UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Benjamim Alves Nepomuceno Neto

Avaliação de Arquiteturas Convolucionais aplicadas ao Reconhecimento de Símbolos Musicais

São Luís

2023

Benjamim Alves Nepomuceno Neto

Avaliação de Arquiteturas Convolucionais aplicadas ao Reconhecimento de Símbolos Musicais

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Geraldo Braz Junior

São Luís

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Alves Nepomuceno Neto, Benjamim.
Avaliação de Arquiteturas Convolucionais aplicadas ao Reconhecimento de Símbolos Musicais / Benjamim Alves Nepomuceno Neto. - 2023.
38 p.

Orientador(a): Geraldo Braz Junior São Luís 2023.
Curso de Ciência da Computação, Universidade Federal do Maranhão, UFMA, 2023.

1. Partituras Musicais Manuscritas. 2. Reconhecimento de Imagens. 3. Redes Neurais. I. Braz Junior São Luís 2023, Geraldo. II. Título.

Benjamim Alves Nepomuceno Neto

Avaliação de Arquiteturas Convolucionais aplicadas ao Reconhecimento de Símbolos Musicais

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho aprovado em São Luís, 19 de julho de 2023:

Prof. Dr. Geraldo Braz Junior
Orientador

Prof. Dr. João Dallyson Sousa de Almeida
Examinador

Profa. Dr. Darlan Bruno Pontes Quintanilha
Examinador

São Luís
2023

Agradecimentos

Agradeço primeiramente a Deus pela oportunidade e pela iluminação nessa fase da minha vida. A minha família pelo apoio, aos meus professores que foram muito compreensivos, presentes e dispostos a responder minhas dúvidas, em especial ao professor Geraldo Braz, professor ao qual cursei varias disciplinas durante a graduação.

"Uma doutrina errada, baseada numa busca sincera, vale muito mais que a contemplativa segurança daqueles que se opõem a tal busca por acreditarem já "saber": saber, sem haverem buscado por si! "

Arnold Schönberg, em *"Harmonia"*

Resumo

A classificação de simbolos musicais manuscritos e digitais é um problema com uma vasta complexidade, visto que uma vasta quantidade de partituras ainda não estão devidamente digitalizadas e classificadas, este trabalho visa cobrir parte desse problema de classificação, em específico de simbolos musicais manuscritos, a qual poderá servir para classificar simbolos musicais manuscritos de autores clássicos e criação de aplicativos de reconhecimento de símbolos de partituras manuscritas. Este estudo investigou o uso de redes neurais para a classificação de simbolos musicais manuscritos em partituras manuscritas. Três arquiteturas de redes neurais foram exploradas: ConvNext, DenseNet e ConvMixer. Essas arquiteturas foram aplicadas a um conjunto de dados de simbolos musicais manuscritas e avaliadas em termos de acurácia. Os resultados obtidos mostraram que a arquitetura ConvMixer alcançou uma acurácia de 98,25%, enquanto a DenseNet obteve uma acurácia de 95,26% e a ConvNext apresentou uma acurácia de 94,2%. Esses resultados destacam a eficácia das redes neurais na classificação de simbolos musicais manuscritos, contribuindo para avanços na área de reconhecimento óptico de notas musicais. Essas abordagens promissoras podem ser exploradas para melhorar a automação e precisão do reconhecimento de partituras manuscritas.

Palavras-chave: Partituras Musicais Manuscritas; Reconhecimento de Imagens; Redes Neurais; ConvMixer; DenseNet; ConvNext.

Abstract

The classification of handwritten and digital musical symbols is a problem with a vast complexity, since a vast amount of scores are not yet properly digitized and classified, this work aims to cover part of this classification problem, specifically of handwritten musical symbols, which can serve to classify handwritten musical symbols of classical authors and create applications for recognizing symbols of handwritten scores. This study investigated the use of neural networks for the classification of handwritten musical symbols in manuscript scores. Three neural network architectures were explored: ConvNext, DenseNet and ConvMixer. These architectures were applied to a dataset of handwritten musical symbols and evaluated in terms of accuracy. The results obtained showed that the ConvMixer architecture achieved an accuracy of 98.25%, while DenseNet obtained an accuracy of 95.26% and ConvNext presented an accuracy of 94.2%. These results highlight the effectiveness of neural networks in classifying handwritten musical symbols, contributing to advances in the area of optical recognition of musical notes. These promising approaches can be exploited to improve the automation and accuracy of handwritten sheet music recognition.

Keywords: Handwritten Sheet Music; Image Recognition; Neural Networks; ConvMixer; DenseNet; ConvNext.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Classificação de imagens de simbolos musicais utilizando redes neurais é um problema relevante na área de processamento de imagens e música. Esse problema ainda requer grandes avanços devido aos varios níveis de dificuldades e condições para ser solucionado, sendo um subproblema da transcrição de partituras, sendo parte essencial para a solução dessa . O objetivo neste tipo de abordagem é desenvolver um sistema automatizado capaz de reconhecer e classificar corretamente as simbolos musicais presentes em uma partitura.

A identificação automática de símbolos manuscritos é um desafio complexo que tem sido abordado com o auxílio de redes neurais. A escrita à mão possui características únicas e variáveis, tornando a tarefa de reconhecimento de símbolos uma tarefa desafiadora. As redes neurais surgem como uma solução promissora devido à sua capacidade de aprender e generalizar a partir de exemplos fornecidos durante o treinamento.

A identificação automática de símbolos manuscritos é uma tarefa desafiadora, mas essencial para facilitar a interação entre humanos e sistemas computacionais. O avanço nessa área é crucial para permitir o desenvolvimento de aplicações mais inteligentes e eficientes que dependem do reconhecimento e interpretação de símbolos escritos à mão (LECUN et al., 1998).

No entanto, o desafio reside na variabilidade intrínseca da escrita à mão, onde diferentes pessoas podem ter estilos de escrita distintos e a mesma pessoa pode variar sua escrita em diferentes momentos. Além disso, há o desafio da presença de ruídos, inclinações, distorções e sobreposições de símbolos que podem dificultar o processo de identificação automática.

As redes neurais podem superar esses desafios por meio de sua capacidade de aprendizado e representação de características relevantes. Ao treinar uma rede neural com um conjunto diversificado de exemplos de símbolos manuscritos, ela pode aprender a extrair informações discriminativas e generalizar para símbolos não vistos anteriormente. O uso de redes neurais convolucionais e redes neurais recorrentes tem se mostrado especialmente eficaz nessa tarefa, permitindo a modelagem de padrões espaciais e temporais presentes na escrita à mão.

À medida que os avanços na área de redes neurais e aprendizado de máquina continuam, espera-se que esses sistemas se tornem cada vez mais precisos e robustos, permitindo uma ampla gama de aplicações práticas que se beneficiam do reconhecimento de símbolos escritos à mão de forma automatizada.

No contexto do acervo de notas musicais, existem bibliotecas e bancos de dados disponíveis que podem ser utilizados para treinar e avaliar os modelos de classificação. Um exemplo conhecido é o dataset de partituras de notas musicais do MusicNet (MUSICNET, 2017). Esse dataset contém milhares de partituras de músicas clássicas e contemporâneas, proporcionando uma base sólida para o treinamento de redes neurais.

Além do dataset do MusicNet, outro importante recurso para o treinamento e avaliação de modelos de classificação de notas musicais é o HOMUS (Handwritten Optical Music Understanding System) (HOMUS. . . , 2016). O HOMUS é um banco de dados de partituras manuscritas, desenvolvido por Rafael Caro Repetto e Gustavo A. Sánchez-Ante no ano de 2016. Ele contém uma variedade de partituras manuscritas de diferentes estilos musicais, permitindo a exploração de desafios adicionais relacionados à variabilidade na escrita e notação musical.

1.1 Trabalhos Relacionados

Nos últimos anos, diversos estudos foram conduzidos na área de classificação de simbolos musicais transcritos utilizando técnicas de aprendizado de máquina. Esses trabalhos visam automatizar o processo de identificação e categorização das notas musicais em transcrições, proporcionando avanços significativos na análise e processamento de partituras. As abordagens empregadas envolvem o uso de redes neurais convolucionais, redes recorrentes, algoritmos de processamento de imagens e técnicas de processamento de sinais, visando aprimorar a precisão e a eficiência na classificação das notas musicais transcritas. Esses estudos contribuem para o desenvolvimento de sistemas de reconhecimento musical mais robustos e precisos, com aplicações em áreas como transcrição automática de partituras, aprendizado de máquina musical e análise computacional de música.

O artigo “A Deep Approach for Handwritten Musical Symbols Recognition” (PEREIRA et al., 2016) apresenta uma metodologia para reconhecimento de notas musicais em partituras manuscritas digitalizadas usando técnicas de aprendizado profundo. O objetivo é desenvolver um sistema de Reconhecimento Óptico de Música (OMR) preciso e robusto para partituras musicais manuscritas. A metodologia proposta utiliza uma rede neural convolucional profunda (Convolutional Neural Network - CNN) e foi testada em um conjunto de dados de referência, alcançando uma taxa mínima de erro de 3,99%, precisão de 96,46% e recall de 96,56% no conjunto de dados HOMUS. A rede neural utilizada na metodologia proposta é a GoogLeNet (SZEGEDY et al., 2015), uma arquitetura avançada de CNN treinada para reconhecer símbolos musicais manuscritos.

