Combate à caça ilegal de mamíferos usando SED

Davi F. Henrique[1], **Mariana M. Blume**[1], **Aline Duarte Riva**[1]

[1]Universidade La Salle
Canoas – RS – Brasil

Abstract. *In order to combat illegal hunting activities, there are approaches that use machine learning to formulate better patrol strategies. As a complementary way, we propose a model using neural networks for the detection of potentially illegal activities, through the processing of the sound captured in regions where there is inhabiting fauna.*

Resumo. *Para o combate às atividades ilegais de caça, existem abordagens que se utilizam de machine learning para formular melhores estratégias de patrulha. Como forma complementar, propomos um modelo utilizando redes neurais para a detecção de atividades potencialmente ilegais, através do processamento do som capturado em regiões onde existe fauna habitante.*

1. Introdução

Cresce em um número alarmante a caça de diferentes espécies de animais ameaçadas de extinção em razão do comércio no mercado ilegal que por sua vez é estimado em ter uma circulação de 8 a 10 bilhões de dólares por ano. No período de 1930 a 1960, houve uma movimentação de 500 milhões de dólares (cotação de 2015) no comércio das 10 principais espécies de animais procuradas no Brasil, conforme a figura 1 desenvolvida por [Antunes et al. 2016]. Embora o Brasil tenha banido oficialmente a caça em 1967, devido à existência de brechas que permitiam o comércio de peles que estavam armazenadas, facilitou a caça ilegal e a exportação até o momento da ratificação da Convenção sobre Comércio Internacional das Espécies da Flora e Fauna Selvagens em Perigo de Extinção (CITES) [Antunes et al. 2016].

Para combater as atividades ilegais de caça, a proposta de utilizar inteligência artificial e *machine learning* se demonstra efetiva. Há trabalhos publicados abordando formas mais eficazes de combate a esta prática, como a patrulha da área onde se encontra a fauna como o descrito algoritmo *LIZARD* [Xu et al. 2020a]. No campo de estudos de prevenção à caça ilegal, um dos pilares para sua efetividade consiste em uma estratégia de patrulha, com rotas que maximizem o número de armadilhas removidas e propiciem o encontro com caçadores [Xu et al. 2020b]. A utilização de câmeras e microfones para a captura de vídeo e áudio, para o monitoramento dos locais protegidos, consiste em uma das práticas realizadas por centros de preservação.

Como forma complementar ao patrulhamento, utilizaremos redes neurais para detectar atividades de caça através do processamento do áudio capturado de regiões onde existe atividade de fauna habitante. O modelo será treinado com uma base de sons de animais e de armas de fogo utilizadas em sua captura e abate. Desta forma, será possível a geração de alertas para uma análise humana dos fatos ocorridos, capturando possíveis atividades ilegais, semelhante à abordagem proposta por [Fulzele et al. 2020].

