Projeto Final - UNBeauty: Banco de imagens para reconhecimento de imagens de diferentes domínios

Frederico Guth (18/0081641)
Tópicos em Sistemas de Computação, ,
Turma TC - Visão Computacional (PPGI)
Universidade de Brasília
Brasília, Brasil
fredguth@fredguth.com

***Resumo*—O enorme avanço na pesquisa em reconhecimento de objetos deve bastante ao desenvolvimento de bancos de imagens de larga escala bem anotados. Entretanto, para ilustrar o mundo "cauda-longa"é impossível obter diversas imagens de treinamento por categoria e será preciso avançar na transferência de conhecimento entre domínios. No presente projeto, criamos uma base de dados de imagens de cosméticos em diferentes domínios. Espera-se que essa base possa ser útil no desenvolvimento de novos algoritmos de reconhecimento que funcionem bem "cross-domain".**

***Index Terms*—base de imagens, reconhecimento de objetos, transferência de domínios**

I. Introdução

Nos últimos anos, houve um enorme avanço nos algoritmos de reconhecimento de objetos [1]. Este resultado só pode ser obtido devido a existência de bancos de imagens grandes e bem anotados, com muitas imagens por classe [1], [2]. Em 8 anos do desafio da Imagenet (ILSVRC), o erro diminuiu uma ordem de magnitude [2] e, em 2017, chegou a apenas 2,3%.

A. Reconhecimento de Objetos

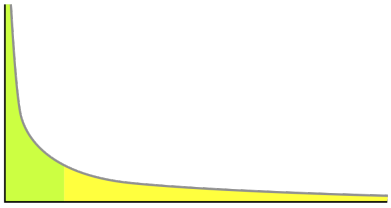

Figura 1: Uma distribuição de cauda longa. Note que as áreas das duas regiões são iguais, onde a região para a direita representa a cauda longa.

Entretanto, uma quantidade importante de comportamentos da natureza e atividades humanas obedecem a uma distribuição estatística assintótica, conhecida como cauda longa (vide Figura 1. Para dar um exemplo, a maioria dos cães que vemos são de algumas poucas raças (categorias) e a maioria das raças tem poucos cães, isto é são raras. Na indústria existem cada vez mais produtos de nicho. Na Amazon, por exemplo, já em 2008 os produtos de nicho representavam 36.7% das vendas [3] e com certeza esse número aumentou nos últimos 10 anos.

Na indústria de cosméticos uma empresa chega a ter milhares de produtos diferentes (*skus*). Além disso, imagens desses produtos aparecem em diferentes contextos, "domínios": fotos de marketing, que são feitas por fotógrafos profissionais e fortemente modificadas digitalmente para diferenciar e "embelezar"os produtos;e fotos reais que são produzidas por fotógrafos amadores com diferentes equipamentos, condições de iluminação e pose; ou seja, cada domínio tem um viés em termos de variedade visual, como a resolução, *viewpoint* e iluminação.

Essa característica do mundo que nos cerca representa alguns desafios para a aplicação prática dos algoritmos de reconhecimento e classificação: (a) o número de categorias pode ser muito grande, (b) o número de amostras para treinamento pode ser muito pequena [1] e (c) os domínios das imagens de treinamento e teste podem ser diferentes.

Neste contexto, bancos de imagens que permitam os algoritmos de reconhecimento a lidar com esses problemas de transferência de domínio e cauda longa são extremamente bem-vindos.

B. Objetivo

Este projeto tem dois objetivos: (1) criar um banco de imagens de produtos de cosméticos com amostras de diferentes domínios e (2) ilustrar, com um algoritmo de reconhecimento de objetos, o impacto da mudança de domínio.

II. Revisão Teórica

Reconhecimento de Objetos lida com a identificação de objetos em imagens. Esta tarefa tão natural aos seres humanos é um grande desafio para algoritmos. Afinal, a Ciência da Computação se desenvolveu até recentemente com grande ênfase no método dedutivo e o arcabouço mental para desenvolvimento de algoritmos era a ideia de instruir o computador, passo a passo, como realizar uma tarefa. Mas como ensinar um computador a enxergar? Como instruir, passo a passo, algo que não sabemos como fazemos?

