Reconhecimento de Som com Redes Neurais Convolucionais

Asiel Aldana Ortiz
Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.
Rio de Janeiro. Brasil
asiel.aldana89@gmail.com

RESUMO

As técnicas de aprendizado profundo tornaram-se um foco de atenção nos últimos anos como uma maneira de transformar entradas de dados em representações mais eficazes usando algoritmos de extração de recursos cada vez mais eficientes. Esse interesse cresceu na área de processamento e classificação automática de sinais sonoros. Este artigo propõe a implementação de um modelo baseado em uma rede neural convolucional (CNN), capaz de aprender padrões específicos a partir da análise espectral de uma sequência de áudio e, assim, classificar cada evento sonoro a partir de sua classe de referência. Para este projeto, dois tipos de abordagens foram considerados com base em diferentes representações espectro-temporais das seqüências sonoras utilizadas (MFC-Spectrogram e MFCCs). Para avaliar o desempenho do modelo, foi utilizado o conjunto de dados disponível ao público (UrbanSound8k), a partir do qual foram obtidos valores médios de acurácia acima de 75% para o total de folds apresentados pelo conjunto de dados.

Palavras-chave

Classificação; rede neural convolucional; espectro.

ABSTRACT

Deep learning techniques have become a focus of attention in recent years as a way of transforming data inputs into more effective representations using increasingly efficient feature extraction algorithms. Such interest has grown in the area of processing and automatic classification of sound signals. This article proposes the implementation of a model based on a convolutional neural network (CNN), capable of learning specific patterns from the spectral analysis of an audio sequence and thus being able to classify each sound event based on its reference class. For this project, two types of approaches were considered, based on different spectrum-temporal representations of the sound sequences used (MFC-Spectrogram and MFCCs). To evaluate the performance of the model, the publicly available dataset (UrbanSound8k) was used, from which average aquatic values above 75% were obtained for the total folds presented by the dataset.

Keywords

Classification; convolutional neural network; spectrum.

1. INTRODUÇÃO

A classificação automática de sons ambientais é uma técnica utilizada em várias aplicações, como vigilância remota, automação residencial e reconhecimento de eventos sonoros em atividades industriais. Essa técnica, juntamente com o recente aumento das tecnologias móveis, está sendo explorada com resultados interessantes implementados em aplicativos que fornecem serviços que vão do reconhecimento de espécies de aves até a classificação de possíveis anomalias em sistemas industriais que avaliam seqüências sonoras associadas a determinados eventos.

Os sons ambientes consistem em vários sons não humanos (excluindo música) da vida cotidiana. Nos últimos anos, várias tentativas foram feitas para reconhecer sons ambientais; atualmente, há um aumento focado em classificá-los usando técnicas de aprendizado profundo [1,2].

A diferença importante entre a fala e o som ambiente é que os primeiros são fortemente estruturados e claramente demarcados, enquanto os segundos não possuem uma estrutura comum [3]. Isso o torna um problema completamente novo.

Lu-Zhang-Li [4] conclui que o SVM fornece uma classificação mais precisa dos sons ambientais do que os vizinhos k-Neighbourhood (kNN) e Gaussian Mixing Models (GMM). Dessa maneira, pode-se pensar que, para esse problema, existe uma solução elegante usando técnicas de aprendizado profundo, pois foi demonstrado que as redes neurais profundas são capazes de lidar com grandes quantidades de dados e modelar características complexas devido aos avanços no poder da computação, incluindo o amplo uso de GPUs.

A abordagem mais comum da classificação de sons baseada em aprendizado profundo é converter uma sequência de áudio em uma imagem e, em seguida, usar uma rede neural para processá-la.

Mostafa-Billor [5] realiza a classificação da música usando redes neurais probabilísticas com resultados satisfatórios. A maioria das abordagens sólidas de classificação usa padrões de reconhecimento supervisionado, no entanto, Zhang e Schuller [6] expressam o problema de que a rotulagem manual de conjuntos de dados é muito cara e recomendam o aprendizado semi-supervisionado como uma solução melhor.

McLoughlin [3] afirma que a classificação do som em ambientes ruidosos realistas é desafiadora e propõe uma rede neural profunda como uma solução viável.

