Estudo de Técnicas de Deep Learning na Classificação de Amostras de Áudio de Instrumentos Musicais

Ranieri W. Moura Gusmão [1], José Augusto M. Nacif [1], Alex Borges Vieira [2]

***Abstract*— Nos tempos modernos, diante da implacável evolução da tecnologia, o ser humano tende a estabelecer uma relação mais próxima das máquinas. Os promissores estudos voltados ao Machine Learning demonstram que a relação humano-máquina evoluiu significativamente. Neste trabalho, será abordado Deep Learning, que é um dos ramos específicos do Machine Learning, para classificação de amostras de áudio de instrumentos musicais. O objetivo é determinar a precisão do desempenho entre Redes Neurais Recorrentes e Convolucionais, realizando um comparativo e identificando fatores significativos que possam interferir no resultado final. Após o treinamento das Redes Neurais Recorrentes e Convolucionais, obteve-se, respectivamente, taxas de precisão de 88% e 96,5%. Percebe-se então, que para classificação de amostras de áudio, as técnicas de convolução empregadas nas Redes Neurais Convolucionais são mais eficientes.**

Palavras-chave: deep learning, áudio, pré-processamento, redes neurais, recorrência, convolução

I. Introdução

Nos tempos modernos, diante da implacável evolução da tecnologia, o ser humano tende a estabelecer uma relação mais próxima das máquinas. Não apenas aprimorando suas características físicas e lógicas, mas criando uma identidade humana em seus complexos cérebros artificiais. Os promissores estudos voltados ao Machine Learning demonstram que a relação humano-máquina evoluiu tão rapidamente que os novos objetivos não almejam apenas facilitar a vida do ser humano em seu meio, mas ensinar a máquina a pensar praticamente como ele. O que pode ser intrigante e nocivo para alguns, mas uma descomunal evolução para outros. De qualquer maneira, algo é certo, nunca se está satisfeito com o que se tem.

Neste trabalho, será abordado Deep Learning, que é um dos ramos específicos do Machine Learning. Como podemos inferir do próprio nome, Deep Learning estabelece o aprendizado mais aprofundado de uma rede neural, através de algoritmos minuciosamente estruturados, embasados em um viés hierárquico que se apoia na teoria de grafos lineares e não lineares. O cenário baseia-se em sons de instrumentos musicais e é composto por 10 classes de amostras de áudio obtidas em [1]. Em uma visão geral, são 300 amostras que serão submetidas a etapas de pré-processamento e aplicadas em diferentes estruturas de redes neurais.

O objetivo é determinar a precisão do desempenho dessas redes, realizando um comparativo e identificando fatores significativos que possam interferir no resultado final. Tal proposta pode favorecer o estudo de outras áreas relacionadas ao assunto, já que manipular com eficiência dados de áudio para treinar redes neurais é um grande desafio.

Em [2], compara-se o desempenho de um método de classificação de áudio que emprega técnicas de Deep Learning, com dois métodos mais simples, o SVM (*Support Vector Machine*) e o GMM (*Gaussian Mixture Models*). Os trabalhos [3] e [4], avaliam o desempenho de Redes Neurais Convolucionais em grandes conjuntos de amostras de imagens. Em [5], é tratado algoritmos aplicados em Redes Neurais Convolucionais na classificação de áudios específicos.

Este artigo está organizado em 6 seções: (I) introdução; (II) referencial teórico; (III) metodologia; (IV) resultados; (V) trabalhos relacionados e (VI) conclusões e trabalhos futuros. Ao final, as referências.

II. Referencial Teórico

A. Transformada de Fourier

O princípio dessa técnica é uma transformada integral que expressa determinada função em termos de bases de senoide (onda seno), consistindo em oscilações repetitivas e suaves. O que ocorre basicamente é uma decomposição de uma função temporal, na forma de sinal ou som, em frequências.

$$F(k) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i k x} dx \tag{1}$$

- i =
- k = Frequência
- x = Tempo

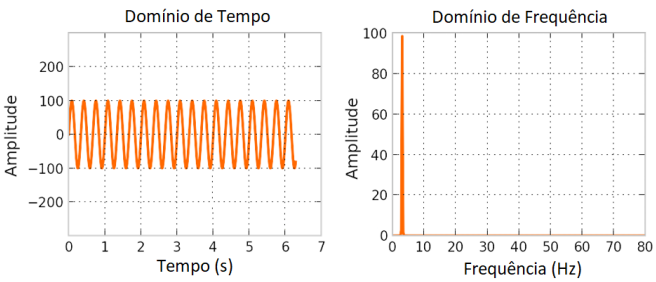

Fig. 1: Transformada de Fourier.

