DEEP LEARNING APLICADA À ANÁLISE FACIAL PARA AUXILAR NO COMBATE À PORNOGRAFIA INFANTIL

JOSÉ ALBERTO SOUZA PAULINO (UFCG, Campus Campina Grande)

E-mails: souzapaulino@gmail.com.

Área de conhecimento: 1.03.03.04-9 (Sistemas de Informação).

Palavras-Chave: redes neurais convolucionais profundas; detecção de faces; combate à pornografia infantil.

1 Introdução

A classificação automática da faixa etária de um indivíduo por meio de reconhecimento facial pode ser útil em várias situações, como por exemplo: analisar possíveis irregularidades na compra de bebidas alcoólicas por adolescentes, verificar idade aproximada de motoristas, controle de conteúdo na internet, além de inúmeras outras aplicações. Esta análise facial pode ser realizada em tempo real ou em mídias já armazenadas, como vídeos em um computador.

Um problema grave e latente em nossa sociedade é o consumo de conteúdo pornográfico envolvendo violência e exploração sexual de crianças e adolescentes (IWF, 2020). Segundo relatório anual da IWF, em 2020 foram identificadas 132.676 URLs contendo imagens de abuso sexual infantil (IWF, 2020). Neste aspecto, várias ações vêm sendo adotadas objetivando coibir a produção de conteúdo desta natureza, além de buscar identificar e punir os produtores e consumidores deste conteúdo. No entanto, ainda é difícil realizar uma análise inteligente e acurada destas mídias, haja visto que se constituem de dados não estruturados como vídeos e fotografias, que por sua natureza despendem custo exacerbado de tempo na eventual necessidade de uma análise, que na maioria das vezes ocorre de forma convencional.

Redes sociais e serviços de *streaming* começaram a adotar mecanismos inteligentes para análise de dados não estruturados, de modo que conteúdos classificados como impróprios, contendo nudez ou violência, dificilmente são publicados, a exemplo de: Facebook, Youtube ou Instagram. E, mesmo que sejam publicados, estes conteúdos não permanecem online por muito tempo. No entanto, em sites próprios para conteúdo adulto, principalmente de cunho sexual, no qual os usuários podem fazer *upload* de arquivos livremente, a filtragem é menos restritiva e acabam sendo porta de entrada para a publicação e consumo de material com pornografia infantil e retroalimentando redes de pedofilia.

Dentro do contexto ora apresentado, esta pesquisa se propõe a desenvolver um modelo classificador capaz de identificar faces de crianças e adolescentes em fotografias ou *frames* de vídeos por meio de aprendizagem profunda.

2 Materiais e Métodos

O estado da arte na classificação de imagens evidencia a superioridade de redes neurais convolucionais profundas (*deep convolutional neural network* ou DCNN) frente à outras técnicas de classificação. Ao longo da última década, diversas arquiteturas de DCNN vêm sendo desenvolvidas. Nesta perspectiva, para o problema abordado nesta pesquisa, adotamos a VGG19, que é um tipo DCNN, que contém 19 camadas das quais 16 são usadas para extração de características e 3 são usadas para classificação, conforme descrito por Zhou et al. (2020).

Apesar do grande potencial da rede VGG19 na extração de características, o problema de identificação de faixa etária em faces de indivíduos não é uma tarefa trivial. Por mais que existam traços característicos em cada uma destas faixas etárias, estes traços são sutis e definir se um indivíduo é adolescente ou adulto, por exemplo, torna-se difícil até mesmo para um ser humano. Desta forma, para que sejam criados padrões que consigam caracterizar cada classe de imagens (faixa etária), são necessários centenas de milhares de exemplo. Estas bases de dados nem sempre estão disponíveis e no desenvolvimento desta pesquisa, para contornar este impeditivo técnico referente ao número limitado de imagens, foi utilizada a transferência de aprendizagem.

A transferência de aprendizagem é uma técnica que consistem em usar uma rede neural treinada em outras bases de dados e preservar este “conhecimento” já adquirido para aplicá-lo em novos problemas. Ao usar esta técnica é possível resolver problemas complexos de classificação com um número substancialmente menor de exemplos.

A base de dados adotada é composta de 3.773 imagens obtidas no Google Imagens, contendo 1.812 imagens de pessoas classificadas como adultos e idosos (**classe 1**) e 1.961 imagens de pessoas definidas como crianças e adolescentes (**classe 0)**. A base de dados foi dividida de forma estratificada em dois subconjuntos: subconjunto de treinamento com, 80% das imagens, e subconjunto de testes, com 20% das imagens.