Piczak [7] e Zhang [1] transmitem a ideia de que redes neurais convolucionais têm as melhores taxas de precisão na análise de espectrograma. Para sinais de áudio, uma representação popular é

o Mel-Spectrogram (Mel Frequency Cepstrum (MFC)) [8,9]. Do anterior, você também pode obter os coeficientes (MFCCs) que, como um todo, caracterizam com precisão um envelope espectral.

Com base nos estudos mais recentes, pode-se sugerir que as redes neurais convolucionais profundas (CNN) [10] podem, em princípio, ser adequadas para o problema da classificação do som ambiental: primeiro, elas são capazes de capturar padrões de modulação de energia em através do tempo e da frequência, quando aplicados a entradas do tipo espectrograma, que, como mencionado, são uma representação importante para distinguir entre sons diferentes, geralmente altos [7]. Segundo, usando kernels convolucionais com um pequeno campo receptivo, a rede deve, em princípio, ser capaz de aprender com sucesso e, em seguida, identificar padrões de tempo do espectro que são representativos de diferentes tipos de som, mesmo que parte do som é mascarado (em tempo / frequência) por outras fontes (ruído).

No entanto, existe uma lacuna na pesquisa sobre o desempenho e a utilidade de redes neurais profundas, projetadas para o reconhecimento normal de objetos, quando se trata de classificar espectrogramas e imagens semelhantes relacionados ao som. O objetivo deste estudo é explorar nessa direção de pesquisa e verificar, a partir da implementação de um modelo, se redes neurais profundas são capazes de fornecer resultados relevantes na classificação de sons ambientais usando suas características de espectro-tempo.

2. METODOS

2.1 Representação do sinal

O Mel-Spectrogram (MFC) é obtido computando a transformada de Fourier do período curto e mapeando sua potência espectral em uma escala de percepção audível (mel-scale) [11]. O uso de um banco de filtros no domínio da frequência é o ponto de partida para calcular os coeficientes característicos das MFCCs [12] de um determinado envelope espectral, a partir do cálculo da transformação linear do cosseno do logaritmo decimal do espectro de potência de uma sequência de áudio em mel-scale, no presente projeto, duas formas de representação de sinal foram usadas com base em suas informações de espectro-tempo (Mel-Spectrogram e MFCCs). Para o MFC, uma frequência **mel** de 128 bandas (eixo vertical) e 128 amostras de tempo (eixo horizontal) foi calculada, com o objetivo de obter uma imagem de 128x128 pixels (para cada Mel-Spectrogram o logaritmo decimal de cada potência foi computado). Para a representação da resolução temporal, considerou-se uma frequência de amostragem de 22050 Hz usando uma Janela Hann de 23 ms de comprimento (tamanho do salto entre janelas (512) dividido pela frequência de amostragem (22050)). Considerando a duração de cada Janela (23 ms), apenas sequências sonoras maiores que 2,95 segundos (2,95 / 0,023 = 128,3) foram analisadas. No caso da representação das MFCCs, foram escolhidos 20 coeficientes para cada faixa de frequência, para obter a resolução apropriada (128x128), foi feita uma extrapolação dos resultados obtidos das MFCCs.

2.2 Modelo proposto

O modelo de rede neural proposto é composto por 4 camadas convolucionais ($l \in \{1,2,3,4\}$) seguidas por três operações de Pooling e 2 camadas totalmente conectadas ($l \in \{5,6\}$), todas as camadas intermediárias com função de ativação Relu(AF) e Softmax na última camada densa (consulte a tabela 1).

Tabela 1. Estructura del modelo

Camada	Número do filtros	Tamanho do filtro	Strided max-pooling	AF
l_1	24	4x4	2x2	Relu
l_2	32	4x4	2x2	Relu
l_3	32	4x4	2x2	Relu
l_4	48	4x4	-	Relu
l_5	64(Hidden Unit)	-	-	Relu
l_6	10(Hidden Unit)	-	-	Softmax

Durante o treinamento, a função de perda utilizada foi "categorical_crossentropy" usando o otimizador Adamax, com um tamanho de lote de 100 amostras selecionadas aleatoriamente. Cada amostra de 2,95 segundos é retirada de uma posição aleatória do conjunto de treinamento. Foi utilizada uma taxa de aprendizado de 0,001, aplicando o Dropout com probabilidade de 0,5 às duas últimas camadas densas. Inicialmente, o modelo foi treinado com valores entre 20 e 100 vezes, embora os resultados revelassem tendências de super aprendizagem depois de 55 épocas, decidido fazer o treinamento final com 55 épocas.

