USO DE DEEP LEARNING PARA DETECÇÃO DE PICHAÇÕES COM GERAÇÃO AUTOMÁTICA DOS DADOS DE TREINAMENTO

DIEGO PORTO JACCOTTET[1]; CRISTIAN CECHINEL[2]; RICARDO MATSUMURA ARAUJO[3]

[1]*Universidade Federal de Pelotas – dpjaccottet@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – contato@cristiancechinel.pro.br*
[3]*Universidade Federal de Pelotas – ricardo@inf.ufpel.edu.br*

1. INTRODUÇÃO

Em cidades, pichações causam muitos prejuízos e elevados custos de limpeza. Com o crescimento da sociedade e das cidades, esse problema vem aumentando cada vez mais, e já pode custar milhões para uma cidade (ALONSO, 1998). Sendo assim existe necessidade por entender melhor esse fenômeno.

Atualmente existem alguns métodos para a captura de imagens de pichações na literatura que implementam diferentes técnicas para organizar essas pichações fotografadas. Alguns trabalhos vêm focando no desenvolvimento de aplicativos comunitários para tirar fotos de pichações, através do uso de *smartphones*, para solucionar o problema da coleta de imagens dessas pichações (PARRA et al., 2012).

As técnicas existentes para coleta de pichações normalmente são manuais (TONG et al., 2011). Porém seria possível realizar essa coleta automaticamente com os bancos de imagens disponíveis atualmente. Nesse sentido, este trabalho apresenta uma investigação sobre a detecção automática de pichações usando redes neurais convolucionais, as quais vêm sendo muito usadas no reconhecimento de imagens (KRIZHEVSKY et al., 2012). Como essas redes são muito profundas, muitas vezes elas necessitam de centenas de milhares de imagens para funcionarem (OORD et al., 2013). Sendo assim, pretende-se também explorar a geração automática de imagens a fim de tornar mais fácil o treinamento dessas redes. De fato, essa geração automática tem grande potencial para facilitar o treinamento de redes neurais, pois a quantidade de imagens necessárias é um fator debilitante para grande parte dos estudos na área. Consequentemente, o aumento da quantidade de imagens classificadas corretamente está diretamente relacionado à qualidade da rede neural no final do treinamento.

O objetivo geral é gerar um dataset e treinar uma rede neural capaz de identificar pichações. Os objetivos específicos são: criar um dataset de imagens reais a partir de banco de imagens e treinar uma rede nesse banco, e depois gerar um conjunto de treinamento sintético para comparação com as imagens reais.

2. METODOLOGIA

A investigação proposta primeiramente faz uso do Google Street View como banco de imagens (ANGUELOV, 2010). Esse banco possui imagens panorâmicas com alta resolução de muitas cidades do mundo. A coleta destas imagens é feita diretamente no navegador de internet, de duas maneiras.

A primeira maneira faz uso de um *script* para automatizar a coleta das imagens. Segundo OORD et al. (2013), 500.000 imagens seria uma quantidade

razoável para treinar uma rede convolucional profunda. Sendo assim, se objetivou coletar 755.000 imagens reais nesse trabalho. Dividindo essa quantidade em 500.000 para o treinamento da rede, 250.000 para a validação e 5.000 para o teste. Todos os conjuntos são divididos na metade como contendo ou não pichação. O conjunto de treinamento é usado para mudar modificar a rede diretamente, já o de validação é usado para saber se está ocorrendo *overfitting* e ajustar os parâmetros para treinamento, e o de teste é usado somente no final de todo o processo, a fim de criar uma estimativa do desempenho da rede no mundo real.

O primeiro script é útil no sentido de que coleta imagens variadas automaticamente, como fotos do mar, do céu, campo, etc. Porém ele coleta poucas imagens especificamente de pichações e outras que sejam parecidas com pichações, como paredes com muitos riscos, que possam confundir a rede.