O momento crucial na melhoria meteórica dos resultados em reconhecimento de objetos se deu em 2012, no desafio *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) [4]. O time liderado por Alex Krizhevsky foi o primeiro a usar redes neurais convolucionais profundas (RCPs) na competição

e ganhou por larga margem [5]. Desde então, técnicas baseadas em RCPs tem sido as mais bem sucedidas para este problema. É importante salientar, entretanto, que tal feito não seria possível se não houvesse uma banco de imagens de larga escala, manualmente anotada e com centenas de imagens por classe, como a ImageNet. *Deep learning* é epistemologicamente muito mais próximo das ciências naturais, uma vez que é um método indutivo, em que o resultado empírico é crucial: o ajuste de hiperparâmetros pode ter tanto ou mais valor do que o desenvolvimento de novas arquiteturas e o desenvolvimento de bancos de imagens é uma fase crucial para aplicações práticas. Andrej Karpathy [6] afirma que a maior diferença entre o trabalho como aluno de doutorado em Stanford e como diretor de IA da Tesla é o foco na criação e manutenção de bancos de imagens (Figura 2). Cada vez mais são os bancos de imagens, e não algoritmos, o fator limitante da aplicação prática e desenvolvimento da inteligência artificial.

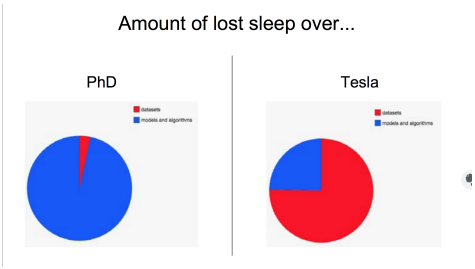

Figura 2: Diferença entre academia e indústria segundo Andrej Karpathy [6]. (Datasets em vermelho; Modelos e algoritmos em azul.)

A. Bancos de Imagens para Reconhecimento de Objetos

Bancos de imagens são vitais para a pesquisa em reconhecimento de objetos. Pode-se dizer que foram um componente chave para o metéorico progresso obtido nos últimos anos, não apenas como fonte de dados para treinamento, mas também como meio de comparação de resultados de pesquisa [7].

Caltech 101 (2004) [8] foi uma dos primeiros bancos de imagens padronizados para reconhecimento de objetos, com 101 classes de objetos e entre 15 e 30 imagens de treinamento por categoria. O banco de imagens Pascal Visual Object Classes (2006) [**?**], Pascal VOC, com 20 categorias e imagens obtidas na internet, popularizou a ideia de associar um banco de imagens padronizado com *ground-truth* e uma competição. Fortemente influenciado pelo Pascal VOC, ImageNet (2009) [9] é um banco de imagens organizado de acordo com a hierarquia da WordNet [**?**]. Apesar das mais de 14 milhões de imagens manualmente anotadas e 21 mil categorias, o subconjunto de sua base destinado à competição ILSVRC, com menos de 2 milhões de imagens e apenas 1000 categorias, é mais utilizado.

B. Viés em Bancos de Imagens e Generalização de Domínios

Faz pouco sentido esperar ausência de viés em qualquer representação do mundo visual. A maioria dos bancos de imagens se propõe a representar o ambiente visual encontrado no dia a dia das pessoas. Mas é impossível avaliar o quão bem um banco de imagens atende esse objetivo, ou qual é o seu viés, sem ter um *ground truth* (que por sua vez seria uma base de dados que também poderia ter viés) [7].

Pode-se dizer, inclusive, que o desenvolvimento de bancos de imagens se dá através da reação a vieses passados: O banco de imagens COIL-100 (100 objetos comuns em um background preto) foi uma reação ao uso de bases modelo e passou a valorizar texturas. A coleção Corel Stock Photos foi uma reação a simplicidade visual de bases como COIL-100. Caltech-101 foi uma reação ao profissionalismo das fotos da Corel e valoriza a diversidade das imagens da internet. Pascal VOC foi uma reação a inexistencia de padrões de treinamento e teste. Imagenet é uma reação a inadequação de bancos de imagens anteriores que eram pequenos demais para a complexidade visual do mundo [7].

Diante deste dilema, uma maneira de se avaliar o viés de bancos de imagens é checar a generalização entre bancos de imagens, por exemplo, treinar com imagens Pascal VOC e testar com imagens da ImageNet [7].

