1.03.99 - Ciência da Computação

DETECÇÃO DE PORNOGRAFIA EM DESENHOS ANIMADOS USANDO REDES NEURAIS PROFUNDAS

Akari Ishikawa[1]*, Mauricio Perez[2], Sandra Avila[1]
1. Estudante do Instituto de Computação da Universidade Estadual de Campinas (RECOD Lab.-IC-UNICAMP)
2. Estudante de Doutorado da Escola de Engenharia Elétrica e Eletrônica da Universidade Tecnológica de Nanyang (NTU, Singapura)
3. Professora do RECOD Lab.-IC-UNICAMP — Orientadora

Resumo

Inúmeras soluções foram propostas para detecção automática de conteúdo pornográfico em vídeos e imagens. Entretanto, pouquíssimas abordagens foram desenvolvidas para lidar com conteúdo sensível em desenhos animados. Neste trabalho, avaliamos como soluções estado da arte para vídeos naturais (com seres humanos) baseadas em redes neurais profundas se comportam quando aplicadas aos desenhos animados. Propomos também um novo método com uma maior acurácia, e mostramos que tratar animações separadamente de vídeos naturais pode melhorar a filtragem de conteúdos impróprios para crianças.

Palavras-chave: Conteúdo Sensível; Aprendizado Profundo; Computação Forense

Apoio financeiro: A. Ishikawa é parcialmente financiada pelo PIBIC/CNPq e da FAEPEX (#2555/18). S. Avila é parcialmente financiada pelo Google Research Awards for Latin America 2018, FAPESP (#2017/16246-0) e FAEPEX (#3125/17). Os autores agradecem o apoio da NVIDIA Corporation com a doação das GPUs Titan Xp.

Trabalho selecionado para a JNIC: UNICAMP

Introdução

Crianças que cresceram rodeadas pela tecnologia, os “nativos digitais”, passam muito tempo na Internet assistindo desenhos animados. De acordo com o Fundo das Nações Unidas para a Infância (UNICEF), crianças e adolescentes representavam em 2016 um terço dos usuários na Internet [1]. Diante disso, a filtragem de conteúdo de forma inteligente tornou-se primordial. Estima-se que 30% de todos os dados transferidos pela Internet sejam de conteúdo pornográfico [2].

As primeiras soluções para filtragem de mídias sensíveis foram baseadas em detecção de nudez [3]. Entretanto, essas abordagens sofrem com o alto número de falsos positivos (e.g., cenas de lutas de boxe, pessoas na praia). Para lidar com esse problema, a detecção de pornografia passou a ser tratada como uma tarefa de classificação e a abordagem baseada no modelo de *Bag of Visual Words* (BoVW) passou então a ser a mais explorada pela literatura [4-9]. No entanto, apesar de exibir altos valores de acurácia, os métodos baseados em BoVW exigem extração de características “feitas à mão” e, principalmente, não conseguem classificar classes mais complexas.

Nos últimos cinco anos, as arquiteturas de aprendizado profundo (DLA, *Deep Learning Architectures*) têm demonstrado precisões de mais de 97% [10]. É importante ressaltar que tais abordagens são treinadas e validadas em bases de vídeos contendo majoritariamente vídeos naturais (com seres humanos), sendo apenas uma pequena parcela de vídeos de animações. Assim, apesar dessas técnicas realizarem boa filtragem em vídeos naturais, os vídeos com conteúdo sensível — mascarados em animações infantis, presentes principalmente em plataformas como o YouTube — não são detectados. Para piorar a situação, pouquíssimos trabalhos na literatura estão voltados para detecção de conteúdo sensível em desenhos animados e para dispositivos móveis [11], que é a maior forma de acesso das crianças à Internet.

Neste trabalho, nós avaliamos a rede neural convolucional (CNN) treinada por um trabalho recente que atingiu 97.9% de acurácia em classificação de pornografia em vídeos naturais [10] e comparamos sua performance com um modelo explicitamente treinado para animações voltado para dispositivos móveis.