A Figura 1 mostra exemplos da base de dados utilizada. Na Figura 1**(a)**, estão algumas amostras da classe adultos e na Figura 1**(b)**, as amostras da classe com crianças e adolescentes.

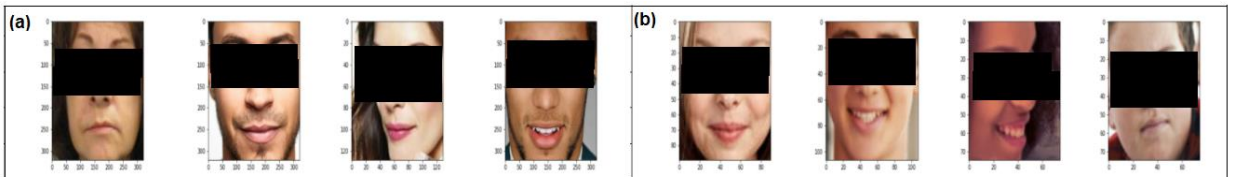

Figura 1: Exemplo de imagens da base de dados utilizada

Sobre essa base de dados, vale destacar dois aspectos importantes: (1) foi adicionada uma tarja no rosto das pessoas em todas as figuras deste artigo pois, apesar de se tratar de imagens disponíveis livremente no Google Imagens, busca-se, assim, minimizar a exposição das pessoas sem autorização prévia; (2) foram adotadas intencionalmente imagens do Google Imagens por não fazer parte de nenhum conjunto de dados com um protocolo rígido de aquisição de imagens. Desta forma, foram usadas imagens com variação no brilho, tonalidade, saturação, resoluções e formatos. O objetivo desta estratégia foi ampliar a capacidade de generalização e robustez da solução desenvolvida.

Foi adotada a linguagem Python para o desenvolvimento e as bibliotecas Tensorflow/Keras (ADIBI, 2016) e Sklearn para a criação do modelo. O fluxo do processo de predição está ilustrado na Figura 2, na qual se tem como entrada a imagem obtida no Google Imagens, em seguida realiza-se a detecção automática do rosto e o resultado é enviado para a DCNN, que por sua vez obtém a predição: classe 0 equivale a um rosto de criança ou adolescente e classe 1 para adultos.

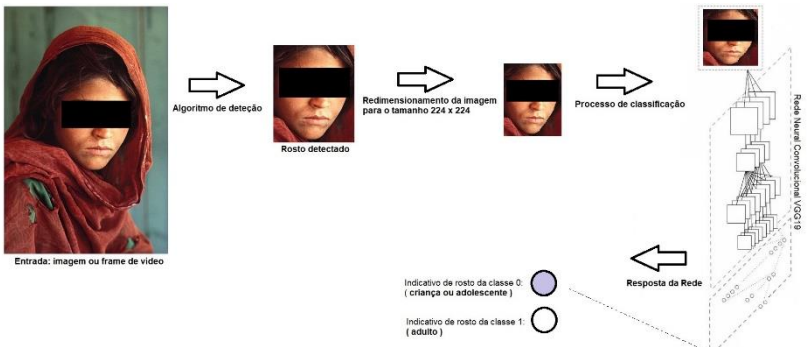

Figura 2: Processo de classificação

3 Resultados e Discussão

Conforme citado na metodologia, foi realizada uma classificação binária na qual a classe 0 corresponde ao grupo de imagens contendo o rosto de crianças e adolescentes e a classe 1 corresponde ao grupo de imagens com o rosto de pessoas adultas e idosas. Os valores obtidos pelo modelo desenvolvido no conjunto de testes, com 755 imagens, estão descritos na Tabela 1.

		Valor previsto		
		Classe 0	**Classe 1**	
Valor Real	**Classe 0**	346	46	392
	Classe 1	43	320	363

Tabela 1: Matriz de confusão para o conjunto de testes

De acordo com a matriz de confusão, foram obtidas as seguintes métricas de avaliação:

- Acurácia: 88,21%;
- Sensibilidade: 88,27%;
- Especificidade: 88,15%

Para alcançar a convergência durante o treinamento, o modelo precisou de pouco mais que 30 épocas. E após a convergência manteve uma acurácia média de 95% e erro de 0,10, conforme Figura 3, o que demonstra que não houve *overfitting* e é um indicativo que a melhora nos resultados pode ser obtida com o aumento da base de dados.

