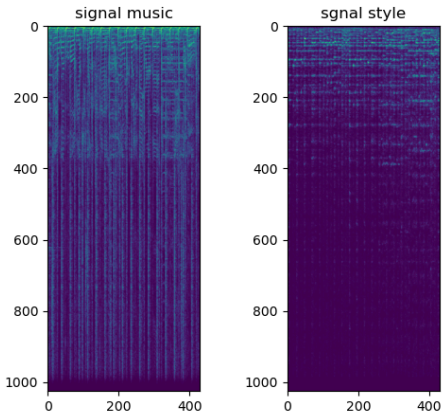

fig. 2 espectrograma do sinal transformado e processado.

C. Arquitetura do modelo

Escolhemos incorporar uma camada convolucional bidimensional devido à transformação da entrada de áudio em um espectrograma. Essa opção foi deliberada, pois a representação dos dados assume o formato de imagem, tornando a aplicação de convolução bidimensional mais apropriada e eficaz para capturar padrões espaciais e correlações inerentes às características do espectrograma. Essa abordagem visa potencializar a capacidade do modelo em extrair informações relevantes, otimizando assim o processo de aprendizado e desempenho na tarefa em questão.

O kernel da camada convolucional foi calculado com o desvio padrão das entradas. Essa prática é usada para escalar os valores gerados aleatoriamente na matriz de pesos kernel, pois o desvio padrão está relacionado à inicialização de pesos em redes neurais e a escolha apropriada do desvio padrão pode afetar o treinamento da rede de maneira positiva.

A fórmula específica usada para calcular o desvio padrão (std) é uma implementação da fórmula de inicialização de pesos proposta por He et al.(inicialização He) Na função usada, o cálculo de std é ajustado para levar em

consideração tanto o número de canais de entrada (1025) quanto o número de filtros (4096) da camada convolucional bidimensional, bem como o tamanho do kernel (11).

$$\text{std} = \sqrt{2} \cdot \sqrt{\frac{2}{\text{channels} + \text{filters} \cdot \text{kenel size}}}$$

Fig. 3 Implementação da fórmula de inicialização de pesos proposta por He et al [7].

Na configuração final do modelo, foram empregados 522 kernels com dimensões de (1x11), os quais foram aplicados em uma única camada convolucional bidimensional. No processo de pré-processamento, as dimensões da imagem foram reduzidas para apenas um canal (escala de cinza), resultando em convoluções realizadas exclusivamente nesse único canal.

O modelo construído é projetado para processar informações em três dimensões, representadas por altura, largura e canais. Notavelmente, a saída do modelo é concebida para manter inalteradas as dimensões espaciais da entrada, evitando qualquer redução na dimensionalidade. Essa abordagem preserva a integridade da representação original.

Além disso, a função de ativação 'ReLU' é incorporada ao modelo, conferindo-lhe não linearidade e o habilitando a aprender padrões complexos durante o processo de treinamento. Essa característica é crucial para a capacidade do modelo de capturar nuances e relações mais intrincadas nos dados, tornando-o mais eficaz na compreensão e interpretação de informações complexas.

D. Treinamento do modelo

Para o treinamento do modelo, foi utilizada a técnica de transferência de estilo neural. implementamos um processo iterativo no qual uma imagem gerada aleatoriamente é otimizada para minimizar uma combinação de perda de conteúdo e perda de estilo em relação a imagens de música que será preservada inicialmente,a perda total é calculada como a soma ponderada dessas duas componentes.

Durante cada iteração do treinamento, os gradientes são calculados em relação à imagem gerada usando a técnica de diferenciação automática fornecida pelo TensorFlow. Esses gradientes são então utilizados para atualizar a imagem gerada através de um otimizador Adam.

A utilização da matriz Gram entra no cálculo da perda de estilo. A função o cálculo da perda de estilo comparando as matrizes Gram das características de estilo da imagem de referência e da imagem gerada. A matriz Gram é uma representação estatística das características de estilo, e a diferença quadrática entre essas matrizes é calculada. Essa perda de estilo é então ponderada por um fator de 0.001 antes de ser somada à perda de conteúdo.

O objetivo final desse processo é gerar uma imagem que preserve o conteúdo da imagem de referência enquanto incorpora o estilo da outra imagem de referência. As perdas de conteúdo e estilo são monitoradas e impressas a cada 50 passos durante o treinamento para avaliar o progresso.

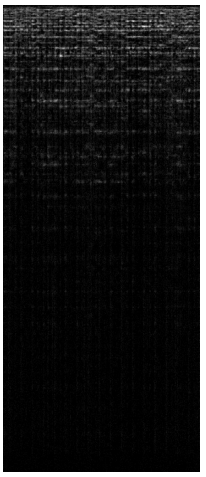

fig 4: , áudio processado com a transferência de estilo.

Ao final do processo obtivemos um áudio rotulado de output.wav e um pnj output, o primeiro é o áudio processado com a transferência de estilos e o segundo é o espectrograma do sinal transformado.

V. Resultados e Conclusões

Ao concluir o trabalho, alcançamos êxito na realização da transferência de estilo musical, destacando-se pelo baixo uso de poder computacional, requerendo apenas 15 minutos para a conclusão eficiente dos processos. Enquanto muitas abordagens se concentram na utilização de autoencoders e autodecoders para a execução da transferência de estilo, nossa abordagem inovadora se baseou em um modelo convolucional.

Os resultados foram notáveis, com nosso modelo impressionando um grupo teste, onde aproximadamente quatro quintos das faixas produzidas foram avaliadas como criativas e totalmente inovadoras, atingindo assim nosso primeiro objetivo. Os detalhes completos e os resultados do nosso trabalho estão disponíveis em Github/Grupo---9-pdsi-/results.

Vale ressaltar que para músicas que possuem performance, nosso modelo não permite o controle de preservá-las, na maioria das vezes o resultado final possui ruído do que era uma vez a voz, mas observamos que em uma minoritária quantidade de vezes a voz foi perfeitamente preservada.

VI. REFERENCIAS

[1]Zhe Xiao, Xin Chen "Music performance style transfer for learning expressive musical performance"19 October 2023

[2] Lazaros Moysis; Lazaros Alexios Iliadis; Sotirios P. Sotiroudis; Achilles D. Boursianis; Maria S"Music Deep Learning: Deep Learning Methods for Music Signal Processing—A Review of the State-of-the-Art"IEE

[3]Christian J. Steinmetz, Nicholas J. Bryan, Joshua D. Reiss"Style Transfer of Audio Effects with Differentiable Signal Processing"18 Jul 2022

[4]Carlos E. Cancino-Chacón, Maarten Grachten, Werner Goebl, Gerhard Widmer"Computational Models of Expressive Music Performance: A Comprehensive and Critical Review"24 October 2018