O artigo intitulado “Handwritten Music Recognition for Mensural Notation with Convolutional Recurrent Neural Networks” (CALVO-ZARAGOZA; TOSELLI; VIDAL, 2019) aborda o reconhecimento de música escrita à mão em um tipo específico de notação

musical chamada "notação mensural"em que os ritmos e durações das notas eram indicados de forma proporcional em relação a uma unidade de tempo básica. utilizando redes neurais convolucionais e recorrentes. O trabalho propõe um sistema baseado em redes neurais profundas que se mostrou suficientemente bem-sucedido para o reconhecimento de notação mensural. Os experimentos realizados demonstraram uma melhoria significativa em relação às abordagens holísticas anteriores, reduzindo a taxa de erro no nível do símbolo de 25,7% para 7,0%.

O artigo intitulado “A Deep Learning-Based Piano Music Notation Recognition Method” (LI, 2022) discute um método para reconhecer notações musicais de piano usando técnicas de aprendizagem profunda. O artigo se concentra na análise de partituras de piano e na extração de características relevantes para obter um reconhecimento preciso da partitura de piano. Os autores utilizam métodos de reconhecimento digital para extrair a matriz de características da partitura digital de piano, incluindo pontos de frequência de multiplicação e a função de envelope. As informações musicais extraídas são então convertidas em arquivos MIDI, permitindo a reconstrução da partitura e a transmissão de áudio. Os resultados experimentais demonstram uma taxa de acerto de 94,4% na extração de informações musicais de partituras de piano, demonstrando a aplicabilidade prática do método proposto. Este artigo é relevante para este contexto sobre classificação de notas musicais transcritas usando redes neurais, ao fornecer ideias sobre uma abordagem específica para reconhecer e analisar notações musicais de piano, o que pode contribuir para o desenvolvimento de técnicas semelhantes nessa linha de pesquisa.

Dito os trabalhos relacionados é importante resaltar as diferenças em relação ao feito nesse trabalho,primeiramente ultilizaremos um dataset de simbolos musicais manuscritos, o que foi feito em 2 trabalhos citados anteriormente (PEREIRA et al., 2016) e (CALVO-ZARAGOZA; TOSELLI; VIDAL, 2019), depois analisaremos três arquiteturas de redes neurais modernas, em um processo similar ao publicado em (PEREIRA et al., 2016), tendo em diferenças que analizaremos três arquiteturas ao invés de apenas uma e a modernidade das arquiteturas analizadas. O presente trabalho difere do publicado em (LI, 2022) por não abodar simbolos digitais, mas tendo o objetivo em comum de classificação destes

1.2 Objetivos

Este estudo tem como objetivo geral avaliar três arquiteturas de redes neurais no contexto da classificação de símbolos musicais manuscritas, um desafio complexo devido à diversidade de formas e variações dos símbolos musicais. A classificação precisa dos símbolos é fundamental para várias aplicações musicais, como a transcrição automática de partituras e a análise computacional de músicas. Através da comparação do desempenho das três arquiteturas em termos de métricas de avaliação, buscamos identificar qual delas é

mais eficaz para esse problema específico. Ao compreender as vantagens e desafios de cada abordagem, poderemos direcionar futuros trabalhos de pesquisa para aprimorar ainda mais a precisão e a eficiência do reconhecimento de símbolos musicais por meio de redes neurais.

A primeira arquitetura selecionada, a ConvMixer, apresenta uma abordagem inovadora ao combinar operações convolucionais e de transformação linear em uma única camada. Essa arquitetura consegue capturar informações espaciais e de frequência nas imagens dos símbolos musicais, permitindo que o modelo aprenda padrões complexos e sutis. Espera-se que a ConvMixer possa aprender efetivamente as características das símbolos manuscritas e realizar uma classificação precisa, superando os desafios impostos pela diversidade e complexidade da notação musical.

A segunda arquitetura, a DenseNet, oferece uma abordagem única para a construção de redes neurais, com conexões densas entre camadas. Essa arquitetura promove uma aprendizagem profunda e eficiente, permitindo que cada camada receba informações das camadas anteriores. No contexto da classificação de símbolos musicais, o DenseNet tem o potencial de aprender eficientemente as características dos símbolos e obter resultados promissores. A conexão densa entre camadas facilita a propagação das informações relevantes, tornando o modelo mais robusto e capaz de lidar com variações nas imagens das notas.

A terceira arquitetura, a ConvNext, utiliza conexões residuais para melhorar o fluxo de informações nas camadas convolucionais. Essas conexões residuais permitem que as informações sejam transmitidas diretamente de uma camada para outra, facilitando o aprendizado de características relevantes das notas musicais. O ConvNext foi selecionado devido à sua capacidade de extrair informações discriminativas das imagens das notas e realizar uma classificação precisa. Espera-se que essa arquitetura contribua significativamente para a solução do problema de classificação de notas musicais manuscritas.

1.2.1 Objetivos Específicos

Destacam-se como objetivos específicos deste trabalho:

- Avaliar as arquiteturas ConvMixer, Densenet e Convnext utilizando o dataset HOMUS, conjunto de dados de símbolos musicais manuscritos.
- Comparar os resultados obtidos pelas três arquiteturas, considerando as métricas de precisão e recall por classe. Serão identificadas as diferenças de desempenho entre as arquiteturas e serão discutidas as vantagens e desafios de cada uma na tarefa de classificação de notas musicais manuscritas.

1.3 Organização do Trabalho

Este trabalho está organizado em cinco capítulos, para apresentar o conteúdo mais claramente, conforme os parágrafos a seguir.

Capítulo 2 aborda a fundamentação teórica dos conceitos assim como as arquiteturas utilizadas nesse trabalho. Nesse capitulo, além da fundamentação teórica das arquiteturas, também é abordado sobre as métricas de avaliação de desempenho das mesmas e formas de visualizar esses dados. O Capítulo 3 apresenta o método e materiais usados na experimentação assim como o modo de sua obtenção e manipulação. No Capítulo 4 os resultados são apresentados seguido pela sua análise, mostrando o desempenho de cada arquitetura, assim como os gráficos de desempenho de cada uma. O capítulo 5 será descrita as conclusões do trabalho e as possiveis novas abordagens para o futuro, logo após as analises dos resultados

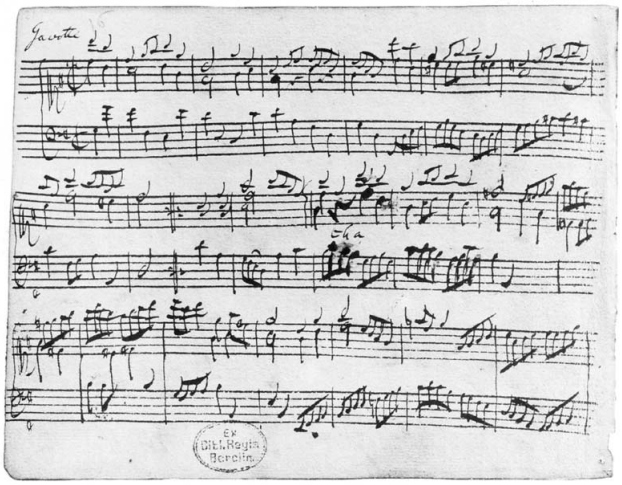
2 Fundamentação Teórica

Neste capítulo serão descritas as principais noções e características da notação musical e um resumo sobre seu aspecto histórico. Também será feita uma introdução a teoria das redes neurais em seus aspectos básicos, assim como uma abordagem teórica das três arquiteturas usadas no estudo.

2.1 Notação Musical

A notação musical, com o passar dos séculos foi evoluindo e se tornando mais complexa. No século XI, surgiu o sistema de notação mensural, que utilizava diferentes formas de notas para representar as diferentes durações musicais. Esse sistema trouxe uma maior clareza e precisão na representação das relações de tempo na música, permitindo uma notação mais refinada e detalhada (HOPPIN, 1978).

Figura 1 – Um exemplo de uma partitura manuscrita.



Fonte: (BACH,).

No século XVII, o sistema de notação moderno foi estabelecido, com o uso de uma pauta com cinco linhas e a adoção dos símbolos musicais que ainda são utilizados atualmente (Figura 1). Foi nesse período que os símbolos musicais como as notas em formato de elipses, as claves, as pausas e as figuras rítmicas ganharam a forma que conhecemos hoje. Essa padronização da notação musical facilitou a comunicação e o ensino da música, permitindo que músicos e compositores compartilhassem suas criações de forma mais precisa e compreensível (APEL, 1997).

No século XVIII, o surgimento dos sinais de dinâmica, como o piano (suave) e o forte (f), trouxe uma nova dimensão à notação musical, permitindo que os compositores

indicassem a intensidade desejada para a execução da peça. Além disso, os sinais de expressão, como legato (ligado) e staccato (curto e destacado), forneceram instruções adicionais para a interpretação musical (ADLER, 2002).

No século XIX, o desenvolvimento da notação musical expandiu-se ainda mais com a introdução de símbolos para indicar técnicas especiais, como arco para instrumentos de cordas, trinado e glissando. Essas adições enriqueceram a capacidade de representar a música de forma mais precisa e detalhada, permitindo que os músicos capturassem as nuances e intenções dos compositores (READ, 1979).

