Detecção de Sentimento Utilizando *Deep Learning* para Auxílio em Análise Forense

Felipe M. S. Maria[1], **Thiago M. Ventura**[1]

[1]Instituto da Computação – Universidade Federal de Mato Grosso (UFMT)
Av. Fernando Corrêa da Costa, nº 2367 - Boa Esperança. Cuiabá - MT - 78060-900

felipe.masao@gmail.com, thiago@ic.ufmt.br

Abstract. *Screams and moans are a specific way of communicating and carry a feeling behind the sound player. Machine Learning techniques were used to detect feelings in various sounds produced by people. It was developed a model using Deep Learning to do this classification achieving a precision of 93,5%.*

Resumo. *Gritos e gemidos são uma forma específica de se comunicar e carregam um sentimento por trás de quem reproduz o som. Foi utilizado técnicas de Machine Learning para fazer a detecção de sentimento nos mais diversos sons produzidos por pessoas. Foi desenvolvido um modelo de Deep Learning para realizar essa classificação, atingindo uma precisão de 93,5%.*

1. Introdução

Uma forma de comunicação humana se dá através de gestos, posturas, gritos e grunhidos. Com o tempo, essa comunicação adquiriu formas mais claras e evoluídas [Machado 2019]. Com isso, é possível afirmar que sentimentos podem ser detectado por meio de expressões faciais e sonoras, produzidas por um indivíduo.

Desta forma um assunto em destaque para estudos na área da computação [Devillers et al. 2005]. Trabalhos como de [Nazir et al. 2018] ou [Karamizadeh and Arabsorkhi 2018] foram encontrados buscando esclarecer essa visão e também contribuíram para a motivação deste trabalho.

Técnicas de *Deep Learning*, ou Redes Neurais Artificiais Profundas, auxiliam neste tipo de trabalho, permitindo o aprendizado automático de vários níveis de representações da distribuição subjacente dos dados a serem modelados [Lauzon 2012]. Um exemplo é o trabalho de [Laffitte et al. 2016], que utilizando técnicas de *Deep Learning* conseguiu classificar automaticamente sons para um sistema de vigilância. Trabalho similar pode ser aplicado para a análise forense, no qual uma grande quantidade de dados deve ser avaliada em dispositivos periciados, tomando muito tempo de um perito caso tenha que ser feito manualmente.

Assim, o objetivo deste artigo é a construção de um modelo de *Deep Learning* capaz de realizar uma análise em sons em busca de sinais de interesse, como alegria, medo, raiva ou prazer, para auxiliar no exame de dispositivos periciados.

2. Materiais e Métodos

Em classificação de imagens em vídeos, os áudios são ignorados. Caso a imagem tenha baixa qualidade, a detecção pode falhar. Mas se o áudio também for analisado, há uma probabilidade maior de detecção das cenas de interesse.

É possível realizar classificação de áudios por meio de espectrogramas, que são imagens geradas pela transformação do sinal de áudio. A Figura 1 mostra o espectrograma de um exemplo de áudio pornográfico. Para conseguir classificar os sons baseados em espectrogramas, são utilizadas redes convolucionais, já que a convolução é comumente utilizada em imagens. O uso da Rede Neural Convolucional (CNN) aplica filtros em dados visuais, mantendo a relação de vizinhança entre os pixels da imagem ao longo do processamento da rede. Esse tipo de rede vem sendo amplamente utilizado, principalmente nas aplicações de classificação, detecção e reconhecimento em imagens ou vídeos [Vargas et al. 2016].

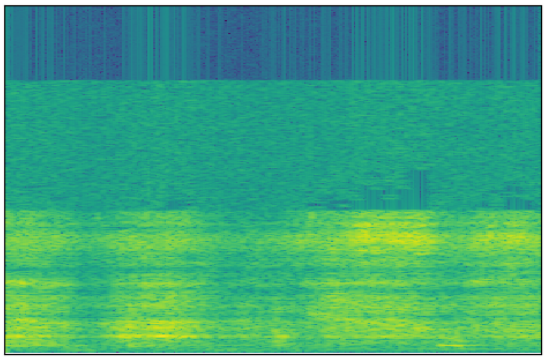

Figura 1. Espectrograma gerado de um áudio pornográfico.