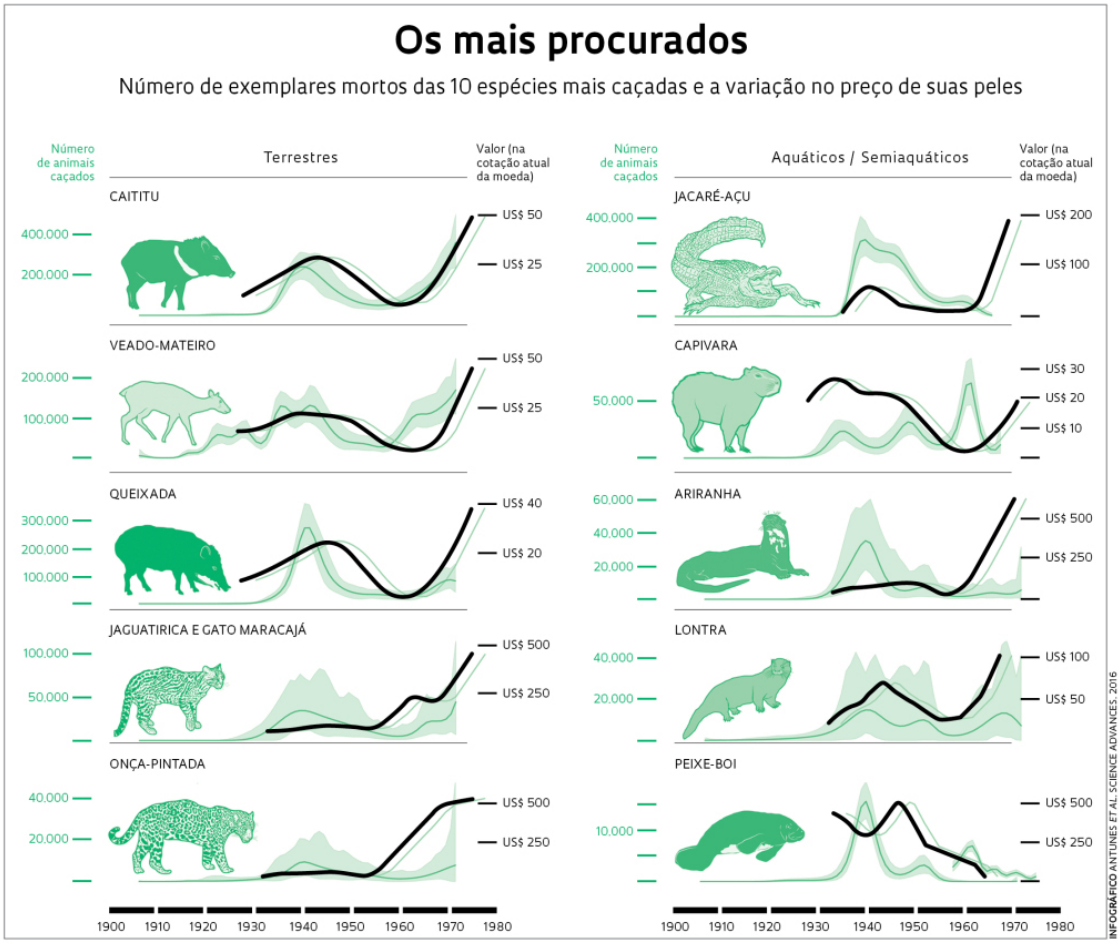

Figura 1. As espécies mais procuradas para caça no Brasil, no período de 80 anos

2. Detecção de eventos em áudio

Um sistema de detecção de eventos em áudio (SED) pode ter vantagens consideráveis, comparado a outros sistemas que realizam a classificação de imagens e vídeos. Devido ao áudio não ter luminosidade, pode-se aplicá-lo mais efetivamente a ambientes noturnos. Se houver objetos oclusos no ambiente, o resultado da detecção do som e sua precisão não serão afetados. Certos tipos de eventos são mais facilmente detectados via ondas sonoras, como o uso de armas de fogo. A computação de áudio utiliza menos recursos, comparada a de imagens e vídeos. O SED pode ser aplicado, além do reconhecimento de fala, em campos como o monitoramento da vida selvagem [Chan and Chin 2020].

Atualmente as implementações de SED que utilizam aprendizado profundo (*deep learning*), possuem em suas arquiteturas três componentes. O primeiro componente tem como principal funcionalidade extrair as características do som, com a finalidade de diminuir a dimensionalidade representativa do mesmo, geralmente implementado usando redes neurais convolucionais (CNN). O segundo componente modela longos contextos de tempo, advindos do primeiro componente, para a identificação de padrões entre classes sonoras sendo geralmente implemen-

tado utilizando redes neurais recorrentes (RNN). O terceiro e último componente, implementado usando função afim, realiza a classificação da amostragem sonora [Drossos et al. 2020].

Como forma de insumo para a detecção de atividades de armas de fogo ou humanas, utilizamos como base o modelo proposto por [Drossos et al. 2020]. Tendo visto que há uma redução no número de parâmetros utilizados pelo modelo construído, tornando-o mais viável para a utilização em um sistema embarcado onde há restrições de recursos, possibilitando um processamento do som junto ao dispositivo que o captura. Por haver uma quantidade menor de parâmetros, o tempo de treinamento diminui facilitando a verificação do modelo.

3. Detecção de atividades de caça

Para detectar atividades de caça utilizando como insumo o SED, utilizaremos o modelo InceptionTime, proposto por [Ismail Fawaz et al. 2020] para a classificação de *time series*, uma vez que o modelo por ser mais escalável, aprende a reconhecer as classes dos dados em um menor tempo.