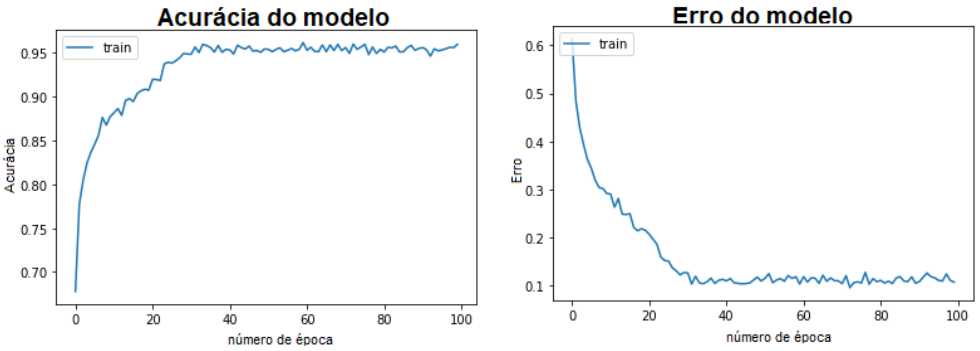

Figura 3: Gráficos de desempenho durante treinamento do modelo

Os testes com o modelo em cenas reais do cotidiano, demonstram a capacidade do modelo para classificação correta. A Figura 4 (a), mostra que o modelo conseguiu reconhecer e classificar corretamente, com 99,99 % de certeza, o rosto de um adolescente em uma fotografia de baixa resolução, inclusive, acompanhada de adultos. Já na Figura 4 (b), o modelo conseguiu classificar, com uma probabilidade de 97,77 % de certeza, em um *frame* de vídeo de um baile funk, o rosto que potencialmente corresponde a um adolescente.

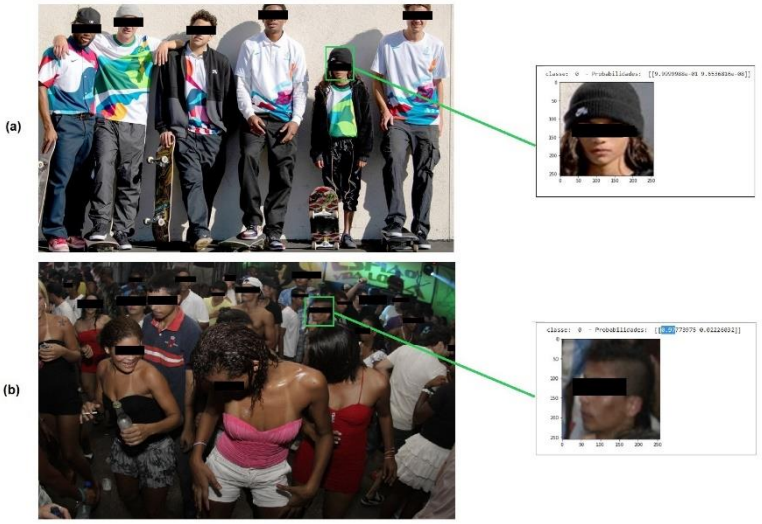

Figura 4: Exemplo de classificação em fotografia e *frame* de vídeo

4. Considerações Finais

Os resultados obtidos com o modelo desenvolvido nesta pesquisa demonstram a capacidade que as técnicas de inteligência artificial têm de analisar dados não estruturados. De modo que este modelo pode ser adotado como uma importante ferramenta no combate a publicação de conteúdo com pornografia infantil e na identificação de material pertencentes a redes de pedofilia. Pautados nos resultados ora apresentados, considera-se que o objetivo proposto foi plenamente atendido, definindo-se assim, como propostas futuras: desenvolvimento de uma ferramenta para análise dinâmica em vídeos, ampliação da base de dados e classificação em quatro classes: criança, adolescente, adulto e idoso.

Referências

ABADI, Martín et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. **arXiv preprint arXiv:1603.04467**, 2016.

IWF (INTERNET WATCH FOUNDATION). Internet Watch Foundation Annual Report 2020: Face the Facts. **[online].** p. 58-65, 2020. Disponível em: <<https://www.iwf.org.uk/report/iwf-2020-annual-report-face-facts>>.

ZHOU, Jianye et al. Multisignal VGG19 Network with Transposed Convolution for Rotating Machinery Fault Diagnosis Based on Deep Transfer Learning. **Shock and Vibration**, v. 2020, 2020.