2.3 Avaliação e resultados do modelo

Para avaliar o comportamento do modelo, foi utilizado o conjunto de dados UrbanSound8k [13], que contém um total de 8732 registros sonoros com suas respectivas anotações, divididos em 10 classes diferentes (air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music), agrupados em 10 folds. A duração máxima de cada áudio é variável, embora nenhum registro exceda 4 segundos de reprodução.

Uma das principais limitações de modelos baseados em aprendizado profundo é precisamente que essa técnica precisa ter um volume de dados grande o suficiente para garantir bons resultados. Ao fazer um estudo do conjunto de dados proposto, pôde-se observar que precisamente as classes menos representadas (car horn e gun shot) em termos de número de amostras (429 e 374) eram exatamente aquelas que continham uma duração média menor

(2.46s e 1.65s). Dessa forma, é necessário pensar em técnicas de aumento de dados para reduzir seu desequilíbrio. Neste trabalho, 2 métodos diferentes foram usados para aumentar os dados [14]:

- Time Stretching(TS): Cada amostra foi reduzida no tempo por três fatores (1,08, 0,82, 0,94).
- Pitch Shifting(PS): Cada amostra foi modificada em 4 semitons (2, -2, -1, 1).

Dessa forma, a biblioteca Librosa foi usada em sua atualização mais recente para o Python, uma vez que cada amostra foi calculada, foram rotuladas das anotações originais.

Depois que todas as amostras com mais de 2,95 segundos foram filtradas, o conjunto de dados original de 8732 amostras foi reduzido para 7479 dessa maneira e, como explicado nesta seção, a classe mais afetada foi o "gun shot", que foi reduzido de 374 para 34 amostras. Depois que o processo de aumento de dados foi aplicado, essa classe foi aumentada para 244 amostras (aplicando TS e PS em 30 amostras e mantendo 4 para adicionar ao conjunto de testes). O processo de normalização de cada pixel individual é realizado para cada uma das amostras (M / 255).

ANÁLISE I:

O primeiro análise do modelo foi realizada dividindo o conjunto de dados em 7000 amostras para treinamento (90%) e validação (10%) e o restante 479 para teste, a fim de verificar como o modelo aprendeu com um conjunto de amostras tiradas aleatoriamente de cada fold, sem adicionar os dados da classe aumentada e para cada uma das duas representações de espectro-tempo (Mel-Spectrogram e MFCCs). Dessa maneira, a **Figura 1** mostra a matriz de confusão obtida para a representação baseada em Mel-Spectrogram e a **Figura 2** mostra a matriz de confusão para MFCCs.

Mel-Spectrogram

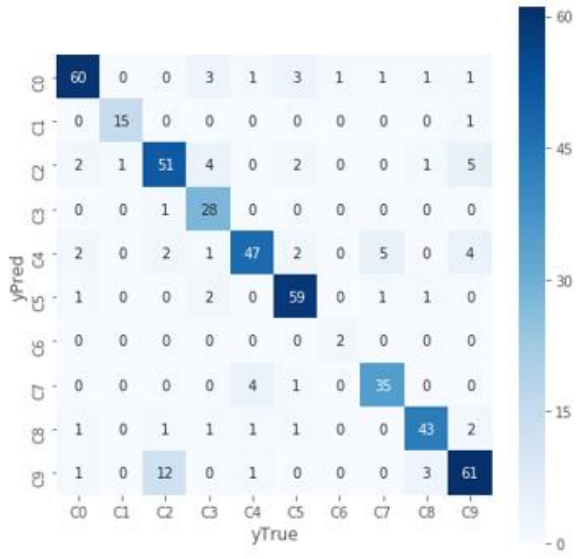

Figura 1. Matriz de Confusão Mel-Spectrogram.