Atualmente, a notação musical continua a evoluir, incorporando símbolos e convenções para representar uma ampla variedade de técnicas contemporâneas, como música experimental, efeitos eletrônicos e improvisação. Além disso, a tecnologia desempenha um papel importante na notação musical, com o desenvolvimento de softwares e aplicativos que facilitam a criação, edição e reprodução de partituras musicais digitalmente.

A notação musical moderna ocidental continua sendo utilizada como uma ferramenta essencial para a escrita, leitura e interpretação da música. Ela desempenha um papel fundamental no ensino, na prática musical e na preservação do repertório musical ao longo dos séculos. Além disso, a notação musical evoluiu juntamente com a tecnologia, permitindo a incorporação de elementos adicionais, como dinâmicas, expressões e indicações de articulação, que enriquecem a interpretação e a compreensão das obras musicais. Assim, a notação musical continua a ser uma forma indispensável de registro e transmissão da música, contribuindo para a preservação e a apreciação da rica herança musical da humanidade.

2.2 Redes Neurais

As redes neurais são modelos computacionais inspirados pelo funcionamento do cérebro humano, capazes de aprender e realizar tarefas complexas a partir de dados. Ao longo das décadas, esses modelos têm sido amplamente estudados e aplicados em diversas áreas, desde o processamento de imagens até o reconhecimento de fala e a tomada de decisões.

As redes neurais são modelos computacionais inspirados pelo funcionamento do cérebro humano, capazes de aprender e realizar tarefas complexas a partir de dados. Ao longo das décadas, esses modelos têm sido amplamente estudados e aplicados em diversas áreas, desde o processamento de imagens até o reconhecimento de fala e a tomada de decisões. Em tarefas de processamento de imagens, as redes neurais convolucionais se destacam, permitindo a identificação de objetos, reconhecimento facial e geração de imagens realistas. Além disso, as redes neurais são aplicadas com sucesso no processamento de linguagem natural, possibilitando a tradução automática, análise de sentimentos e

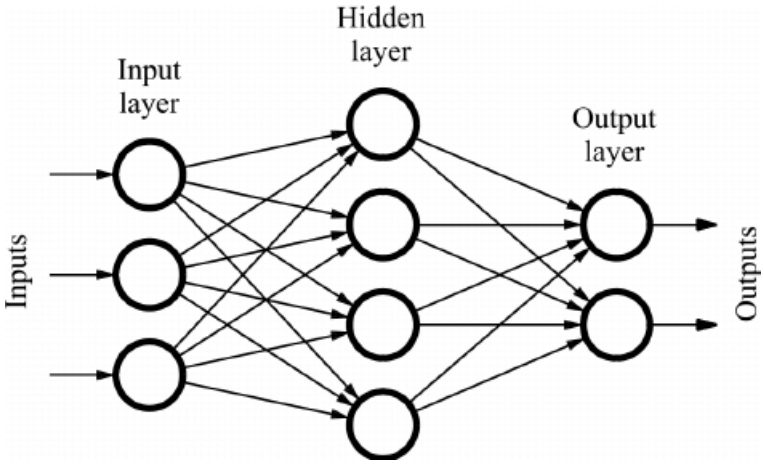
chatbots. Elas desempenham um papel crucial em avanços como a inteligência artificial e a aprendizagem profunda, alcançando resultados impressionantes em áreas como reconhecimento de voz, diagnóstico médico e previsão de demanda. O campo das redes neurais continua em constante evolução, com novas arquiteturas e algoritmos surgindo regularmente, impulsionando ainda mais a pesquisa e o desenvolvimento nessa área fascinante (NIELSEN, 2015).

Suponha que temos um conjunto de dados com informações sobre três tipos de flores: rosas, lírios e tulipas. Cada flor é descrita por dois atributos: comprimento das pétalas e largura das pétalas. Um exemplo simples que pode ser resolvido por uma rede neural, com arquitetura básica (Figura 2), que possui uma input layer, um hidden layer e um output layer é a classificação de flores com base em suas características.

Um subtipo de redes neurais projetado para o processamento de imagens são as redes neurais convolucionais (GOODFELLOW; BENGIO; COURVILLE, 2016). Estas são amplamente reconhecidas por sua capacidade de processar e extrair informações de imagens eficientemente. Essas arquiteturas conseguem identificar objetos, reconhecer rostos e até mesmo gerar imagens realistas com base nos padrões aprendidos durante o treinamento. Elas têm se mostrado extremamente úteis em diversas áreas, como na condução de carros autônomos, onde a capacidade de detectar e reconhecer objetos é essencial para a segurança e o desempenho do veículo. Além disso, as redes neurais convolucionais são aplicadas com sucesso no campo da medicina, auxiliando no diagnóstico médico assistido por computador. Essas redes podem analisar imagens médicas, como radiografias e tomografias, e identificar padrões que indiquem a presença de doenças ou anomalias. Com sua capacidade de processar grandes volumes de dados de forma rápida e precisa, as redes neurais convolucionais têm revolucionado como lidamos com imagens e mostram um potencial significativo para aplicações futuras em diversas áreas.

O sucesso das redes neurais se deve, em parte, aos avanços na disponibilidade de dados e no desenvolvimento de algoritmos de treinamento mais eficientes. As redes neurais podem aprender a partir de grandes volumes de dados, ajustando os pesos das conexões entre os neurônios conforme os exemplos apresentados. Com isso, elas conseguem realizar tarefas complexas e obter resultados precisos em diferentes domínios. A pesquisa e o desenvolvimento em redes neurais continuam avançando rapidamente, com novas descobertas e aplicações emergindo constantemente. À medida que essas tecnologias se aprimoram, novas possibilidades surgem, impulsionando avanços em áreas como inteligência artificial, aprendizado de máquina e ciência de dados. As redes neurais estão se tornando cada vez mais indispensáveis em nossa sociedade, impactando desde a medicina até a indústria, e sua influência continuará a crescer à medida que exploramos seu potencial em um mundo cada vez mais orientado por dados (BISHOP, 2006).

Figura 2 – Um exemplo de rede neural simples.



O objetivo é treinar uma rede neural para aprender a classificar corretamente as flores com base nesses atributos. A rede neural teria um input layer com dois neurônios correspondentes aos atributos de comprimento e largura das pétalas. Em seguida, teríamos um hidden layer com um número variável de neurônios que podem ser ajustados durante o treinamento. Por fim, teríamos um output layer com três neurônios, representando as três classes possíveis: rosa, lírio e tulipa.

Durante o treinamento, alimentaríamos a rede neural com as informações das flores do conjunto de treinamento, ajustando os pesos das conexões entre os neurônios para minimizar a diferença entre as saídas esperadas (a classe correta da flor) e as saídas preditas pela rede neural. Esse processo é repetido várias vezes até que a rede neural possa fazer previsões precisas para novas flores.

Após o treinamento, a rede neural estaria pronta para receber as informações de comprimento e largura das pétalas de uma flor desconhecida e prever sua classe correta. Por exemplo, se fornecermos as informações de uma flor com comprimento de pétala 5.0 e largura de pétala 2.5, a rede neural pode prever que se trata de uma rosa.

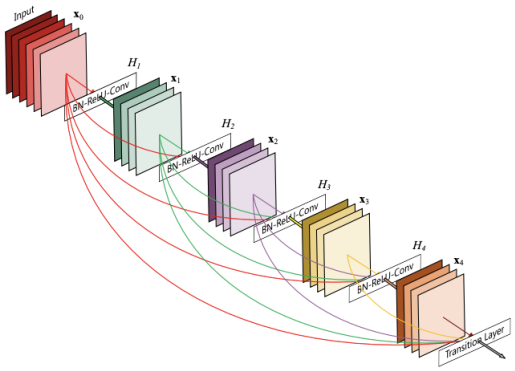
Esse exemplo simples ilustra o processo básico de uma rede neural com um input layer, um hidden layer e um output layer na resolução de um problema de classificação. É importante destacar que, na prática, problemas mais complexos podem exigir redes neurais com arquiteturas mais elaboradas e técnicas adicionais, mas esse exemplo fornece uma visão geral da teoria das redes neurais e como elas podem ser aplicadas.

2.2.1 DenseNet

A DenseNet (HUANG et al., 2017) (Figura 3) é uma arquitetura de rede neural convolucional que se destaca por sua abordagem inovadora na conexão entre as camadas. O objetivo principal da DenseNet é resolver o problema do desvanecimento do gradiente e promover um fluxo eficiente de informações durante o treinamento da rede. Para isso, a

arquitetura DenseNet introduz conexões densas entre todas as camadas, permitindo que cada camada receba informações diretas de todas as camadas anteriores. Essa conexão densa promove a reutilização e o compartilhamento de características, resultando em representações mais ricas e robustas.

