Classificação Multilabel de Instrumentos Musicais

Francieli M. de Carvalho[1], Jonas G. S. Júnior[2], Matheus C. F. Brakes[3], Victor G. Pimenta[3]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Av. Esperança, s/n - Chácaras de Recreio – 74690-900 – Goiânia – GO – Brasil

{francielimoreira, jonasjunior, brakes_fares, victorguerreiro}@discente.ufg.br

Resumo. *Este trabalho investiga a performance de diferentes abordagens para a classificação multilabel de instrumentos musicais utilizando o dataset IRMAS. Para isso, utilizamos o modelo Wav2vec2, pré-treinado no AudioSet e ajustado no dataset IRMAS. Além disso, experimentamos com o dataset Instrument-UFG e aplicamos técnicas de mixagem de áudio para criar exemplos multilabel. Nossos experimentos mostraram que o aumento de dados, através da combinação de diferentes datasets e da aplicação de técnicas de mixagem, resultou em melhorias significativas na acurácia e na redução da perda durante o treinamento e validação. No entanto, observamos que métricas de balanceamento, como o F1 Score, ainda necessitam de melhorias. Estes resultados destacam a aplicabilidade das nossas abordagens em áreas como identificação automática de instrumentos e recomendação musical.*

1. Introdução

Diversos trabalhos abordaram a classificação multilabel de instrumentos musicais no dataset IRMAS. [1] propôs um sistema de classificação baseado em redes neurais convolucionais (CNNs) e pooling temporal utilizou uma arquitetura de rede neural recorrente (RNN) com attention combinou CNNs e RNNs.

Embora esses trabalhos apresentam resultados promissores, há espaço para aprimorar a performance através da utilização de técnicas mais recentes de aprendizado profundo, como o modelo Wav2vec2 e o HuBERT, e da exploração de novas abordagens, como o aumento de dados. Um estudo recente, [2], avaliou e comparou diversos modelos de aprendizado profundo para tarefas de representação de áudio, incluindo o Wav2vec2 e o HuBERT. O estudo demonstra que esses modelos, pré-treinados no AudioSet, alcançam performance de estado da arte em diversos datasets de classificação, incluindo o IRMAS.

Este trabalho tem como objetivo pesquisar e avaliar abordagens recentes com Deep Learning, utilizando representações de modelos de estado-da-arte para treinar um classificador. Envolve a criação de um dataset com rótulos de instrumentos e a geração de dados multilabel a partir da mixagem de áudios single label. O foco está na análise do impacto do aumento de dados na performance da classificação. Os resultados

indicam que o aumento de dados contribui para a melhora da performance, especialmente quando combinado com técnicas de aprendizado profundo.

2. Datasets

2.1. Dataset IRMAS

O dataset IRMAS (Instrument Recognition in Musical Audio Signals) é um conjunto de dados amplamente utilizado para a pesquisa em reconhecimento de instrumentos musicais. Ele contém gravações de áudio de performances musicais etiquetadas com os instrumentos predominantes em 11 classes: Violoncelo, Clarinete, Flauta, Guitarra acústica, Guitarra elétrica, Órgão, Piano, Saxofone, Trompete, Violino e Voz humana. Cada gravação pode conter múltiplos instrumentos, tornando-o adequado para tarefas de classificação multilabel. As gravações estão distribuídas em diferentes gêneros musicais, o que adiciona variabilidade e complexidade ao dataset.

2.2. Dataset Instrument-UFG

Para aumentar a diversidade e a quantidade de dados multilabel, foi criado o dataset Instrument-UFG. Este dataset é composto por gravações selecionadas a partir de diferentes fontes, incluindo:

- 3.108 músicas selecionadas do YouTube, abrangendo uma variedade de gêneros musicais.
- 3.341 gravações específicas de instrumentos, coletadas do SoundCloud. Estas gravações foram escolhidas para cobrir uma ampla gama de timbres e estilos de execução.