Com base nas próximas figuras, é possível visualizar o comportamento de uma amostra de áudio antes (Figura 2) e depois da aplicação da Transformada de Fourier (Figura 3).

[1] Universidade Federal de Viçosa – Florestal, MG, Brasil
[2] Universidade Federal de Juiz de Fora – MG, Brasil

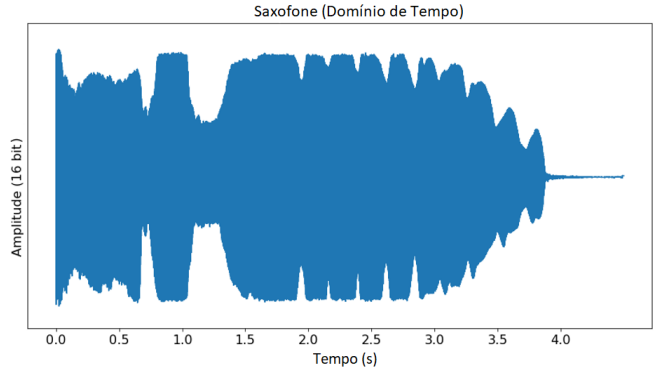

Fig. 2: Comportamento de uma amostra de áudio antes de ser submetida a uma Transformada de Fourier.

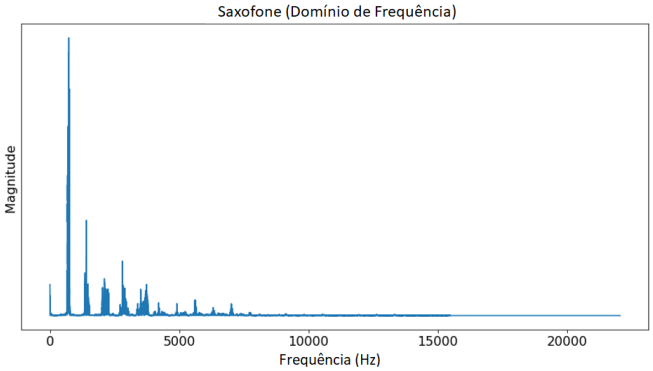

Fig. 3: Comportamento de uma amostra de áudio após ser submetida a uma Transformada de Fourier.

B. Coeficientes Cepstrais de Frequência de Mel (MFCC)

Elaborado e divulgado em [6], é um recurso amplamente utilizado até os dias de hoje na identificação e classificação de sons. Para determiná-los com êxito, é essencial seguir algumas etapas.

1) Dividir o sinal em pequenos quadros. Considerando amostras de 16 KHz e cada quadro em 25 ms (padrão), temos como resultado 400 quadros (16000 $\times$ 0.025). Configura-se também uma tolerância, ou *frame step*, de 10 ms (padrão) entre cada conjunto de 400 quadros, evitando assim a sobreposição dos valores;
2) Para cada quadro no passo 1, calcular uma estimativa da densidade espectral do sinal, um periodograma, no espectro de potência. Para isso, aplicamos a **Transformada de Fourier**, na Equação 1, usando o valor resultante como parâmetro para a **Fórmula de Estimativa do Periodograma**, na Equação 2, onde N é o valor de um Quadro de Análise de Amostra, configurado em 512 pontos;

$$P_i(k) = \frac{1}{N}|F(k)|^2 \tag{2}$$

- $F(k)$ = Transformada de Fourier
- N = Quadro de Análise de Amostra

3) Determinar os bancos de filtros de Mel através da definição de um conjunto (padrão) de 26 filtros triangulares (Figura 4). Estipula-se um limite inferior e superior de frequência, que será respectivamente 0 Hz e 4410 Hz. O próximo passo é aplicar a Equação 3 que converterá frequência (Hz) em Mel. Os resultados são somados a 40 pontos adicionais espaçados linearmente entre os limites definidos no início dessa etapa. Esse número é determinado avaliando um espaço de 1s dividido por 25 ms (1 quadro). Enfim, aplica-se a Equação 4, que realiza a operação inversa da Equação 3.