Figura 3 – Arquitetura Densenet



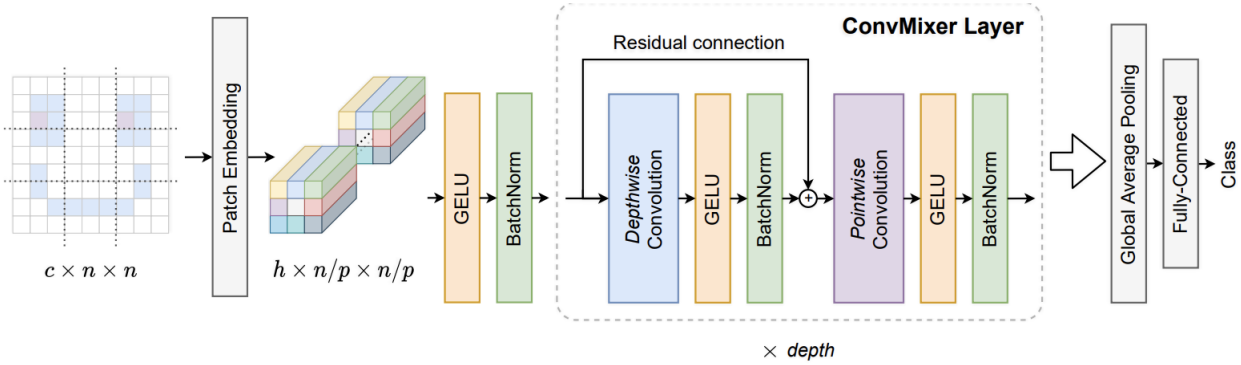
Fonte : (HUANG et al., 2017)

Uma das particularidades da DenseNet é a sua estrutura em blocos densos. Cada bloco é composto por camadas de convolução em sequência, seguidas por uma conexão densa que combina as saídas de todas as camadas anteriores. Essa estrutura promove a propagação direta do gradiente, permitindo que os gradientes sejam transmitidos de maneira mais eficiente durante o treinamento. Além disso, a conexão densa ajuda a mitigar o desaparecimento do gradiente, permitindo que informações importantes sejam preservadas em todas as camadas.

Outra particularidade da DenseNet é a sua eficiência em termos de parâmetros e uso de memória. Devido à conexão densa, a quantidade de parâmetros na DenseNet é significativamente reduzida em comparação com outras arquiteturas convolucionais tradicionais. Essa redução de parâmetros é especialmente benéfica em cenários com recursos computacionais limitados, permitindo o treinamento de modelos mais profundos e complexos sem aumentar excessivamente os requisitos de memória.

A DenseNet tem sido amplamente utilizada e mostrou resultados promissores em tarefas de visão computacional, como classificação de imagens, segmentação e detecção de objetos. Sua arquitetura densa e eficiente na propagação de informações tem sido reconhecida como uma solução eficaz para lidar com o desvanecimento do gradiente e melhorar o desempenho das redes neurais convolucionais. À medida que a pesquisa e o desenvolvimento avançam, é provável que a DenseNet continue sendo uma escolha popular em aplicações de visão computacional, impulsionando avanços adicionais nessa área.

Figura 4 – Arquitetura ConvMixer



Fonte : (NG et al., 2022)

2.2.2 ConvMixer

A ConvMixer (NG et al., 2022) (Figura 4) tem como ideia central aplicar convoluções em diferentes resoluções espaciais para capturar padrões locais e globais em uma imagem. Em vez de usar camadas convolucionais seguidas por camadas totalmente conectadas, o ConvMixer usa apenas camadas convolucionais em cascata. Isso permite que a rede aprenda representações complexas sem a necessidade de camadas totalmente conectadas, tornando o modelo mais leve e eficiente em termos computacionais.

A arquitetura ConvMixer consiste em uma série de blocos repetidos para criar uma rede neural profunda. Cada bloco do ConvMixer é composto por duas camadas principais: uma camada de convolução em patches e uma camada de transformação linear. A combinação dessas camadas permite que o modelo capture informações contextuais em diferentes escalas e distâncias.

- Camada de convolução em patches: A entrada para o bloco é um tensor que representa um conjunto de patches extraídos de uma imagem de entrada. A camada de convolução em patches é aplicada a cada patch individualmente, tratando-os como "palavras" em uma sequência.Essa camada de convolução em patches é responsável por capturar informações locais e criar representações de características mais ricas para cada patch. Normalmente, a camada de convolução em patches é composta por uma convolução 1D ou 2D seguida de uma função de ativação, como ReLU.

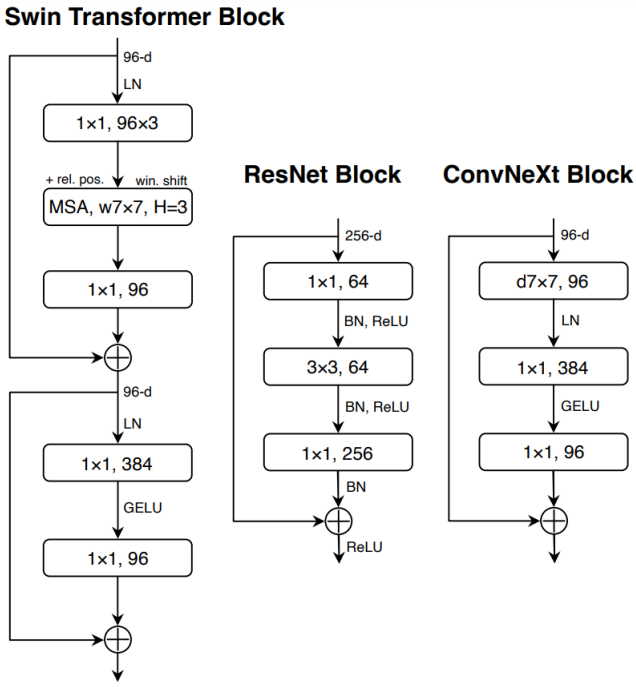
- Camada de transformação linear: Após a camada de convolução em patches, os patches convolucionais são linearmente transformados para capturar informações contextuais em escalas maiores. Essa camada de transformação linear é responsável por capturar informações globais e criar representações de características de escopo mais amplo. A transformação linear é geralmente realizada por meio de uma convolução 1x1 ou uma operação de projeção linear. Uma função de ativação, como

ReLU, pode ser aplicada após a transformação linear para introduzir não-linearidade na representação. Esses dois componentes principais (camada de convolução em patches e camada de transformação linear) são repetidos em cada bloco do ConvMixer. Além disso, técnicas como camadas de normalização, atenção ou pooling podem ser adicionadas entre esses componentes para melhorar o desempenho e a capacidade de representação do modelo.

A ConvMixer tem se mostrado eficaz em uma variedade de tarefas de visão computacional, como classificação de imagens, segmentação e detecção de objetos. Sua arquitetura flexível e eficiente torna o ConvMixer uma escolha atraente para problemas onde a captura de relacionamentos espaciais e o uso eficiente de recursos computacionais são essenciais. À medida que a pesquisa e o desenvolvimento avançam, é provável que o ConvMixer continue evoluindo e sendo aplicado em várias aplicações de processamento de imagem e visão computacional.

2.2.3 ConvNext

Figura 5 – Arquitetura ConvNext



Fonte : (LIU et al., 2022)

A arquitetura ConvNeXt (LIU et al., 2022) (Figura 5) é uma arquitetura de rede neural desenvolvida com base em convoluções convencionais (ConvNets) e inspirada nas vantagens dos Transformers. O objetivo era criar uma arquitetura de rede que pudesse competir com as redes Transformers em termos de desempenho e escalabilidade, mas mantendo a simplicidade e eficiência das ConvNets.

A arquitetura ConvNeXt segue uma trajetória de "modernização de uma ResNet (uma ConvNet padrão) em direção a uma arquitetura de Transformer hierárquica, como a Swin Transformer. Ao longo desse processo, várias modificações são feitas na arquitetura original da ResNet para incorporar elementos e *design choices* (ou escolhas de design, em português) encontrados nas arquiteturas dos Transformers.

Principais blocos e modificações da ConvNeXt:

- Stem Patchify: Substituição do bloco inicial da ResNet por um bloco "Patchify" semelhante ao usado nas arquiteturas Transformers. Esse bloco divide a imagem de entrada em patches e os processa separadamente.
- Stage Compute Ratio: A distribuição de computação entre as etapas da rede é modificada para se adequar à arquitetura Swin Transformer. Isso envolve ajustar o número de blocos em cada estágio e sua complexidade computacional.
- Inverted Bottleneck: Introdução de um bloco de "bottleneck invertido" inspirado nas arquiteturas ResNeXt e MobileNetV2. Esse bloco tem uma dimensão oculta quatro vezes maior que a dimensão de entrada, permitindo uma maior representação de recursos.
- Large Kernel Sizes: Exploração do uso de tamanhos de kernel maiores do que o padrão de 3x3 usado nas ConvNets. Isso envolve o deslocamento de uma camada convolucional em cada bloco para permitir o uso de convoluções com tamanhos maiores, como 7x7.
- GELU Activation Function: Substituição da função de ativação ReLU pela função GELU (Gaussian Error Linear Unit), que é comumente usada nas arquiteturas Transformers.
- Layer Normalization: Substituição do Batch Normalization (BN) pela Layer Normalization (LN), uma técnica de normalização usada nas arquiteturas Transformers.
- Separate Downsampling Layers: Introdução de camadas separadas de downsampling (redução da resolução espacial) entre as etapas da rede para melhorar o desempenho e a estabilidade do treinamento.

Obs: MSA - Em um bloco Swin Transformer, "MSA"significa "Multi-Scale Attention"(Atenção Multi-Escala) e é um componente-chave para capturar relacionamentos de longo alcance entre diferentes partes de uma imagem.

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Acurácia

$$\frac{TP}{TP + FP} \tag{2.2}$$

Precisão

$$\frac{TP}{TP + FN} \tag{2.3}$$

Recall

2.3 Métricas de Avaliação

As métricas de avaliação são ferramentas essenciais para medir o desempenho de redes neurais e determinar a eficácia de seus modelos. Existem várias métricas comumente utilizadas para avaliar diferentes aspectos de problemas de classificação como precisão (Equação 2.2), recall (Equação 2.3), acurácia (Equação 2.1), F1-score (Equação 2.4). Essas métricas fornecem uma visão geral do quão bem o modelo está realizando as previsões e ajudam a identificar possíveis áreas de melhoria.