As gravações passaram por processos de remoção de silêncio, segmentação em trechos de 3 segundos e resample para a taxa de 16kHz, compatível com o Wav2vec2, garantindo a uniformidade dos dados para o treinamento do modelo.

2.3. Dataset com Mixagem

Os exemplos multilabel são criados a partir de dados originalmente single-label, são utilizadas técnicas de mixagem de áudio. Áudios de diferentes instrumentos presentes no conjunto de treino são mixados, criando novos exemplos de treinamento que contêm múltiplas etiquetas. Esse processo envolve a combinação de duas ou mais gravações de diferentes instrumentos, gerando novos exemplos de treinamento com múltiplos instrumentos tocando simultaneamente. O dataset resultante é composto por 1.000 áudios mixados, aplicando essa técnica tanto ao dataset IRMAS quanto ao dataset Instrument-UFG.

Table 1. Quantidade de dados de cada dataset

Instrumentos	IRMAS	Instrument-UFG (YouTube)	Instrument-UFG (SoundCloud)	Total
Violoncelo	388	630	53	1.071
Clarinete	505	173	160	838
Flauta	451	412	354	1.217
Guitarra Acústica	637	x	647	1.284
Guitarra Elétrica	760	x	72	832
Órgão	682	252	346	1.280
Piano	721	336	1.200	2.257
Saxofone	626	181	210	1.017
Trompete	577	129	197	903
Violino	580	995	102	1.677
Voz Humana	778	x	x	778
Total Geral	**6.705**	**3.108**	**3.341**	**13.154**

3. Metodologia

3.1. Composição dos Datasets e Treinamento

Nesta subseção, detalhamos a composição dos conjuntos de dados utilizados no treinamento do modelo. As combinações testadas foram:

- **Dataset IRMAS train:** Conjunto de treinamento original do IRMAS.
- **Dataset IRMAS train + Dataset Instrument-UFG:** Conjunto de treinamento original do IRMAS combinado com o dataset Instrument-UFG.
- **IRMAS train + samples mixados do IRMAS train.**
- **Dataset IRMAS train + Dataset Instrument-UFG + samples mixados (IRMAS train + Instrument-UFG):** Conjunto de treinamento original do IRMAS combinado com o dataset Instrument-UFG e amostras mixadas dos dois datasets para criar dados multilabel.

Utiliza-se o modelo Wav2vec2 e realiza-se um ajuste fino específico para o dataset IRMAS. Este ajuste fino é essencial para adaptar o modelo às características específicas do IRMAS, melhorando a precisão na classificação dos instrumentos musicais presentes nas gravações.

Combina-se o Dataset de treinamento original do IRMAS com o dataset Instrument-UFG. Esta combinação é utilizada para treinar o modelo, visando melhorar a capacidade do modelo de reconhecer uma variedade mais ampla de instrumentos musicais.

Para criar exemplos multilabel a partir de dados originalmente single-label, utilizamos a técnica Mixagem. Áudios de diferentes instrumentos foram combinados, gerando novos exemplos de treinamento com múltiplas etiquetas. Para balancear os dados, calculamos uma probabilidade de amostragem que prioriza classes minoritárias e garante ao menos um par para cada combinação possível de classes. Essa técnica foi aplicada aos datasets IRMAS e Instrument-UFG.

O treinamento é conduzido por 10 epochs, salvando o checkpoint com menor loss de validação. Cada modelo foi então treinado usando todo o conjunto de treino em questão e avaliado no conjunto de teste do IRMAS. Para comparar as diferentes abordagens, utilizamos as métricas de precisão, recall e F1-score.

O classificador utilizado é uma camada de cabeça personalizada para o modelo Wav2vec2, projetada para a tarefa de classificação de áudio. A estrutura do classificador inclui:

1. **Input Features**: As características extraídas pelo modelo Wav2vec2.
2. **Dropout Layer**: Uma camada de dropout para reduzir o overfitting, aplicando uma taxa de dropout.
3. **Dense Layer**: Uma camada densa (linear) que transforma a dimensionalidade das características extraídas.
4. **Tanh Activation Function**: Uma função de ativação tanh para introduzir não-linearidade.
5. **Dropout Layer**: Outra camada de dropout para maior regularização.
6. **Output Projection**: Uma camada linear final que projeta a saída para o número de etiquetas desejadas, correspondendo aos instrumentos musicais.