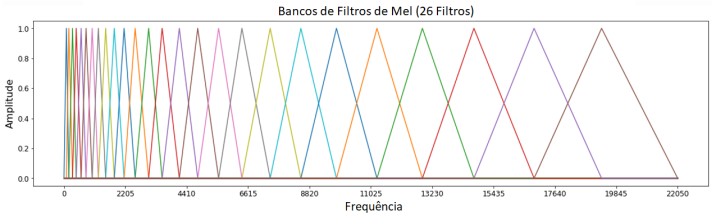

Fig. 4: Bancos de Filtros de Mel.

$$M(f) = 1125\ ln(1 + \frac{f}{700}) \tag{3}$$

$$M^{-1}(f) = 700\ (exp(\frac{m}{1125}) - 1) \tag{4}$$

- f = Frequência
- m = Escala de Mel

4) Calcular o logaritmo de todas as 26 energias do banco de filtros no passo 3. Essa etapa amplifica o som de maneira que possa ser factível para a audição humana. O uso do logaritmo, ao invés de outra operação matemática, viabiliza a manipulação de coeficientes cepstrais que obtém-se na etapa final, o que caracteriza uma técnica de normalização de canal.
5) Aplicar a Transformada Discreta de Cosseno (DCT) em cada um dos 26 resultados obtidos no passo 4, para obter os 26 coeficientes cepstrais. Essa etapa correlaciona as energias, possibilitando que matrizes de covariância diagonal sejam utilizadas para modelar os recursos. Consideraremos os 13 menores valores que serão os **Coeficientes Cepstrais de Frequência de Mel (MFCC)**.

C. Redes Neurais Abordadas

A escolha das redes neurais utilizadas tem como intuito, demonstrar a diferença do comportamento entre elas mediante a manipulação do conjunto de dados proposto para estudo. Neste artigo, aborda-se os seguintes tipos, os quais serão definidos de acordo com suas características peculiares:

1) **Redes Neurais Recorrentes:** como o próprio nome indica, são redes que, através de recorrência, manipulam os dados fornecidos a elas. Consideram tempo e sequência, apresentando uma dimensão temporal.

Se comportam de maneira que analisam como entrada tanto as novas informações, quanto tudo que foi percebido anteriormente no tempo.

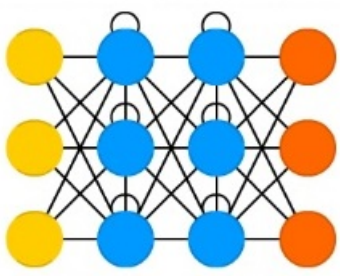

Fig. 5: Redes Neurais Recorrentes.

Na ilustração da Figura 5, os nós em amarelo representam as entradas, enquanto os nós em laranja são as saídas. Os nós em azul denotam recorrência, ou seja, tomam como parâmetros de aprendizado informações atuais e antigas, aprimorando o resultado.

2) **Redes Neurais Convolucionais:** nesses tipos de redes, o foco está na convolução. O desenvolvimento desse algoritmo teve como princípio, o córtex visual de animais. São milhões de agrupamentos celulares complexos, sensíveis a pequenas sub-regiões do campo visual, denominados campos receptíveis.
De acordo com [7], pode-se dividir a estrutura dessas redes em três objetivos principais:
   - **Extração de características:** cada neurônio recebe sinais de entrada de um campo receptível da camada anterior, viabilizando a extração de características locais;
   - **Mapeamento de características:** cada camada da rede é composta por diversos mapas de características, que consistem em regiões onde os neurônios compartilham os mesmos pesos (filtros) e dão robustez ao modelo;
   - **Subamostragem:** após cada camada de convolução, aplica-se uma camada de subamostragem, que é a coleta de amostras de cada mapa de característica.

   Na Figura 6, semelhante à Figura 5, os nós em amarelo representam as entradas e os nós em laranja as saídas. Já os nós em rosa (e rosa com circunferência de borda preta) são, respectivamente, os filtros e as convoluções. Os nós em verde representam as camadas de convolução.

III. Metodologia

Com base no referencial teórico, foi implementado um algoritmo para atender cada etapa do projeto, disposta no diagrama da Figura 8:

A) **Detecção de Nível de Ruído:** é eliminado dados redundantes ou irrelevantes das amostras de áudio;
B) **Pré-processamento de Dados:** aplica-se as técnicas de Transformada de Fourier e Coeficientes Cepstrais de Frequência de Mel (MFCC);
C) **Aplicação em Redes Neurais:** é realizado o treinamento de Redes Neurais Recorrentes e Convolucionais.