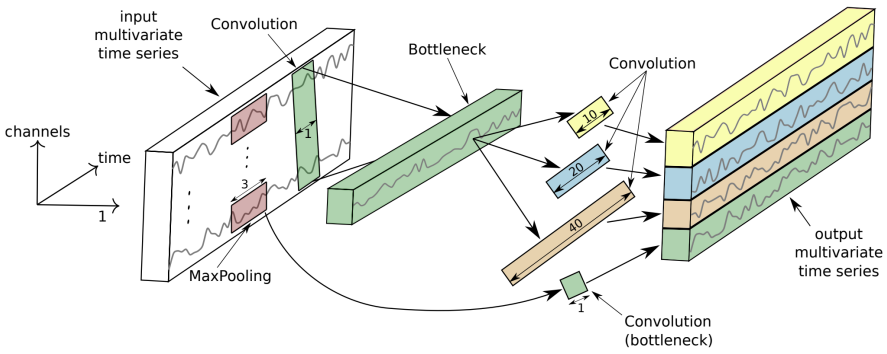

Figura 2. Estrutura do modelo InceptionTime

4. Base de dados utilizadas

Dentre as bases de dados utilizadas para o treinamento, selecionamos a [Piczak 2015] para as amostras de sons da natureza como chuva, trovão, pássaros. Utilizando conjuntamente aos sons de pássaros, disponível através do Xeno-Canto[1] os quais foram extraídos de forma automatizada[2]. Para a detecção do uso de armas de fogo foi utilizado o conjunto de amostras disponíveis pelo [Dogan 2021], o qual contém 2310 amostras de disparos de armas de fogo, munido ao publicado *dataset* de armas [Lilien 2018]. Para a detecção de fala se utilizou o Commons Voice Corpus 6.1 [Ardila et al. 2020]. Além do uso da base Urban8K [Salamon et al. 2014] para as amostras de sons produzidos por atividade humana, como sons de motores em funcionamento, sirenes, buzinas de carro.

[1]https://xeno-canto.org
[2]https://github.com/AgaMiko/xeno-canto-download

Tabela 1. Amostras sonoras utilizadas

Nome	Quantidade
Armas de fogo	5532
Atividades de caça (armas de fogo)	548
Pássaros cantando	3262
Conversas	752
Motor	1000
Fogos de artifício	40
Chuva	1841
Sirene	929
Buzina de carro	429
Motosserra	40
Trovão	40
Helicóptero	40
Vento	1241
Total	15694

4.1. Construção da base de dados de caça

Utilizou-se como base de dados os vídeos publicamente disponíveis no YouTube, onde há a utilização de armas de fogo em atividades de caça. Construímos um processo semi-supervisionado para fazer a classificação das amostras para a construção do *dataset*. Para tal, particionamos os vídeos em amostras de um tamanho máximo pré-definido, onde em cada uma destas, executamos o SED para obtermos a predição das categorias sonoras da amostra. Manualmente selecionamos as amostras onde há uma confiança pequena. Se houver uma maior confiança, as amostras são automaticamente adicionadas à base de dados.

Um total de 548 amostras sonoras foram coletadas, dentre vídeos de caça de Coiotes (*Canis latrans*), Javalis (*Sus scrofa*), Lince-Pardo (*Lynx rufus*). Construímos um *dataset* contendo todas as amostras em formato MP3 devido aos vídeos inicialmente utilizarem compressão, não sendo possível recuperar a sua qualidade original. Disponibilizamos este *dataset* na plataforma Zenodo por uma licença que permite o livre uso [Henrique 2021].

4.2. Pré-processamento das amostras sonoras

Para reduzir a representação dimensional do som, preservando suas principais características, para todas as amostras sonoras se utilizou da representação em espectrograma de Mel. A escala de Mel se assemelha a como os seres humanos percebem as frequências sonoras, sendo uma escala logarítmica, por sua vez não linear [Purwins et al. 2019]. Sendo amplamente utilizada no campo de representação das características do som em tarefas relacionadas a análise de áudio. Utilizando uma configuração de 64 filtros para projetar as frequências no espectro.