Esta arquitetura transforma as representações de áudio em previsões de classes específicas, melhorando a precisão na classificação de múltiplos instrumentos musicais simultaneamente.

4. Resultados

As figuras 1 e 2 apresentam a evolução da acurácia e do F1 Score ao longo das épocas para diferentes combinações de datasets utilizados no treinamento do modelo. A comparação inclui: IRMAS + Mixagem, IRMAS + UFG + Mixagem, IRMAS + UFG, e IRMAS.

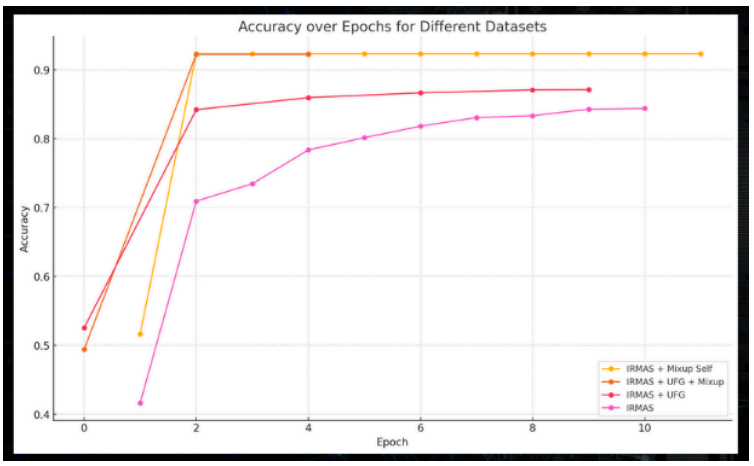

Figura 1. Evolução da Acurácia ao Longo das Épocas para Diferentes Combinações de Datasets

A acurácia é medida ao longo de 10 épocas, mostrando que o uso combinado de IRMAS com UFG e a técnica Mixagem atinge uma acurácia superior desde as primeiras épocas e se mantém estável. O uso de Mixagem dentro do próprio IRMAS também mostra um aumento significativo na acurácia, enquanto o uso isolado do IRMAS e a combinação com UFG sem Mixagem apresentam uma evolução mais gradual.

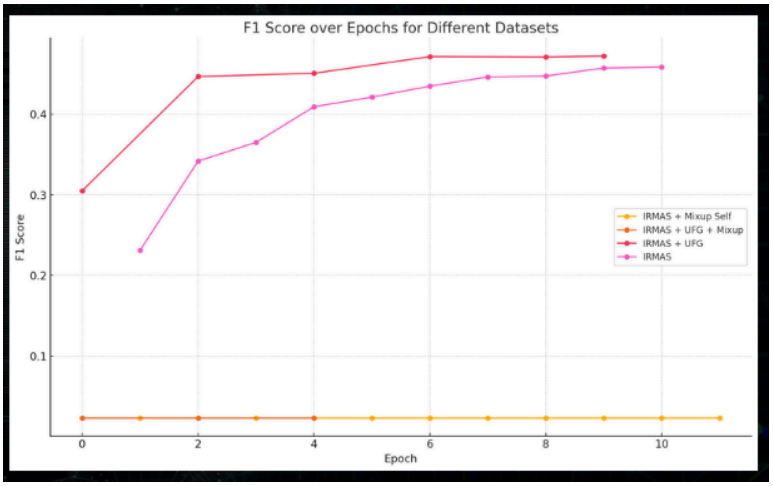

Figura 1. Evolução do F1 Score ao Longo das Épocas para Diferentes Combinações de Datasets

O F1 Score é medido ao longo de 10 épocas, mostrando que o uso combinado de IRMAS com UFG e a técnica Mixagem atinge um F1 Score superior desde as primeiras épocas e se mantém estável. O uso de Mixagem dentro do próprio IRMAS apresenta um F1 Score baixo e estável, enquanto o uso isolado do IRMAS e a combinação com UFG sem Mixagem mostram uma evolução gradual.