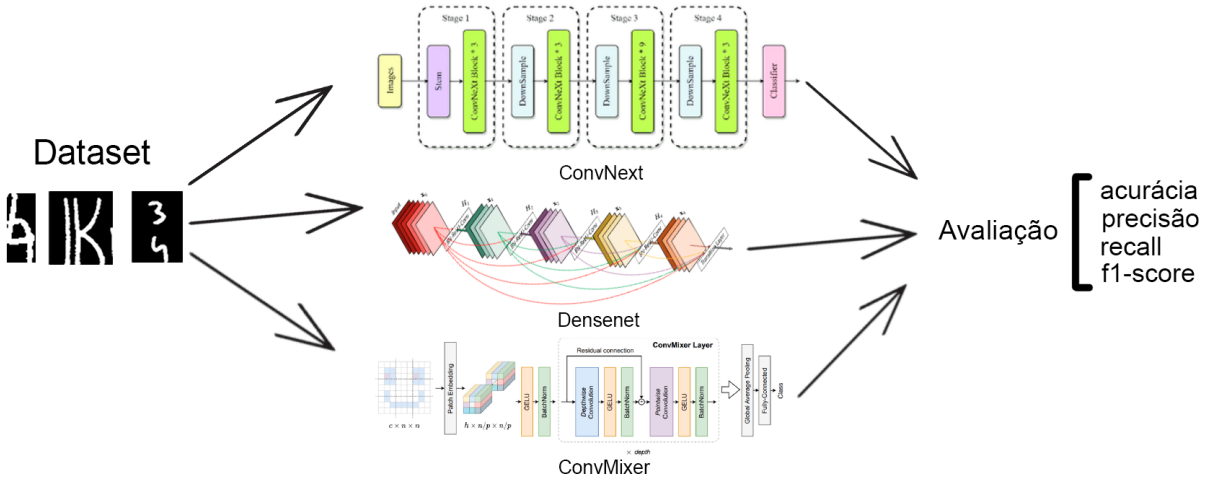
$$\frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.4}$$

F1-score

3 Materiais e Método

A metodologia adotada neste trabalho, apresentada pela Figura 6, consiste em utilizar um conjunto de dados de imagens de símbolos musicais manuscritas para treinar e avaliar três arquiteturas de redes neurais. O objetivo principal é analisar e comparar o desempenho dessas arquiteturas na tarefa de classificação das notas musicais presentes nas imagens.

Figura 6 – Etapas da metodologia proposta



Foram selecionadas três arquiteturas de redes neurais recentes e com bom desempenho em tarefas de classificação: ConvNetXt, DenseNet e ConvMixer. Cada uma dessas arquiteturas possui características e abordagens distintas para extrair e processar informações das imagens.

Na etapa de treinamento, cada arquitetura foi alimentada com o mesmo conjunto de dados e ajustadas com hiperparametros com os melhores desempenhos para o modelo. Após o treinamento, as arquiteturas foram submetidas a uma fase de avaliação utilizando um conjunto separado de dados de teste. Esse conjunto de teste contém imagens de símbolos musicais não vistas anteriormente pelas redes neurais durante o treinamento. Dessa forma, é possível avaliar a capacidade de generalização e o desempenho das arquiteturas em dados desconhecidos. Todo o processo é explicado nas seções subsequentes.

3.1 Aquisição dos Dados

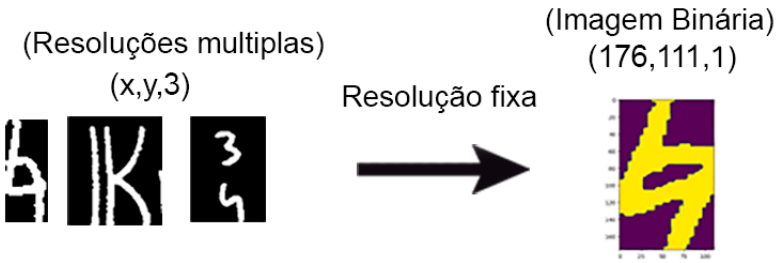
Foi utilizado o banco de dados HOMUS (HOMUS. . . , 2016). Na utilização desse banco de dados foram selecionadas 4 partituras de cada músico, então foi utilizada uma

segmentação manual para extração de 4000 símbolos musicais em 20 classes. A Tabela 1 apresenta alguns exemplos das imagens e classes da base. Cada classe, na base, conta com 200 imagens cada.

3.2 Pré-processamento

Na etapa de pré-processamento das imagens para a classificação de símbolos musicais, foi necessário realizar algumas transformações para garantir a consistência e eficiência no processamento dos dados. Um dos passos fundamentais foi redimensionar todas as imagens para uma dimensão específica. Nesse caso, optou-se por um tamanho de 176 pixels de altura por 111 pixels de largura (Exemplo na Figura 7). Esses valores foram escolhidos com base na maior imagem em termos de altura e largura encontrada no conjunto de dados, com isso as imagens são "esticadas"até a resolução desejada caso sejam menores. Dessa forma, todas as imagens foram ajustadas para um formato padronizado, facilitando o processamento posterior.

Figura 7 – Pré processamento das imagens.



Fonte: Acervo do autor

Além do redimensionamento, também foi realizada uma modificação nos canais de cores das imagens. Originalmente, as imagens possuíam três canais de cores (RGB). No entanto, para economizar processamento e simplificar a análise, optou-se por converter as imagens para tons de cinza. Essa conversão para escala de cinza reduz a quantidade de informações a serem processadas, sem afetar significativamente a capacidade de distinguir as notas musicais.

Além disso, os pixels das imagens foram convertidos para o formato binário. Essa conversão é adequada para o contexto em que o fundo da imagem é completamente preto e a figura que representa a nota musical é completamente branca. Ao converter os pixels para o formato binário, os valores dos pixels são ajustados para valores extremos (preto ou branco) com base em um determinado limiar. Essa abordagem simplifica a análise dos padrões e características da nota musical, uma vez que os pixels são representados por valores binários que indicam a presença ou ausência da nota em cada região da imagem.

Tabela 1 – Tabela das Classes

Digital	Manuscrita	Nome	Classe
		Diminuendo	1
		Barra	2
		Bemol	3
		Semínima	4
		Barra	5
		Semibreve	6
		Clave de Sol	7
		Pausa de Semínima	8
		Mínima	9
		Sustenido	10
		Tempo	11
		Semicolcheia	12
		Pausa de Colcheia	13
		Clave de Dó	14
		Clave de Fá	15
		Crescendo	16
		Cabeça da Nota	17
		Bequadro	18
		Duas Semicolcheias Ligadas	19
		Duas Colcheias Ligadas	20

Essas transformações no pré-processamento das imagens visam garantir que todas as imagens tenham as mesmas dimensões e estejam em um formato adequado para a extração de características e a classificação das notas musicais. Ao padronizar o tamanho, converter para escala de cinza e formato binário, é possível simplificar o processamento das imagens e facilitar a detecção dos padrões relevantes para a classificação. Essas etapas são importantes para preparar os dados de entrada e obter um desempenho mais eficaz dos modelos de classificação aplicados posteriormente.

3.3 Construção do modelo

Para contribuir no problema de classificação de notas musicais por meio de visão computacional, foram adotadas três arquiteturas de redes neurais: Convmixer, Densenet e Convnext. A escolha dessas arquiteturas é devido ao ótimo desempenho em problemas de classificação em múltiplas classes e pela suas vantagens estarem ligadas principalmente ao aproveitamento e reconhecimento aprimorado de características de cada classe.

A ConvMixer foi escolhida pelo uso de convolução para capturar a informação espacial das imagens e as operações lineares para capturar informações de frequência mais alta, mas ao invés de utilizar camadas convolucionais tradicionais, ela utiliza camadas convolucionais em um formato misto, combinando operações de convolução 1x1 e convolução linear. Essa abordagem permite capturar informações contextuais e relacionamentos espaciais entre as características presentes nos dados de entrada, resultando em representações mais ricas e significativas. A DenseNet foi escolhida pelas suas conexões densas entre todas as camadas, permitindo que cada camada receba informações diretas de todas as camadas anteriores. Essa conexão densa promove a reutilização e o compartilhamento de características, resultando em representações mais ricas e robustas. A ConvNext foi escolhida pela capacidade de representação e a capacidade discriminativa da rede, permitindo melhores resultados em tarefas de classificação de imagem. Ela utiliza vias (paths) que controlam a diversidade e a complexidade das interações entre os canais de entrada.

A arquitetura Convmixer, com aproximadamente 1,2 milhões de parâmetros, foi configurada com uma profundidade de 16 camadas, um tamanho de kernel de 5 e um tamanho de patch de 2. Também foi usado um learning rate de 0.005, batch size de 16 e 10 épocas. No modelo DenseNet, com cerca de 350 mil parâmetros, foi utilizado um valor de "growth rate"(taxa de crescimento) de 8, um learning rate de 0.005, batch size de 16 e 10 épocas.