MFCCs

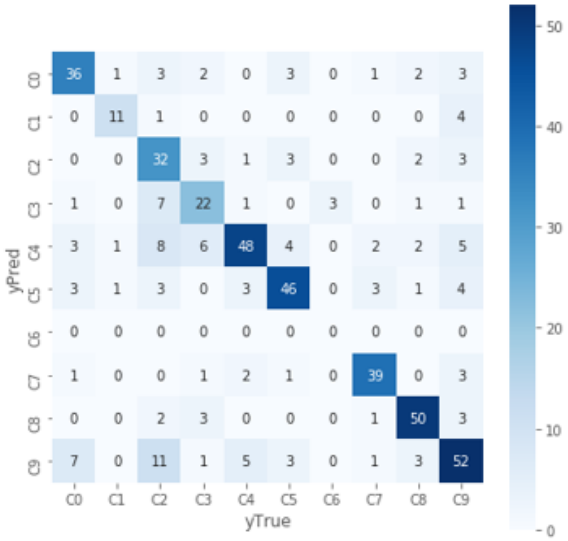

Figura 2. Matriz de Confusão MFCCs.

Ambos representam o melhor desempenho obtido para diferentes conjuntos de treinamento e teste, com base nas duas abordagens propostas. A acurácia medida neste caso para o Mel-Spectrogram foi de 83,7% e para os MFCCs de 70,1%.

Assim, conforme mostrado na Tabela 2 para a Análise I, **Acc0** representa o valor de "acurácia média" obtido para as duas representações antes de aplicar o aumento de dados e **Acc1** representa o valor de " acurácia média" obtido para as duas representações depois de aplicar o aumento de dados. As Figuras 3,4 e 5 mostram as métricas obtidas para as duas representações em cada uma das classes.

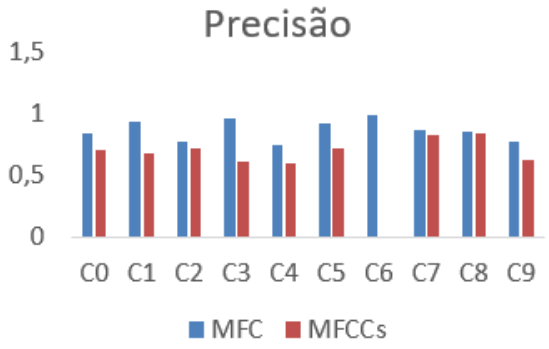

Figura 3. Precisão para MFC y MFCCs.

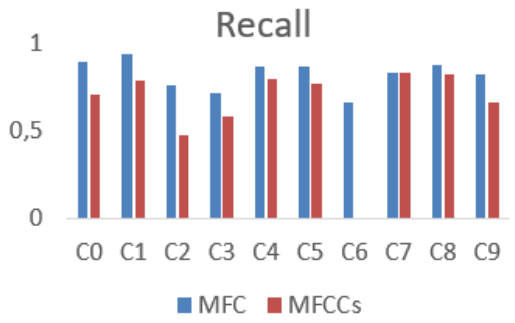

Figura 4. Recall para MFC y MFCCs.

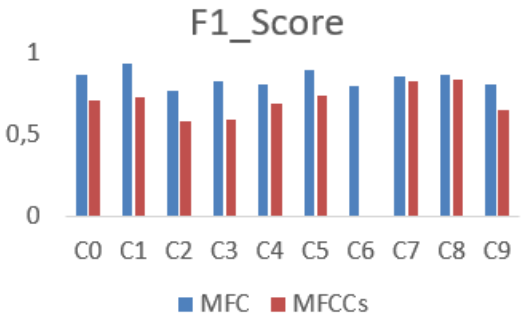

Figura 5. F1-Score para MFC y MFCCs.

ANÁLISE II:

Analisando o conjunto de dados dividido em 10 folds, pode-se ver como cada uma delas contém uma representação de cada classe, dessa maneira ao fazer validação cruzada usando o procedimento proposto na ANÁLISE I (Selecionar amostras aleatórias de cada fold), pode correr o risco de localize um número maior de amostras relacionadas nos conjuntos de treinamento e teste e, assim, altere os valores de precisão calculados.

Por esse motivo, para a segunda análise, cada uma das 10 folds foi dividida em conjuntos de treinamento, validação e teste. Usando no primeiro caso, o Fold1 até Fold9 como treinamento e validação e o Fold10 como Teste, repetindo o mesmo procedimento para os demais até usar as 10 folds como conjunto de treinamento e calculando a acurácia final como o valor médio medido entre todos os testes feitos. Os valores médios obtidos são mostrados na **Tabela 2**, antes e depois da aplicação do aumento de dados.