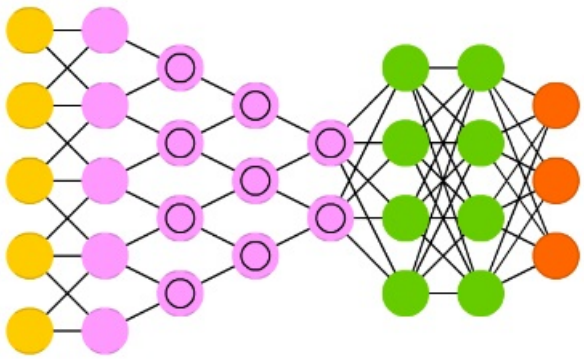

Fig. 6: Redes Neurais Convolucionais.

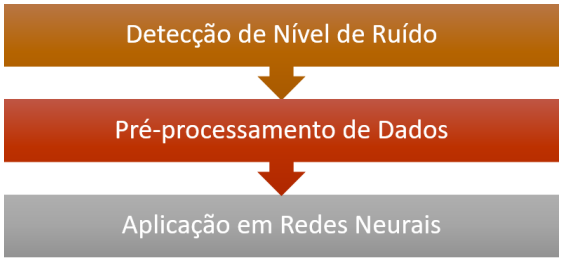

Fig. 7: Etapas do projeto.

A. *Detecção de Nível de Ruído*

Para a otimização dos resultados, realiza-se previamente a eliminação de dados redundantes ou irrelevantes. Tal procedimento pode ser denotado como *Noise Floor Detection* ou **Detecção de Nível de Ruído**.

O algoritmo que será utilizado nesse tópico, parte de um princípio bem simples: identificar amplitudes baixas ou nulas e eliminar da amostra de aúdio. Esse procedimento, se efetuado de maneira correta, reduz o tamanho do conjunto de dados e pode aumentar o desempenho das redes neurais.

Por mais simples que seja o princípio do algoritmo, implementá-lo pode ser uma tarefa desafiadora. No caso, determinamos um limite (*threshold*) desprezível da amplitude para implementar uma cobertura de sinal (*signal envelope*). No contexto de Física e Engenharia, a cobertura de sinal é uma onda regular delineando seus extremos que permite estimar uma amplitude constante.

Estipula-se uma taxa de 0.0005s como limite e, na Figura 8, pode ser visto o resultado de uma amostra (Figura 2) após ser submetida a esse algoritmo.

B. *Pré-processamento de Dados*

É a etapa principal, onde os dados são preparados para se adequarem aos procedimentos aplicados durante o treinamento das redes neurais. Um dado no formato de áudio manifesta-se através de ondas, característica a qual não viabiliza a fácil identificação individual de um determinado som sem processamento prévio.

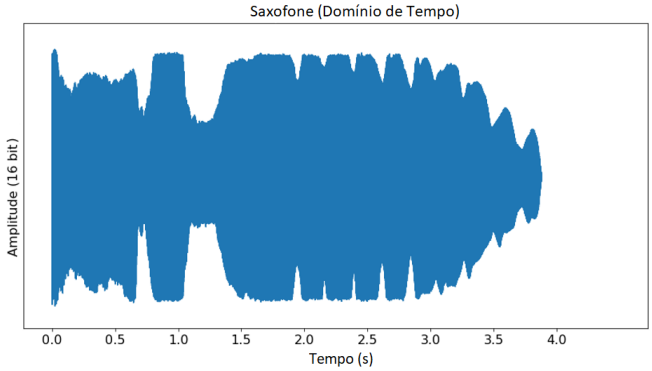

Fig. 8: Amostra de áudio (Figura 2) após a Detecção de Nível de Ruído. Perceba que a parte final foi eliminada.

A frequência de uma classe de áudio é determinada através do total de duração de suas amostras. É importante salientar que nem sempre a frequência que uma suposta amostra aparece em um conjunto, determina a eficácia em classificá-la. A qualidade do áudio, por exemplo, é uma característica essencial para definir o quão fácil será a sua classificação.

Na Figura 9 é ilustrada a distribuição das classes de áudio antes e depois da execução do algoritmo de Detecção de Nível de Ruído.