Já no modelo Convnext, também com aproximadamente 1,2 milhões de parâmetros, foram utilizadas configurações específicas, como tamanho de filtro, tamanho de pooling e número de camadas. Esses parâmetros foram ajustados empiricamente para otimizar o

desempenho do modelo na tarefa de classificação de notas musicais. Ainda foram usados um learning rate de 0.005, batch size de 32 e 20 épocas.

A escolha desses hiperparametros em comum para as duas das arquiteturas, convmixer e densenet, é devido a melhor adaptação dos pesos ao learning rate em questão. Nos experimentos, com um learning rate maior, as arquiteturas entravam em vales e tinha dificuldade na generalização das características. Com learning rates menores, as arquiteturas também tinha dificuldades em encontrar os mínimos globais e seu tempo de treinamento tinha que ser aumentado drasticamente. O batch size de 16 se deve aos modelos se adaptarem bem a esse tamanho de lote, e 10 épocas devido aos pesos estarem bem ajustados nessa época. Na ConvNext o aprendizado se mostrou mais lento, devido isso foi necessário um batch size e um número de épocas maior. O tamanho do batch size foi de 32 e o treinamento foi de 20 épocas.

Em resumo, a construção do modelo envolveu a implementação de três arquiteturas de redes neurais, cada uma com suas próprias configurações e parâmetros. Os conjuntos de treinamento, validação e teste foram definidos para garantir uma avaliação adequada do desempenho do modelo. Os parâmetros específicos de cada arquitetura foram selecionados com base em experimentações e ajustes para obter o melhor desempenho possível na classificação de notas musicais. Essas escolhas foram feitas considerando características das arquiteturas, como a capacidade de aprendizado profundo, eficiência e capacidade de processar informações visuais.

3.4 Treinamento e Avaliação

O conjunto de dados utilizado consiste em 20 classes, totalizando 4000 imagens. Essas imagens foram divididas em três subconjuntos: treinamento, validação e teste.

No conjunto de treinamento, usado para atualizar os pesos das conexões durante o treinamento, foi reservado 70% das imagens disponíveis, onde foi feita a divisão estratificada por classe. O conjunto de validação, composto por 20% das imagens, foi utilizado para ajustar os parâmetros do modelo e monitorar o desempenho durante o treinamento. Já o conjunto de teste, com 10% das imagens, foi utilizado para avaliar o desempenho final do modelo em dados não vistos. Os modelos foram avalidos com base na acurácia, precisão, recall e f1-score

3.5 Ambiente de Experimentação

O ambiente utilizado para manipulação da base de dados e treinamento das arquiteturas foi o google colab, um ambiente de programação em nuvem voltado a ciência

de dados e inteligência artificial. No plano gratuito é ofertado uma GPU T4, suficiente para o nosso problema em questão.

As bibliotecas utilizadas foram as: OpenCV (Disponivel em: <https://opencv.org/>) para pré-processamento das imagens, Keras (Disponivel em: <https://keras.io/>) para a criação das arquiteturas, matplotlib para plot dos gráficos e tabelas, e numpy para a processo de manipulação numérica.

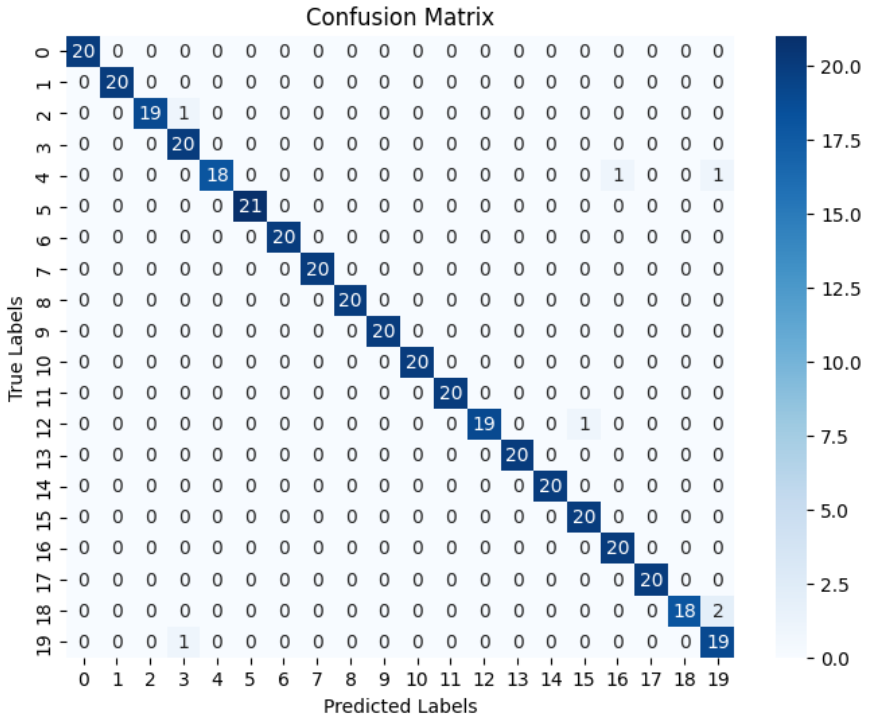
Todo o processo de manipulação da base de dados, criação das arquiteturas e treinamento foi feito no ambiente colab.

4 Resultados

A base de dados foi inicialmente pré-processada. Em seguida, os modelos configurados conforme descrito, e o processo de treinamento foi iniciado. No processo de treinamento dos modelos, foram adotadas diferentes parametrizações no ambiente do Google Colab, utilizando uma GPU T4.

No experimento utilizando a arquitetura ConvMixer, observou-se uma curva de aprendizado que demonstrou um rápido aumento no desempenho do modelo durante as primeiras épocas de treinamento. A Figura 8 apresenta a matriz de confusão dos resultados.

Figura 8 – Matriz de confusão para a convMixer



A Figura 9 apresenta as curvas loss. Observamos que a partir da terceira época, o ganho de desempenho diminuiu, indicando que o modelo estava convergindo para um resultado próximo do ótimo. A curva de aprendizado apresentou uma tendência de estabilização, com pequenas oscilações no desempenho ao longo das épocas. Ao final do treinamento, o modelo alcançou resultados promissores durante o teste. A métrica de acurácia atingiu um valor de 0,9651, indicando que o modelo conseguiu classificar corretamente 96,51 por cento das amostras de teste. Além disso, as métricas de precisão e recall obtiveram valores de 0,9675% e 0,9651%, respectivamente. Isso demonstra a capacidade do modelo de realizar corretamente as previsões positivas (precisão) e de recuperar corretamente as instâncias positivas (recall). Esses resultados Figura 8 sugerem

que a arquitetura ConvMixer conseguiu aprender efetivamente as características relevantes das notas musicais e realizar uma classificação precisa.

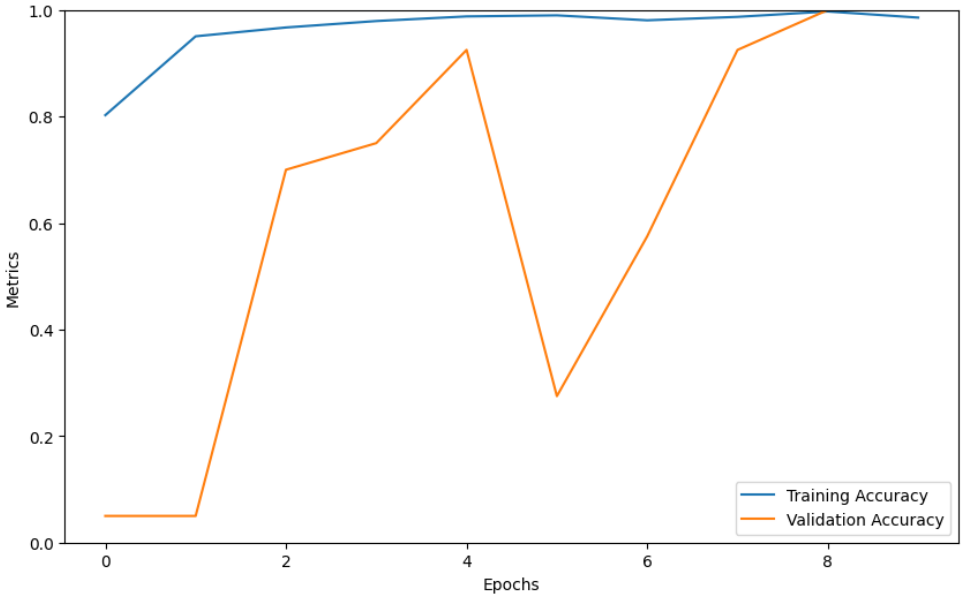


Figura 9 – Curva de perda e validação (ConvMixer)

No experimento utilizando a arquitetura DenseNet, observou-se uma curva de aprendizado que demonstrou um aumento gradual e consistente no desempenho do modelo ao longo das épocas de treinamento (Figura 11). A curva de aprendizado exibiu uma tendência positiva, com a perda de treinamento diminuindo progressivamente e a acurácia aumentando. Ao final do treinamento, o modelo obteve resultados promissores durante o teste. A métrica de acurácia alcançou um valor de 0,9576%, indicando que o modelo pôde classificar corretamente 95,76% das amostras de teste. Além disso, as métricas de precisão e recall obtiveram valores de 0,9575% e 0,9551%, respectivamente. Isso indica a capacidade do modelo de realizar corretamente as previsões positivas (precisão) e de recuperar corretamente as instâncias positivas (recall). A curva de aprendizado consistente indica que o modelo foi capaz de generalizar bem os padrões presentes nos dados de treinamento para realizar previsões precisas nos dados de teste.