Tabela 2. Valores de análise

	ANÁLISE I		ANÁLISE II	
	Acc0	**Acc1**	**Acc0**	**Acc1**
MFC	0.804	0.833	0.758	0.761
MFCCs	0.695	0.712	0.667	0.671

Usando UrbanSound8k, podemos comparar os resultados deste estudo com abordagens publicadas anteriormente que foram avaliadas nos mesmos dados. O modelo convolucional proposto por PiczakCNN [7] possui 2 camadas convolucionais seguidas por 3 camadas densas, a rede opera em 2 canais de entrada: log do Mel-Spectrogram e seus deltas. O valor de acurácia obtido foi de 0,73 calculado como a acurácia média de todos os folds, sem usar o aumento de dados. Para este estudo sem aumento de dados, o modelo implementou uma acurácia de 0,758. Embora a diferença não seja tão pronunciada, elas podem estar relacionadas ao número de camadas convolucionais e à quantidade de bandas **mel** usadas em cada caso. O PiczakCNN usa 60 bandas e neste experimento 128. Analisando também que a quantidade de dados submetidos a treinamento e teste pode ter diferenças.

Como pode ser verificado para cada uma das duas análises descritas, as abordagens baseadas nos MFCCs relatam métricas inferiores às relatadas pelo MFC-Spectrogram.

3. CONCLUSÕES

Neste projeto, uma arquitetura de rede neural convolucional profunda foi proposta para a classificação de sons ambientais. A partir deste estudo, a acurácia da classificação foi avaliada por duas abordagens diferentes (MFC e MFCC), nas quais se observou que, a partir da representação baseada no MFC-Spectrogram, o modelo forneceu melhores métricas.
Verificou-se a influência do aumento de dados na Precisão do modelo para as classes menos representadas e concluiu-se que cada tipo de classe é influenciado diferentemente por cada conjunto de aumento, o que pode significar que o desempenho do modelo pode ser aprimorado mesmo mais aplicando o aumento de dados em cada classe.

Dessa forma, foi possível verificar que técnicas baseadas em redes neurais profundas podem ser um método eficaz para a classificação de sons ambientais, mesmo quando são influenciados por ruídos externos.

4. REFERÊNCIAS

[1] H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” in Proceedings of the IEEE

[2] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, “Robust Environmental Sound Recognition for Home Automation,” in IEEE Transactions on Automation Science and Engineering, vol. 5, no. 1, pp. 25–31, Jan. 2008.

[3] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, “Robust Sound Event Classification Using Deep Neural Networks,” in IEEE/ACM. Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 540–552, Mar. 2015.

[4] L. Lu, H.-J. Zhang, and S. Z. Li, “Content-based audio classification and segmentation by using support vector machines,” in Multimedia Systems, vol. 8, no. 6, pp. 482–492, Apr. 2003.

[5] M. M. Mostafa and N. Billor, “Recognition of Western style musical genres using machine learning techniques,” in Expert Systems with Applications, vol. 36, no. 8, pp. 11378–11389, Oct. 2009.

[6] Z. Zhang and B. Schuller, “Semi-supervised learning helps in sound event classification,” in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 2012, pp. 333–336.

[7] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” in Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1–6.

[8] D. Stowell and M. D. Plumbley, “Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning,” PeerJ, vol. 2, pp. e488, Jul. 2014.

[9] S. Dieleman and B. Schrauwen, “Multiscale approaches to music audio feature learning,” in 14th ISMIR, Curitiba, Brazil, Nov. 2013.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[11] S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” JASA, vol. 8, no. 3, pp. 185–190, 1937.

[12] T. Ganchev, N. Fakotakis, and G. Kokkinakis, “Comparative evaluation of various MFCC implementations on the speaker verification task,” in 10th Int. Conf. on Speech and Computer, Greece, Oct. 2005, pp. 191–194.

[13] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in 22nd ACM International Conference on Multimedia(ACM-MM’14), Orlando, FL, USA, Nov. 2014, pp. 1041–1044

[14] B. McFee, E. Humphrey, and J. Bello, “A software framework for musical data augmentation,” in 16th Int. Soc. for Music Info. Retrieval Conf., Malaga, Spain, Oct. 2015, pp. 248–254.