A base de dados deste trabalho foi composta de áudios extraídos de fragmentos de vídeos. Esses dados foram normalizados, mantendo todos os áudios com 1 segundo de duração. A base de dados foi tratada e gerou uma base normalizada com 663 imagens, sendo: 35 de felicidade; 113 de medo; 140 de raiva; 375 de prazer.

Para o uso do CNN foi utilizada a programação com a linguagem *Python*, contendo pacotes de pré-processamento, criação de modelos e análise dos resultados. O framework *Keras* [Chollet et al. 2015] foi utilizado para a criação do modelo. O modelo criado neste trabalho segue uma estrutura que ao todo consiste em 23 camadas, onde 6 são do tipo Conv2D, 3 são do tipo MaxPooling, 2 do tipo Dense e 1 do tipo Flatten. Após cada camada do tipo Conv2D há uma camada de ativação e para cada camada do tipo MaxPooling foi adicionada uma camada de Dropout.

3. Resultados

Após o tratamento dos dados e configuração do modelo, foi possível treinar a rede neural convolucional. O modelo foi treinado com 100 épocas. O resultado do treinamento pode ser visto na Figura 2. Analisando o gráfico pode ser visto que a partir da 40ª época, a rede neural se estabilizou, sofrendo reduções na taxa de variação". A precisão na última época foi de 93,5%.

4. Considerações Finais

O objetivo deste artigo foi desenvolver uma rede neural com aprendizado profundo que fosse capaz de classificar um sentimento no som produzido por um humano, seja por fala, gritos ou gemidos. Para os dados testados o modelo conseguiu uma precisão superior à 90%, mostrando potencial para evolução e possibilidade de aplicação prática.

Como trabalhos futuros, pretende-se buscar uma forma de classificar vários sentimentos em vídeos de comprimentos maiores. Não somente, buscar um classificador que tente diferenciar o sentimento transmitido pela pessoa e diga a faixa etária. Assim, pode-se separar conteúdos pornográficos entre pedofilia ou não.

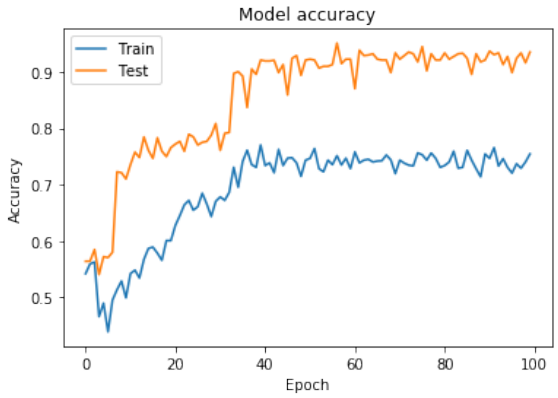

Figura 2. Gráfico baseado na precisão do modelo.

Referências

Chollet, F. et al. (2015). Keras. `https://keras.io`. [Online; acessado 20-Julho-2019].

Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

Karamizadeh, S. and Arabsorkhi, A. (2018). Methods of pornography detection. In *Proceedings of the 10th International Conference on Computer Modeling and Simulation*, pages 33–38. ACM.

Laffitte, P., Sodoyer, D., Tatkeu, C., and Girin, L. (2016). Deep neural networks for automatic detection of screams and shouted speech in subway trains. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464. IEEE.

Lauzon, F. Q. (2012). An introduction to deep learning. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1438–1439. IEEE.

Machado, G. M. (2019). História da Comunicação Humana. `https://www.infoescola.com/historia/historia-da-comunicacao-humana/`. [Online; acessado 15-Setembro-2019].

Nazir, S., Awais, M., Malik, S., and Nazir, F. (2018). A review on scream classification for situation understanding. *International Journal of Advanced Computer Science and Applications*, 9(8).

Vargas, A. C. G., Carvalho, A. M. P., and Vasconcelos, C. N. (2016). Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. In Cappabianco, F. A. M., Faria, F. A., Almeida, J., and Körting, T. S., editors, *Electronic Proceedings of the 29th Conference on Graphics, Patterns and Images (SIBGRAPI'16)*, São José dos Campos, SP, Brazil.