No experimento utilizando a arquitetura ConvNext, observou-se uma curva de aprendizado que demonstrou um aumento gradual e consistente no desempenho do modelo ao longo das épocas de treinamento (Figura 12). A curva de aprendizado exibiu uma tendência positiva, com a perda de treinamento diminuindo progressivamente e a acurácia aumentando. Ao final do treinamento, o modelo obteve resultados promissores durante o teste.

A métrica de acurácia alcançou um valor de 0,9501, indicando que o modelo. Além disso, as métricas de precisão e recall obtiveram valores de 0,9616 e 0,9377, respectivamente. Isso indica a capacidade do modelo de realizar corretamente as previsões positivas (precisão)

Figura 10 – Matriz de confusão para a DenseNet

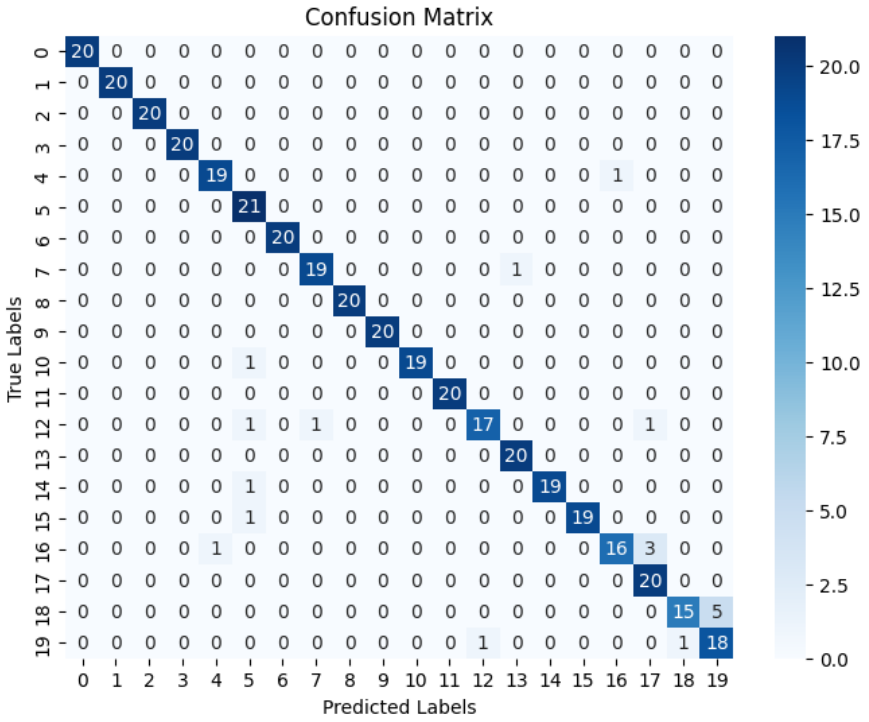
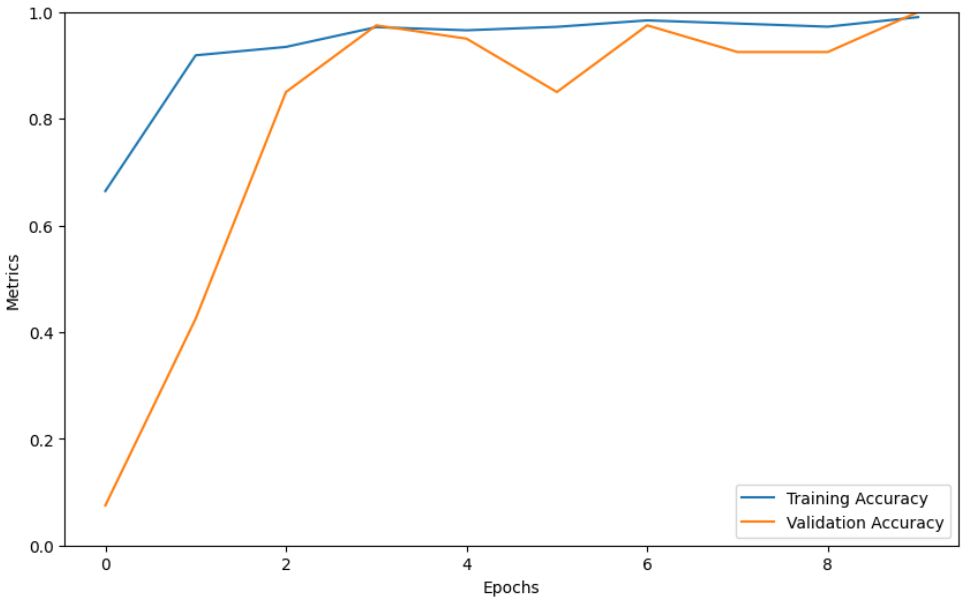


Figura 11 – Curva de perda e validação (DenseNet)



e de recuperar corretamente as instâncias positivas (recall). Esses resultados sugerem que a arquitetura ConvNext foi capaz de aprender efetivamente as características das notas musicais e realizar uma classificação precisa. A curva de aprendizado consistente indica que o modelo conseguiu generalizar bem os padrões presentes nos dados de treinamento para realizar previsões precisas nos dados de teste.

O desempenho das três arquiteturas (Tabela 2) pode ser comparado com base nas métricas obtidas durante o treinamento e validação. No caso do ConvMixer, observamos

Figura 12 – Curva de perda e validação (ConvNext)

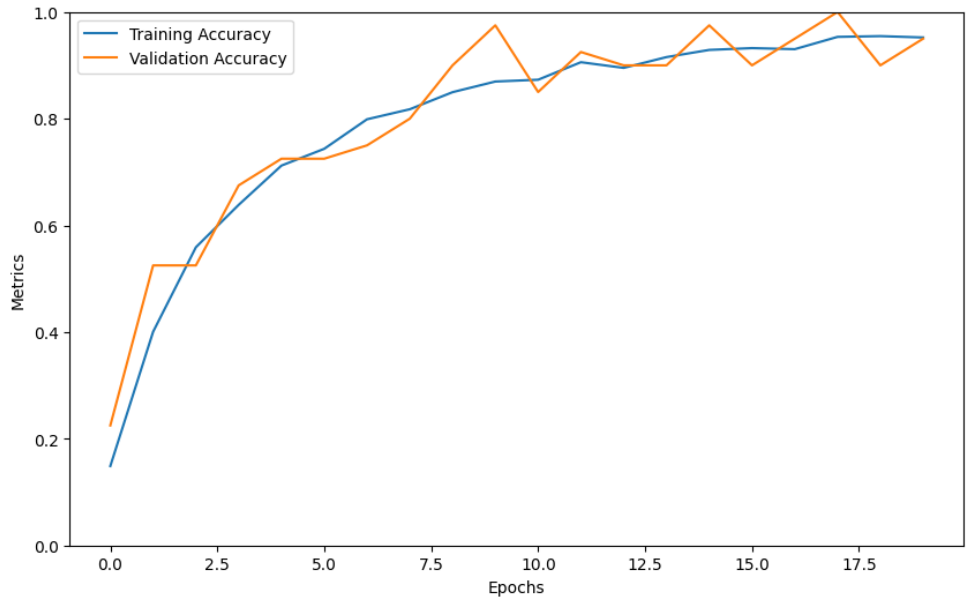
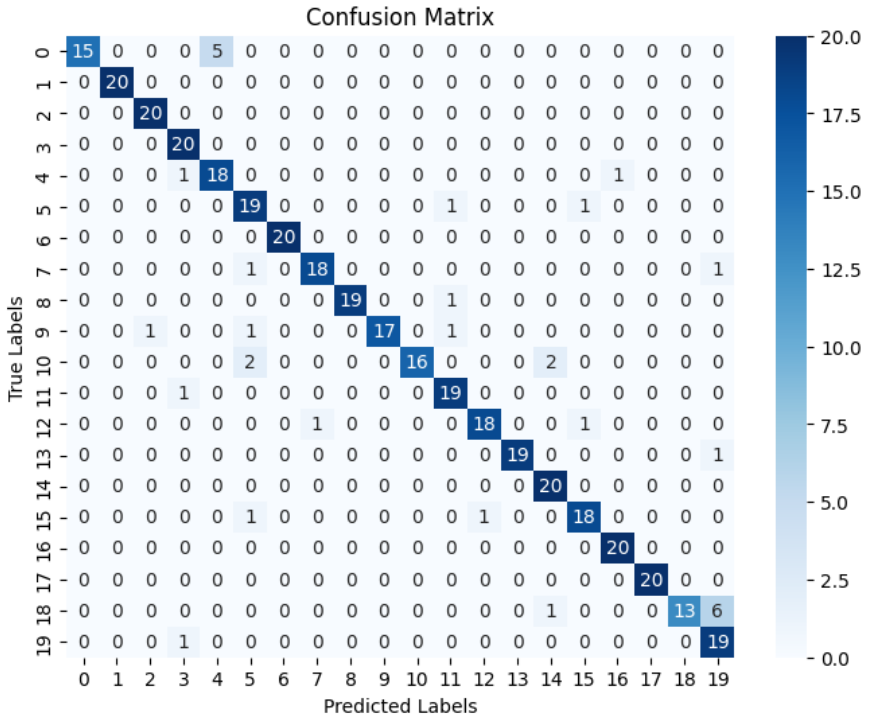


Figura 13 – Matriz de confusão para a ConvNext



um bom desempenho com alta precisão e recall nas primeiras épocas. No entanto, conforme o treinamento progride, a precisão e recall diminuem na validação, indicando um possível caso de overfitting. A acurácia se mantém alta, mas a perda aumenta, sugerindo uma instabilidade do modelo em relação à generalização. Para o Densenet, também observamos um bom desempenho, com alta precisão e recall desde as primeiras épocas. A acurácia e a perda também mostram resultados satisfatórios, com uma tendência de melhoria ao longo das épocas. Esses resultados indicam que o modelo está aprendendo bem os dados

de treinamento e consegue generalizar para os dados de validação. Quanto ao ConvNext, o desempenho melhora gradualmente ao longo das épocas. A acurácia, precisão e recall apresentam uma tendência positiva, e a perda diminui. Isso sugere que o modelo está aprendendo efetivamente os dados de treinamento e consegue generalizar para os dados de validação. Em resumo, enquanto o ConvMixer mostra sinais de overfitting e instabilidade nos resultados, o Densenet e o ConvNext apresentam um desempenho mais consistente e melhor capacidade de generalização. Ambas as arquiteturas demonstram boa acurácia, precisão e recall.