Metodologia

Base de Dados

Por sermos os primeiros a abordar pornografia em desenhos animados, a primeira etapa desse trabalho consistiu em criar uma base de vídeos de desenhos animados. Coletamos 544 (cerca 50 horas) vídeos não-sensíveis do YouTube de canais oficiais (e.g., Disney Channel, Cartoon Network) e 195 vídeos (cerca de 25 horas) de desenhos animados impróprios retirados de sites pornográficos (e.g., PornHub,

Xvideos). O desbalanceamento entre as classes foi proposital para tentar criar uma base representativa da realidade. Na Figura 1 exibimos uma amostra de quadros da base.

Figura 1: Amostra dos vídeos não sensíveis (primeira linha) e pornográficos (segunda linha).

Metodologia Proposta

Utilizamos DLAs para detectar conteúdo perturbador em vídeos de desenhos animados. Como ponto de partida, avaliamos o *pipeline* proposto por Perez et al. [10] (trabalho desenvolvido pelo nosso grupo de pesquisa para detecção de pornografia), nos vídeos de desenhos animados com conteúdo pornográfico.

Perez et al. propuseram utlizar não só a informação estática do vídeo (quadros dos vídeos) mas também utilizar a informação dinâmica extraindo fluxos ópticos [12]. Para tal, eles avaliaram um método de fusão chamado *late fusion*, que combina a informação estática e a dinâmica combinando os *scores* emitidos pelos classificadores de cada informação. Em seu trabalho, Perez et al. utilizaram para todos os experimentos a arquitetura de rede neural profunda GoogLeNet [13].

Na Figura 2 ilustramos uma visão geral do *pipeline* da abordagem de Perez et al. que aplicamos em nosso trabalho. Primeiro, os quadros e informações de movimento são extraídos dos vídeos. Então, a DLA é utilizada para extrair características. Estas são agregadas pelo processo de *pooling* e então classificadas por um método de classificação convencional (e.g., *Support Vector Machines*).

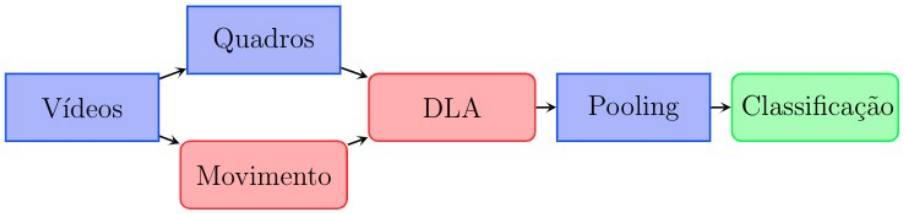

Figura 2: Visão geral da abordagem de extração de características espaço-temporais.

Na primeira etapa do trabalho, propusemos testar a hipótese a respeito da transferência de aprendizado: Uma rede treinada para vídeos pornográficos naturais é capaz de classificar desenhos animados? E se fossemos criar um modelo específico para desenhos animados, deveríamos partir (*transfer learning*) de um modelo treinada para pornografia ou de uma base genérica (i.e., ImageNet)? Para isso, avaliamos como a GoogLeNet, treinada por Perez et al. [10] para vídeos pornográficos com humanos, se comporta na tarefa de classificar nossa base de dados contra a GoogLeNet treinada para a base da ImageNet.

Apesar dos seus bons resultados, a GoogLeNet é uma rede neural convolucional (CNN) destinada a aplicações em plataformas *desktop,* devido ao seu grande número de parâmetros (aproximadamente 6 milhões). Como nosso objetivo principal é propor uma solução para plataformas móveis, na segunda etapa do projeto reproduzimos o *pipeline* de experimentos utilizando a MobileNetV2 [14], uma rede com cerca de 2.3 milhões de parâmetros desenvolvida para dispositivos móveis.

Resultados e Discussão

Transferência de Aprendizado

No primeiro experimento, aplicamos uma transferência de aprendizado com os modelos pré-treinados para pornografia e para a ImageNet para classificar nossa base de desenhos animados. A comparação de suas performances pode ser vista na Figura 3.

Deste experimento notamos duas observações importantes. O modelo que exibia 97% de acurácia para pornografia em vídeos naturais agora atinge 90% de acurácia para desenhos animados, uma queda significativa de desempenho. Confirmamos então nossa hipótese de que os modelos para vídeos naturais não são representativos o suficiente para cobrir os dois domínios (vídeos com humanos e animações).