Sendo assim, se fez necessário desenvolver um segundo método para coleta de imagens. Esse método usa um script para coletar imagens de 456x456 pixels do navegador ao apertar de um botão, a pichação deve estar centralizada na imagem. Depois outro script recorta essa imagem com uma janela deslizante para criar 100 imagens de 256 pixels. A Figura 1 mostra como independente de onde se recorte a imagem de 256x256 pixels, ou o quão pequena seja a pichação, ela sempre vai aparecer na imagem recortada.

Figura 1. Exemplo de um recorte em uma imagem capturada.

As imagens devem possuir 256x256 pixels, pois essa é a entrada usual das redes convolucionais estado da arte. A investigação proposta usa o *framework* de Deep Learning Faster-RCNN para realizar a detecção das pichações nas imagens (REN; GIRSHICK; SUN, 2015). Esse *framework* exige uma rede pré-treinada para classificação no Caffe. Uma das melhores redes do Faster-RCNN é a VGG16, por este motivo ela foi a escolhida neste trabalho.

Após realizar o treinamento da VGG16 no Caffe usando às 755.000 imagens de pichações, 377.500 contendo pichações e 377.500 sem pichações. O Faster-RCNN faz o *fine-tuning* da rede pré-treinada no Caffe, para isso ele precisa de pelo menos 10.000 imagens rotuladas. Neste trabalho se optou por gerar essas imagens automaticamente.

A geração automática das imagens rotuladas é feita com 100 imagens focadas em pichações, todas com fundo transparente, e 100 imagens sem pichações. Tenta-se sobrepor a pichação em uma posição randômica na imagem sem pichação, cerca de 20 vezes. A cada tentativa, é calculada a mudança que ocorreu na imagem sem pichação. O posicionamento com a mudança mais

significativa constitui uma imagem das 10.000 que serão usadas para o treinamento do Faster-RCNN, e o rótulo é gerado automaticamente para a posição da pichação na imagem.

3. RESULTADOS E DISCUSSÃO

Completada a coleta das 755.000 imagens, o primeiro experimento foi o treinamento da rede Caffenet, com 5 camadas convolucionais, para classificação. O resultado foi 91,76% de acerto nas 5.000 imagens de teste, com 90 falsos positivos e 322 falsos negativos. Em seguida, foi usada a mesma rede, porém fazendo o *fine-tuning* de modelo pré-treinado no Imagenet. Essa segunda tentativa alcançou 93,82% no mesmo conjunto de teste, com 33 falsos positivos e 276 falsos negativos. Depois, foi usada a rede VGG16 já pré-treinada no Imagenet, que alcançou 99,32% no mesmo conjunto de teste, com 5 falsos positivos e 29 falsos negativos.

É provavel que isso ocorra pela VGG16 possuir mais camadas, e também por sorte na escolha da taxa de aprendizado da rede. Essa taxa não pode ser alta demais, senão o treinamento diverge. Ela também está relacionada ao tamanho de *batch*, quanto menor for o *batch size* menor tem que ser o *learning rate*, já que o aprendizado é menos eficiente com *batch sizes* menores. A taxa também tem que ser diminuída no momento certo do treinamento, para que siga ocorrendo convergência. Treinamentos muito lentos ou muito rápidos tendem a não acabar nos melhores resultados (SZEGEDY et al., 2015).

Outro fator é que não foram usadas imagens de pichações muito difíceis de identificar, muito difíceis são pichações que mal aparecem na imagem, nesse primeiro experimento. Só foram usadas pichações óbvias nas imagens, aquelas que uma pessoa não conhecedora conseguiria identificar imediatamente. A Figura 2 mostra a entropia sobre o conjunto de treinamento para cada 500 iterações na rede VGG16, *learning rate* inicial de 0,002 e dividido por fator de 10 a cada 10.000 iterações.