Adotada a taxa de amostragem de 22050 Hz com a fórmula de Hidden Markov Toolkit para a conversão entre Hz e Mel, com a utilização de janelas de Hamming. Utilizando de valores próximos aos reportados pelo estudo sobre redução de

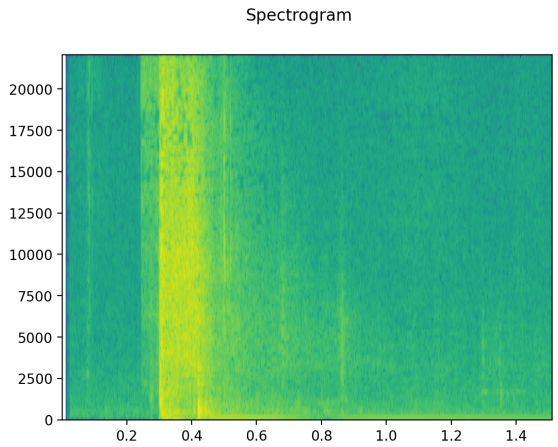

Figura 3. Espectrograma de um disparo de arma de fogo

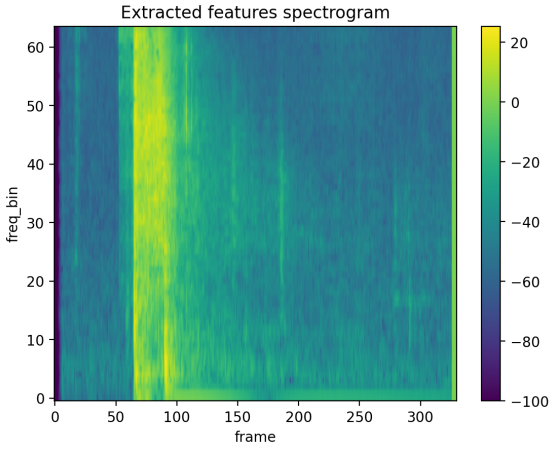

Figura 4. Espectrograma de Mel de um disparo de arma de fogo

frequência e resolução de tempo na classificação de música em arquiteturas de redes neurais convolucionais [Ferraro et al. 2019].

A figura 3 demonstra um espectrograma de um disparo de arma de fogo em níveis de frequência e duração do som, a figura 4 ilustra o resultado da aplicação do processamento sonoro descrito.

5. Treinamento do modelo

O treinamento do modelo ocorreu em duas etapas, sendo a primeira o aprendizado do SED com as amostras coletadas, e a segunda o modelo para a detecção de atividades de caça. As amostras sonoras foram separadas em 3 grupos, tendo o grupo de treinamento 70%, o grupo de teste 20% e o grupo de validação 10%. Os modelos foram treinados utilizando o otimizador de Adam com *binary cross-entropy loss*. Sendo executados em uma NVIDIA GeForce GTX 1660 TI.

Para o treinamento do SED utilizou-se o modelo separável em profundidade com RNN (*Depth-wise separable with RNN*) [Drossos et al. 2020] com *batches* de tamanho 14, tendo a RNN uma dimensão de entrada de 2048, contendo 15 classes sonoras.

No modelo InceptionTime se utilizou de *batches* de tamanho 14, 1 canal de *bottleneck*, 3 blocos, 64 canais de entrada, 2 canais de saída e um kernel de tamanho 8. Tendo menos amostras utilizadas durante o treinamento, apenas o grupo de atividades de caça, vento, pássaros e sons da natureza.

6. Resultados

Utilizamos a abordagem descrita em [Drossos et al. 2020] para calcular a precisão do modelo SED e InceptionTime. Os resultados foram quantificados na tabela 2 utilizando dois parâmetros, F_1 para a taxa de frames corretos e ER para a taxa de erro.

Apesar da alta precisão dos modelos no domínio das bases de dados, em testes exploratórios com outras amostras demonstraram algumas incongruências com a

Tabela 2. Resultados observados

Nome	F_1	ER
DESSED	0.97	0.05
InceptionTime	1.00	0.00

porcentagem descrita. Comportamento ilustrado pela figura 5, na predição de um vídeo de caça disponível no YouTube.