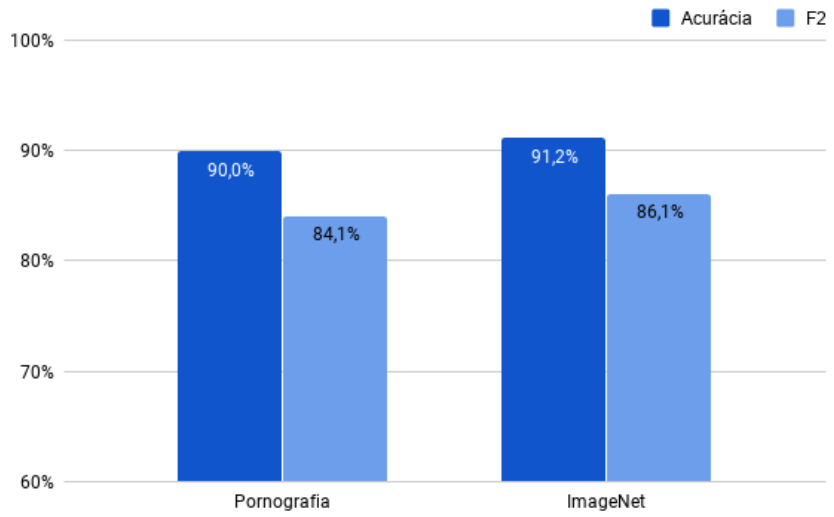

Figura 3: Comparação do desempenho das redes treinadas para pornografia e na base da ImageNet classificando a base de desenhos animados através de uma transferência de aprendizado.

O segundo ponto observado foi a performance sutilmente superior do modelo treinado na base da ImageNet em relação ao modelo treinado na base de pornogafia. Essa informação é especialmente importante pois demonstra que não precisamos retreinar os modelos de redes neurais profundas para um conjunto de vídeos pornográficos antes de utilizá-los em nossa tarefa de classificação de desenhos animados. É importante lembrar que os modelos de DLA em sua grande maioria já possuem modelos treinados para a ImageNet disponibilizados pela comunidade científica.

Finetuning

Dadas as evidências encontradas na primeira etapa do trabalho, exploramos outras arquiteturas de redes convolucionais para plataformas móveis cujos pesos pré-treinados para a ImageNet estavam disponibilizados pela comunidade científica. Escolhemos a MobileNetV2 [14] por sua alta performance dentre as redes propostas para plataformas móveis e por seu *design* baseado em *linear bottlenecks* que reduz o custo computacional de suas convoluções.

Neste experimento, como *baseline* realizamos apenas uma transferência de aprendizado em que utilizamos a MobileNetV2 treinada para a ImageNet para classificar nossa base de dados de desenhos pornográficos. Alcançamos uma acurácia de 88,4% e a medida F2 de 81,4%.

Aplicamos então um *finetuning*, em que inicializamos a rede com os pesos pré-treinados da ImageNet e retreinamos o modelo com a nossa base de desenhos pornográficos. Os resultados foram ainda melhores do que os obtidos pela GoogLeNet.

A comparação da *baseline* com o experimento com o *finetuning* pode ser vista na Figura 4. Vale ressaltar que nestes experimentos foi utilizado o mesmo *pipeline* descrito na seção de métodos.

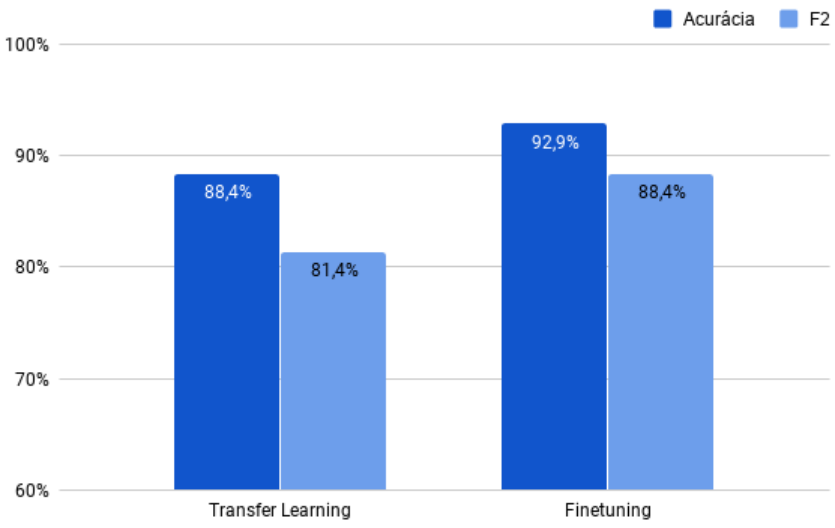

Figura 4: Comparação do experimento com apenas a transferência de aprendizado e o experimento com o *finetuning* na MobileNetV2.