A tabela 2 apresenta os resultados da melhor época (epoch) de treinamento e validação para diferentes combinações de datasets e técnicas utilizadas no modelo. Cada

linha da tabela corresponde à combinação de datasets e técnicas que resultaram nas melhores métricas de desempenho observadas durante o treinamento.

Table 2. Resultados de Treinamento e Validação para Diferentes Combinações de Datasets

Dataset	Epoch	Training Loss	Validation Loss	Accuracy	F1 Score	Precision	Recall
IRMAS	4	0.178200	0.356863	0.783798	0.409268	0.364975	0.678431
IRMAS + Mixagem	8	0.031000	0.039814	0.923230	0.022747	0.014139	0.090909
IRMAS + Dataset UFG	2	0.097600	0.379907	0.842222	0.446743	0.473119	0.580971
IRMAS + Mixagem + Dataset UFG	4	0.030700	0.042722	0.922590	0.022936	0.014257	0.091667

Esses resultados mostram como diferentes combinações de datasets e técnicas influenciam a performance do modelo em termos de generalização e precisão. A melhor performance em termos de acurácia foi alcançada com a combinação de IRMAS e Mixagem, enquanto a melhor performance em termos de F1 Score foi alcançada com a combinação de IRMAS e o Dataset UFG.

5. Conclusões

Neste trabalho, investigamos a performance de diferentes abordagens para a classificação multilabel de instrumentos musicais utilizando os datasets IRMAS e Instrument-UFG. Aplicamos técnicas avançadas de aprendizado profundo, incluindo o modelo Wav2vec2, para otimizar a precisão da classificação.

Criamos um dataset rotulado e desenvolvemos uma abordagem de mixagem de áudios single-label para gerar dados multilabel, aumentando significativamente a complexidade e a variedade do dataset. Essa técnica de mixagem foi aplicada tanto no dataset IRMAS quanto no dataset Instrument-UFG, resultando em um total de 13.154 exemplos de treinamento.

Nossos experimentos mostraram que o uso combinado de datasets e técnicas de mixagem de dados resultou em uma melhoria significativa na acurácia do modelo. Em particular, a combinação do IRMAS com o dataset Instrument-UFG e a aplicação da técnica Mixagem demonstrou a melhor performance em termos de acurácia e F1 Score ao longo das épocas de treinamento. No entanto, também observamos que a adição de dados e a aplicação de técnicas de mixagem exigem atenção especial às métricas de balanceamento, como o F1 Score, para garantir uma performance equilibrada em todas

as classes. Embora a acurácia geral tenha melhorado, as métricas de precisão e recall indicaram a necessidade de ajustes adicionais para evitar o desbalanceamento.

Em resumo, nossos resultados indicam que a utilização de técnicas de mixagem de dados e a integração de datasets adicionais são estratégias promissoras para a classificação multilabel de instrumentos musicais. Esses métodos não só aumentam a complexidade do dataset, mas também melhoram a generalização do modelo, proporcionando um avanço significativo na área de reconhecimento automático de áudio.

Referencias

[1] Han, Y., Kim, J., & Lee, K. (2016). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. Disponivel em: https://arxiv.org/pdf/1605.09507

[2] La Quatra, M., Koudounas, A., Vaiani, L., Baralis, E., Cagliero, L., Garza, P., & Siniscalchi, S. M. (2024). Benchmarking Representations for Speech, Music, and Acoustic Events. Disponivel em: https://arxiv.org/pdf/2405.00934

[3] Kratimenos, A., Avramidis, K., & Garoufis, C. (2020). Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain. Disponivel em: https://ieeexplore.ieee.org/document/9287745.

[4] Ajayakumar, R., & Rajan, R. (2020). Predominant Instrument Recognition in Polyphonic Music Using GMM-DNN Framework. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain. Disponivel: https://ieeexplore.ieee.org/document/9179626.