Na métrica de comparação F1 a convmixer obteve o melhor desempenho, tendo erros pontuais em classes como a 4 e 18, mas tendo uma elevada taxa de acerto nas outras classes, beirando os 100%.A DenseNet teve erros principalmente na classe 18,tendo uma taxa de acerto de apenas 75%, na Convnext tbm teve uma alta taxa de erro na classe 18 tendo apenas 65% de acerto nessa classe.

Ao observar a matriz de confusão das três arquiteturas pode-se observar que a classe 18 foi o maior fator de erro em comum, depois a classe 4. Uma possivel melhora no desempenho nessas classes poderia ser tido usar outras técnicas de divisão de batchs e/ou treinamento específicos dessas classes, alem de tratamento no próprio dataset de treinamento.

Tabela 2 – Resultados das redes ConvMixer, DenseNet e ConvNext

Arquitetura	**Acuracia**	**Precisão**	**Recall**	**F1**
ConvMixer	0,9825	0,9838	0,9824	0,9830
DenseNet	0,9526	0,9565	0,9524	0,9544
ConvNext	0,942	0,9466	0,9430	0,9447

5 Conclusão

Neste trabalho, foram avaliadas diferentes arquiteturas de redes neurais convolucionais para a tarefa de classificação de notas musicais. Utilizamos a biblioteca Keras e a base de dados HOMUS para realizar os experimentos.

No processo de avaliação, foram implementadas e treinadas três arquiteturas diferentes: ConvMixer, DenseNet e ConvNext. Essas arquiteturas foram selecionadas com base em estudos e pesquisas que relataram resultados consistentes em tarefas de classificação e processamento de imagens. Durante o treinamento, dividimos a base de dados em conjuntos de treinamento, validação e teste.

Ao analisar os resultados, observamos que todas as arquiteturas alcançaram desempenho promissor na classificação de notas musicais. A arquitetura ConvMixer obteve uma acurácia de 98,25%, a DenseNet alcançou 95,26%, e a ConvNext atingiu 91,77% de acurácia. Esses resultados indicam que os modelos conseguiram aprender efetivamente as características relevantes das notas musicais e realizar uma classificação precisa.

Em relação aos objetivos propostos, podemos concluir que os principais foram atingidos. Avaliamos as arquiteturas de redes neurais convolucionais propostas, realizamos o treinamento e a avaliação dos modelos, e obtivemos resultados promissores de desempenho. Esses resultados contribuem para a área de processamento de imagens e classificação de notas musicais, fornecendo insights sobre o uso de diferentes arquiteturas em tarefas relacionadas.

No entanto, também identificamos algumas limitações e oportunidades para trabalhos futuros. Durante o experimento, observamos indícios de overfitting na arquitetura ConvMixer, o que pode ser mitigado com técnicas de regularização e aumento de dados. Além disso, outros modelos e técnicas de pré-processamento podem ser explorados para melhorar ainda mais o desempenho na classificação de notas musicais.

Este trabalho demonstrou a eficácia de diferentes arquiteturas de redes neurais convolucionais na classificação de notas musicais. Os resultados obtidos mostram a capacidade dos modelos em aprender as características relevantes das imagens e realizar previsões precisas. Esses resultados contribuem para o avanço da área de processamento de imagens aplicado à música e fornecem uma boa referencia para pesquisas futuras nesse domínio.

Além dos resultados promissores alcançados nesta pesquisa, há vários benefícios e possíveis melhorias que podem ser explorados na abordagem de classificação de notas musicais utilizando redes neurais convolucionais. Uma das principais vantagens dessa

abordagem é a capacidade de generalização dos modelos treinados. Uma vez treinados em um conjunto diversificado de dados musicais, esses modelos podem ser aplicados a diferentes contextos musicais, como partituras de diferentes estilos e períodos históricos. Isso torna essa abordagem flexível e útil em várias aplicações, como reconhecimento automático de partituras, análise musical computacional e geração de música assistida por computador.

Além disso, existem várias melhorias que podem ser consideradas para aprimorar ainda mais a precisão e a eficiência desses modelos. Por exemplo, técnicas avançadas de pré-processamento de imagens podem ser exploradas para melhorar a qualidade das imagens de notas musicais antes da classificação. Além disso, estratégias de aumento de dados, como rotação, translação e deformação elástica, podem ser aplicadas para aumentar a robustez dos modelos em relação a variações de posição e escala nas imagens das notas. O uso de técnicas de transferência de aprendizado também pode ser explorado, permitindo que os modelos se beneficiem do conhecimento prévio adquirido em tarefas semelhantes de visão computacional.

Em suma, a abordagem de classificação de notas musicais utilizando redes neurais convolucionais apresenta uma série de benefícios e oferece oportunidades para melhorias futuras. Essa abordagem mostra-se promissora na aplicação de processamento de imagens em contextos musicais, com potencial para avançar nas áreas de reconhecimento automático de partituras, análise musical computacional e geração de música assistida por computador. Com a contínua pesquisa e desenvolvimento nesse campo, podemos esperar avanços significativos na precisão e eficiência dos modelos, abrindo caminho para aplicações ainda mais amplas e inovadoras no campo da música e tecnologia.

Referências

ADLER, S. *The Study of Orchestration*. [S.l.]: W. W. Norton & Company, 2002. Citado na página 17.

APEL, W. *Harvard Dictionary of Music*. [S.l.]: Harvard University Press, 1997. Citado na página 16.

BACH, J. S. *Gavotte from his French Suite No. 5*. [S.l.: s.n.]. Citado na página 16.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado na página 18.

CALVO-ZARAGOZA, J.; TOSELLI, A. H.; VIDAL, E. Handwritten music recognition for mensural notation with convolutional recurrent neural networks. *Pattern Recognition Letters*, v. 128, p. 115–121, 2019. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865519302338>. Citado 2 vezes nas páginas 12 e 13.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016. Citado na página 18.

HOMUS: Handwritten Optical Music Understanding System. [S.l.: s.n.], 2016. Dataset. Citado 2 vezes nas páginas 12 e 25.

HOPPIN, R. H. *Medieval Music*. [S.l.]: W.W. Norton Company, 1978. Citado na página 16.

HUANG, G.; LIU, Z.; MAATEN, L. van der; WEINBERGER, K. Q. *Densely Connected Convolutional Networks*. [S.l.: s.n.], 2017. Citado 2 vezes nas páginas 19 e 20.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER. Gradient-based learning applied to document recognition. proceedings of the ieee. In: . [S.l.: s.n.], 1998. Citado na página 11.

LI, C. A deep learning-based piano music notation recognition method. In: *Computational Intelligence and Neuroscience, vol. 2022, Article ID 2278683, 9 pages*. [S.l.: s.n.], 2022. Citado na página 13.

LIU, Z.; MAO, H.; WU, C.-Y.; FEICHTENHOFER, C.; DARRELL, T.; XIE, S. *A ConvNet for the 2020s*. [S.l.: s.n.], 2022. Citado na página 22.

MUSICNET. In: . [S.l.: s.n.], 2017. Citado na página 12.

NG, D.; CHEN, Y.; TIAN, B.; FU, Q.; CHNG, E. S. *ConvMixer: Feature Interactive Convolution with Curriculum Learning for Small Footprint and Noisy Far-field Keyword Spotting*. [S.l.: s.n.], 2022. Citado na página 21.

NIELSEN, M. *Neural Networks and Deep Learning*. San Francisco, CA: Deterministic AI, 2015. Citado na página 18.

PEREIRA, R. M. P.; MATOS, C. E.; JUNIOR, G. B.; ALMEIDA, J. a. D. de; PAIVA, A. C. de. A deep approach for handwritten musical symbols recognition. In: *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: Association for Computing Machinery, 2016. (Webmedia '16), p. 191–194. ISBN 9781450345125. Disponível em: <https://doi.org/10.1145/2976796.2988171>. Citado 2 vezes nas páginas 12 e 13.

READ, G. *Music Notation: A Manual of Modern Practice*. [S.l.]: Taplinger Publishing, 1979. Citado na página 17.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 1–9, 2015. Citado na página 12.