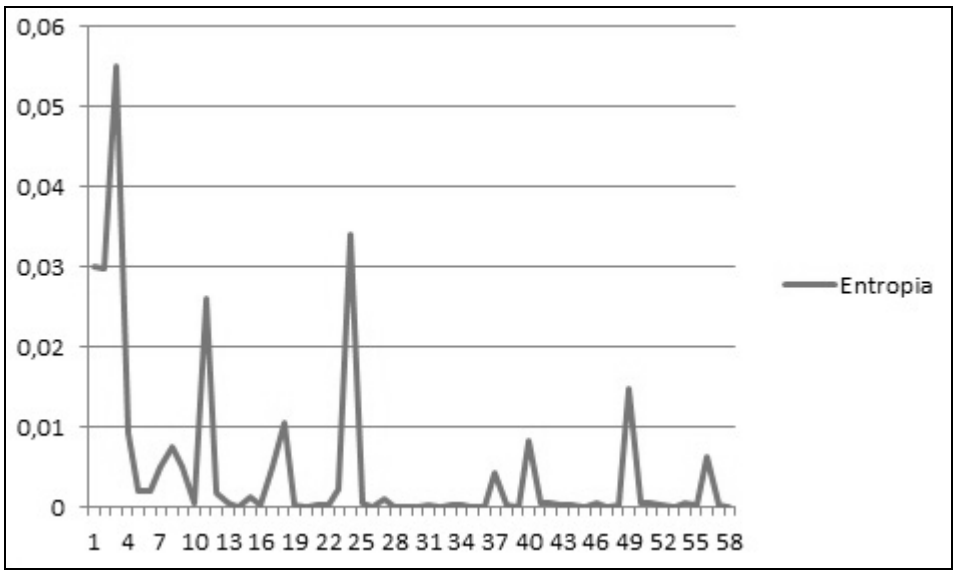

Figura 2. Entropia Sobre o Conjunto de Treinamento.

Os resultados são promissores para a detecção de pichações, mas até o momento só foi feito um primeiro teste para ver a possibilidade de classificação de imagens contendo pichações. Ainda pode-se melhorar o *dataset* e alcançar taxas de classificação melhores, incluindo para pichações difíceis. Além disso, falta

completar testes de detecção que determinem exatamente onde e quantas pichações existem em uma dada imagem, para isso será usado dataset sintético de 10.000 imagens, 5.000 para treinamento, 2.500 para teste e validação. Até esse momento só foram usadas imagens reais.

4. CONCLUSÕES

As principais inovações neste trabalho são: a criação de um dataset de pichações com 755 mil imagens, mais 10.000 imagens sintéticas, apropriado para treinamento de redes neurais convolucionais; o teste da hipótese de que é possível detectar pichações, mesmo quando estas se encontram em ambientes confusos com paredes muito antigas; o teste da hipótese de se geração de imagens artificiais é útil para o reconhecimento de pichações.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALONSO, A. Urban graffiti on the city landscape. **San Diego State University**, 1998.

ANGUELOV, D.; DULONG, C.; FILIP, D.; FRUEH, C.; LAFON, S.; LYON, R.; OGALE, A.; VINCENT, L.; WEAVER, J. Google street view: Capturing the world at street level. **Computer**, v.6, p.32-38, 2010.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G.E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, p. 1097-1105, 2012.

OORD, V.D; DIELEMAN, S.; SCHRAUWEN, B. Deep content-based music recommendation. **Advances in Neural Information Processing Systems**, Nevada, p. 2643- 2651, 2013.

PARRA, A.; BOUTIN, M.; DELP, E.J. Location-aware gang graffiti acquisition and browsing on a mobile device. **IS&T/SPIE Electronic Imaging**, p. 830402-830402. International Society for Optics and Photonics, 2012.

REN, S.; HE, K.; GIRSHICK; R. AND SUN, J. Faster R-CNN: Towards real-time object detection with region proposal networks. **Advances in Neural Information Processing Systems**, p. 91-99, 2015.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the inception architecture for computer vision. **arXiv preprint arXiv:1512.00567**, 2015.

TONG, W.; LEE, J. E.; JIN, R.; JAIN, A. K. Gang and moniker identification by graffiti matching. **Proceedings of the 3rd international ACM workshop on Multimedia in forensics and intelligence**, p. 1-6, 2011.