Conclusões

Nossos experimentos sugerem que as soluções propostas para pornografia com vídeos naturais apresentam uma queda de performance significativa ao classificarem pornografia em desenhos animados. Adicionalmente, ao tratarmos os diferentes tipos de vídeos (naturais e animados) de forma especial, obtemos resultados melhores.

Dentro do âmbito de modelos de redes neurais profundas, nossos resultados demonstraram que boas performances podem ser obtidas apenas realizando transferência de aprendizado de modelos pré-treinados na ImageNet, sem a necessidade de se transferir de um modelo treinado outra base de conteúdo sensível (e.g., pornografia), economizando custo computacional e abrindo mais possibilidades de experimentações com outras arquiteturas cujos pesos treinados na ImageNet foram disponibilizados pela comunidade científica.

Infelizmente, pela falta de outros trabalhos na literatura envolvendo pornografia em desenhos animados, não conseguimos comparar nossos métodos e nossos resultados obtidos. Esperamos que este trabalho motive a produção de outros no tema.

Para trabalhos futuros, temos vários horizontes a serem investigados, como explorar outras arquiteturas de aprendizado profundo, implantar a solução em plataformas móveis e disponibilizá-la para pais e crianças. Explorar outras formas de conteúdo sensível em desenhos animados, como o recente fenômeno *Elsagate*, em que personagens infantis (e.g., Mickey, princesas da Disney) são caracterizados realizando atividades impróprias para crianças (e.g., fumando, bebendo, roubando, piadas envolvendo higiene pessoal e doenças) tem apresentado resultados promissores [11].

Referências Bibliográficas

[1] Livingstone, Sonia, John Carr, and Jasmina Byrne. "One in three: Internet governance and children's rights." (2015).

[2] Short, Mary B., et al. "A review of Internet pornography use research: Methodology and content from the past 10 years." *Cyberpsychology, Behavior, and Social Networking* 15.1 (2012): 13-23.

[3] Forsyth, David A., and Margaret M. Fleck. "Automatic detection of human nudes." *International Journal of Computer Vision* 32.1 (1999): 63-77.

[4] Lopes, Ana P. B., et al. "A bag-of-features approach based on hue-sift descriptor for nude detection." *17th European Signal Processing Conference*. IEEE, 2009.

[5] Avila, Sandra, et al. "BOSSA: Extended BoW formalism for image classification." *18th IEEE International Conference on Image Processing*. IEEE, 2011.

[6] Avila, Sandra, et al. "Pooling in image representation: The visual codeword point of view." *Computer Vision and Image Understanding* 117.5 (2013): 453-465.

[7] Caetano, Carlos, et al. "A mid-level video representation based on binary descriptors: A case study for pornography detection." *Neurocomputing* 213 (2016): 102-114.

[8] Moreira, Daniel, et al. "Pornography classification: The hidden clues in video space–time." *Forensic Science International* 268 (2016): 46-61.

[9] Moreira, Daniel, et al. "Multimodal data fusion for sensitive scene localization." *Information Fusion* 45 (2019): 307-323.

[10] Perez, Mauricio, et al. "Video pornography detection through deep learning techniques and motion information." *Neurocomputing* 230 (2017): 279-293.

[11] Ishikawa, Akari, Edson Bollis, and Sandra Avila. "Combating the Elsagate phenomenon: Deep learning architectures for disturbing cartoons." *7th IAPR/IEEE International Workshop on Biometrics and Forensics*, 2019 (aceito para publicação).

[12] Brox, Thomas, et al. "High accuracy optical flow estimation based on a theory for warping." *European Conference on Computer Vision*. Springer, 2004.

[13] Szegedy, Christian, et al. "Going deeper with convolutions." *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015.

[14] Sandler, Mark, et al. "MobileNetv2: Inverted residuals and linear bottlenecks." *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.