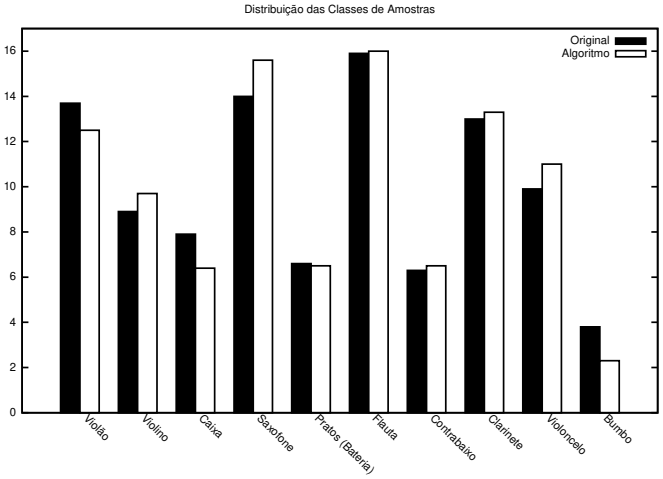

Fig. 9: Distribuição das classes de áudio antes (preto) e depois (branco) da execução do algoritmo de Detecção de Nível de Ruído.

Na literatura, duas técnicas de pré-processamento geralmente são aplicadas: **a Transformada de Fourier** e os **Coeficientes Cepstrais de Frequência de Mel (MFCC)**. Preza-se também a eficácia do algoritmo do tópico anterior. Uma implementação equivocada causaria instabilidade na amplitude das amostras e poderia reduzir significativamente o desempenho das redes neurais do próximo tópico.

C. Aplicação em Redes Neurais

Devido ao grande reconhecimento no cenário, serem intuitivamente simples de descrever, explicar e implementar, os tipos de redes neurais abordados foram:

- **Redes Neurais Recorrentes;**
- **Redes Neurais Convolucionais.**

Outro fator relevante é que ambas são pioneiras de outros tipos de redes neurais, que geralmente são variações embasadas em algum objetivo específico.

Esta é a última etapa, onde as amostras já estão aptas para serem submetidas aos algoritmos de treinamento.

IV. Resultados

A máquina disponibilizada para o projeto é um computador pessoal com as seguintes especificações:

- **Memória:** DDR3 6 GB 667 MHz
- **GPU:** NVidia GeForce GTX 1030 2 GB
- **Sistema Operacional:** Windows 10 Professional

Para o desenvolvimento do código que gerou os resultados deste estudo foi utilizada a linguagem de programação Python. Esta tem sido muito reconhecida no âmbito de Machine Learning e aplicada em inúmeros projetos. É importante citar que para facilitar a implementação, modificação e extensibilidade do código, o projeto foi criado na IDE Spyder, obtida em um framework denominado Anaconda.

Mais precisamente com relação ao desenvolvimento do código desse projeto, as bibliotecas Keras e TensorFlow foram instaladas. Sendo a última, muito útil na manipulação de amostras de áudio para treinamento de redes neurais.

Em ambos treinamentos, os parâmetros gerados pelos modelos foram armazenados previamente na memória, favorecendo assim o processamento. Em caso de redes neurais mais robustas e complexas, isso se tornará inviável, podendo causar overflow. Para propósitos de validação, a proporção empregada de amostras foi 75% e 25%.

As Redes Neurais Recorrentes demandaram durante cada período, em média, 15 segundos para concluir o treinamento. A partir do quarto período, as classificações começaram a se tornar gradativamente mais precisas. No final, obteve-se uma taxa de precisão de aproximadamente 88%. Veja Figura 10.

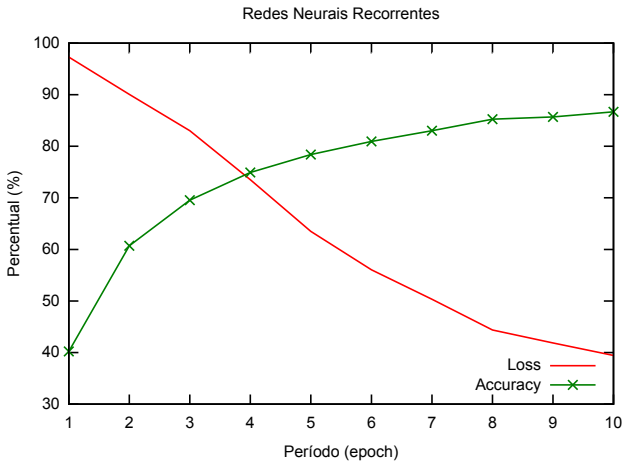

Fig. 10: Desempenho das Redes Neurais Recorrentes.

No que se refere às Redes Neurais Convolucionais, foram aplicadas 4 camadas de convolução. Para cada período,

despendeu-se cerca de 20 segundos. A partir do segundo, o desempenho evoluiu significativamente, alcançando no final do treinamento, uma taxa de precisão de aproximadamente 96,5%. Veja Figura 11.