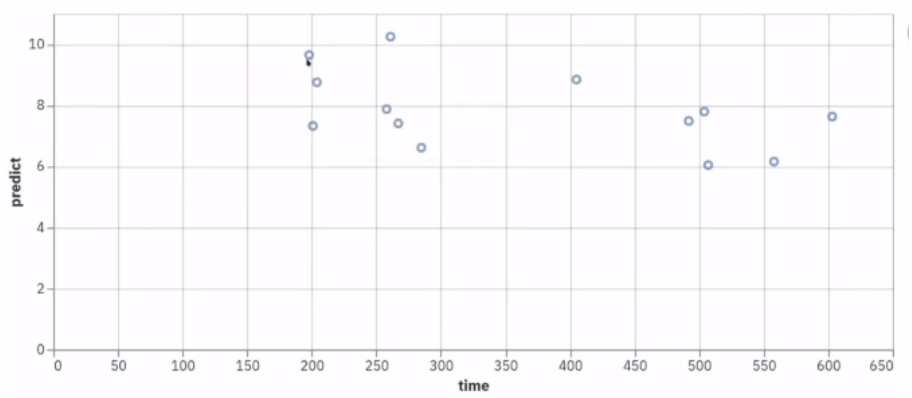

Figura 5. Predição do modelo sobre um vídeo de caça

7. Conclusão

A performance do modelo SED se demonstrou efetiva na detecção do uso de armas de fogo, com a diversidade das amostras pode-se ter uma confiança maior do modelo sobre amostras em domínios diferentes dos apresentados. Poderia se ter utilizado do modelo SED separável em profundidade com convoluções dilatadas proposto por [Drossos et al. 2020], possivelmente haveria resultados diferentes.

Para a detecção de atividades de caça, a classificação de *time series* demonstrou grande precisão em prever os eventos. Pelo conjunto de amostras coletadas de atividades de caça é possível ter uma previsibilidade da performance do modelo em exemplos reais. Devido à distribuição desigual das amostras sonoras do dataset formado, houve um desbalanceamento entre as categorias sonoras o que pode ser observado durante os testes, pela taxa de erro maior em outros domínios.

Em futuros trabalhos, seria possível de utilizar o sistema proposto por [Liang et al. 2019] para ter uma previsibilidade da distância do atirador ao local onde o som está sendo capturado. Possibilitando uma assertividade maior sobre a possível área onde ocorreu o incidente, facilitando a verificação dos fatos ocorridos. A utilização de um sistema multi agente neste cenário pode trazer uma maior sinergia com outros sistemas para a construção de um sistema robusto para o combate de atividades ilegais.

Referências

Antunes, A. P., Fewster, R. M., Venticinque, E. M., Peres, C. A., Levi, T., Rohe, F., and Shepard, G. H. (2016). Empty forest or empty rivers? A century of commercial hunting in Amazonia. *Science Advances*, 2(10):e1600936.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus.

Chan, T. K. and Chin, C. S. (2020). A comprehensive review of polyphonic sound event detection. *IEEE Access*, 8:103339–103373.

Dogan, S. (2021). A new fractal h-tree pattern based gun model identification method using gunshot audios. *Applied Acoustics*, 177:107916.

Drossos, K., Mimilakis, S. I., Gharib, S., Li, Y., and Virtanen, T. (2020). Sound Event Detection with Depthwise Separable and Dilated Convolutions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, Glasgow, United Kingdom. IEEE.

Ferraro, A., Bogdanov, D., Jeon, J. H., Yoon, J., and Serra, X. (2019). Music auto-tagging using cnns and mel-spectrograms with reduced frequency and time resolution. *CoRR*, abs/1911.04824.

Fulzele, V., Kulkarni, Y., and Aras, S. (2020). Conservation of wildlife from poaching by using sound detection and machine learning. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY*.

Henrique, D. (2021). Human animal hunting audio dataset.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., and Petitjean, F. (2020). Inception-time: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962.

Liang, J., Aronson, J. D., and Hauptmann, A. (2019). Technical Report of the Video Event Reconstruction and Analysis (VERA) System – Shooter Localization, Models, Interface, and Beyond. *arXiv:1905.13313 [cs]*. arXiv: 1905.13313.

Lilien, R. (2018). Development of computational methods for the audio analysis of gunshots.

Piczak, K. J. (2015). ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S., and Sainath, T. N. (2019). Deep learning for audio signal processing. *CoRR*, abs/1905.00078.

Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA.

Xu, L., Bondi, E., Fang, F., Perrault, A., Wang, K., and Tambe, M. (2020a). Dual-Mandate Patrols: Multi-Armed Bandits for Green Security. *arXiv:2009.06560 [cs, stat]*. arXiv: 2009.06560.

Xu, L., Gholami, S., McCarthy, S., Dilkina, B., Plumptre, A., Tambe, M., Singh, R., Nsubuga, M., Mabonga, J., Driciru, M., Wanyama, F., Rwetsiba, A., Okello, T., and Enyel, E. (2020b). Stay Ahead of Poachers: Illegal Wildlife Poaching Prediction and Patrol Planning Under Uncertainty with Field Test Evaluations (Short Version). In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1898–1901, Dallas, TX, USA. IEEE.