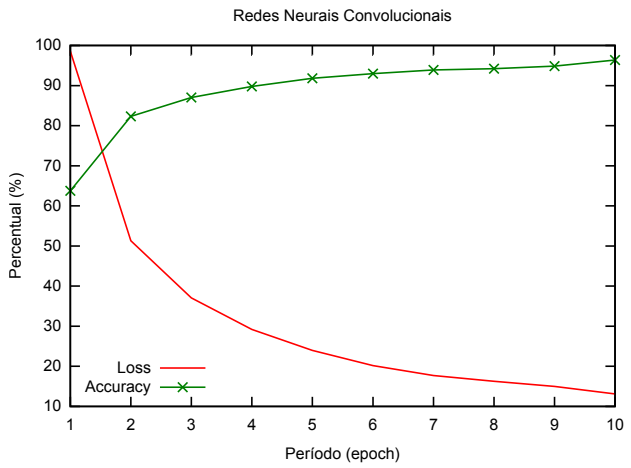

Fig. 11: Desempenho das Redes Neurais Convolucionais.

As métricas de *loss* estão relacionadas aos falsos positivos e falsos negativos que causam inconsistência na classificação das classes. Por outro lado, as métricas de *accuracy* tratam os acertos obtidos pelas redes neurais durante o treinamento.

V. Trabalhos Relacionados

No decorrer da evolução dos estudos embasados em Deep Learning, pode-se considerar e destacar os objetivos de alguns trabalhos. Estes que abordam assuntos com algumas semelhanças, mas também diferenças em determinados pontos de vista. Tudo isso é muito válido para desenvolver novas ideias ou obter conclusões de situações controversas.

Em [2], é utilizado um conjunto de dados de eventos comuns, composto por 84 arquivos (228 minutos) divididos em 4 classes. O principal objetivo é comparar o desempenho de um método de classificação que emprega técnicas de Deep Learning, com dois métodos mais simples, o SVM (*Support Vector Machine*) e o GMM (*Gaussian Mixture Models*), métodos vetoriais e probabilísticos, respectivamente.

Abordando como cenário grandes conjuntos de dados no formato de imagens, em [3] e [4], são realizados experimentos que avaliam o desempenho de Redes Neurais Convolucionais em diferentes GPUs, visando obter o comportamento e a eficácia diante das inúmeras classes de imagens envolvidas.

Em um estudo de algoritmos aplicados em Redes Neurais Convolucionais, o [5] trata inúmeros conjuntos de dados visando a identificação do locutor de uma mensagem e seu gênero sexual durante chamadas telefônicas. Em outra abordagem, há também a classificação de gêneros musicais e seus respectivos artistas.

Diante dos trabalhos citados, nosso projeto trata ambos tipos de redes neurais abordadas em sua forma nativa, levando em consideração prioritariamente a manipulação das amostras de aúdio antes de sua aplicação nas redes neurais.

VI. Conclusões e Trabalhos Futuros

As Redes Neurais Recorrentes não são tão viáveis para a classificação de amostras de áudio. Isso não significa que são inaplicáveis a esse âmbito, pois podem ser válidas em questões específicas, como reconstrução ou geração de partes de áudio. De qualquer forma, no quesito classificação de amostras de áudio, as Redes Neurais Convolucionais são mais eficientes.

Em novos estudos, serão investigadas novas técnicas mais confiáveis de pré-processamento de aúdio para a eliminação de partes de áudios redundantes ou irrelevantes. O uso de um conjunto de dados mais robusto e a comparação de outros tipos de redes neurais também são futuras propostas para o andamento do projeto. Além disso, pode-se avaliar o desempenho das redes neurais para cada instrumento individualmente.

References

[1] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, Xavier Serra. "General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline". Proceedings of the DCASE 2018 Workshop (2018)

[2] Kons, Zvi & Toledo-Ronen, Orith. (2013). Audio event classification using deep neural networks. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 1482-1486.

[3] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.

[4] krishna, M & Neelima, M & Mane, Harshali & Matcha, Venu. (2018). Image classification using Deep learning. International Journal of Engineering & Technology. 7. 614. 10.14419/ijet.v7i2.7.10892.

[5] Lee, Honglak & Pham, Peter & Largman, Yan & Ng, Andrew. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. NIPS. 1096-1104.

[6] Davis, Stan and Paul Mermelstein. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Se." (1980).

[7] Haykin, S. S. (2009). Neural networks and learning machines. Upper Saddle River, NJ: Pearson Education.