Entretanto, a questão da generalização de domínio é tratada como um caso especial do reconhecimento de objetos chamado adaptação de domínio e menosprezada na grande maioria dos desenvolvimentos de bases de imagens. Praticamente não se encontra comparações de resultados em-domínio (*in-domain*) e entre-domínios (*cross domain*) para os algoritmos baseados em redes neurais convolucionais profundas.

III. Metodologia

A. Materiais

Foram utilizados:

- Servidor Paperspace/Fastai: GPU 8GB, 30GB RAM, 8 CPU
- NVIDIA Quadro P4000 com 1792 CUDA cores.
- Python 3.6.4 :: Anaconda custom (64-bit)
- Pytorch 0.3.0
- OpenCV 3.4.0
- Programa RectLabel
- 5 Notebooks Jupyter
- 12 programas python.
- celular Samsung S8
- um pano vermelho de 2m x 1,2m
- 168 produtos da empresa O Boticário

Todos os arquivos do projeto estão publicamente disponíveis em git@github.com:fredguth/unb-cv-3183.git

B. Construção da Base de Imagens

A base de dados UNBeauty foi construída com dois domínios. Importante mencionar que no domínio das fotos profissionais já dispunhamos de imagens de várias revistas dO Boticário.

1) Domínio vídeo de celular
    a) Obtivemos 168 diferentes produtos (*skus*) da marca O Boticário;

b) Montamos uma mesa apoiada em uma parede e pregamos um pano vermelho da parede até cobrir totalmente a mesa;
c) Com o celular Samsung S8, fizemos um vídeo de poucos segundos (6 a 15s por produto) da parte frontal de cada produto nesse cenário. Quando o produto era transparente, também fizemos um vídeo do mesmo contra um fundo branco;
d) Para cada vídeo, anotamos o código do produto e alteramos o nome do vídeo para o nome do código do produto. Quando mais de um vídeo foi feito para o mesmo produto, simplesmente acrescentamos uma letra (b, c, etc);
e) Convertemos os vídeos em imagens organizadas por categoria (movie2images.py);
f) Fizemos uma calibração de cores simples em todas as imagens (balanceDataset.py).

2) Domínio das Imagens de Revista
a) Baixamos as imagens das revistas de um repositório S3 (importMags.py);
b) Com o programa RectLabel anotamos com bounding boxes os produtos para os quais tínhamos imagens no domínio do vídeo de celular;
c) A partir do json gerado pelo RectLabel, usamos o programa boti-4.py para gerar um dataset chamado mags_test organizado por categoria.

Um aspecto importante a salientar é que escolhemos domínios com características bem diferentes. O domínio do vídeo do celular permite várias imagens por classe, mas todas muito similares. O domínio das imagens de revista apresenta poucas amostras por classe e imagens que são bastante modificadas em termos de cores e composição. Esperamos que essa escolha possa ilustrar a diferença da classificação *in-domain* da *cross-domain*.

C. Classificador de objetos

O método de reconhecimento de objetos desenvolvido nesse projeto foi fortemente influenciado por [10] e constitui-se das seguintes etapas:

1) Definimos Resnet-50 como arquitetura RCP e obter rede pré-treinada com imagens da base ImageNet;
2) Usando o programa split_dataset.py, dividimos o dataset em bases estratificadas de treinamento, validação e teste, com 9084, 2272 e 2839 imagens, respectivamente;
3) Relaxamos as últimas 2 camadas restantes da Resnet-50;
4) Otimizamos a Taxa de Aprendizado (Learning Rate);
5) Otimizamos o resultado em Tempo de Teste (Test Time Augmentation)
6) Rodamos 10 vezes o teste para obter a média e o desvio padrão da acurácia;
7) Repetimos os passos acima usando a base mags_test com imagens do domínio das fotos profisionais
8) Repetimos os passos acima aplicando *Data Augmentation*

(a) Linha Coffee

(b) Linha Floratta

(c) Linha Match

Figura 3: Produtos de uma mesma linha podem ser bastante similares.

IV. Resultados

Nesta seção apresentamos os resultados obtidos. Todos os dados podem ser acessados no repositório do projeto (III-A).

A. Base de Imagens UNBBeauty

Com o método apresentado, obtivemos uma base de dados com:

- 14.195 imagens de produtos no domínio do vídeo de celular; divididas em
- 168 categorias, nomeadas pelo sku que representavam
- uma seleção estratificada de treinamento, validação e teste com 9.084, 2.272 e 2.839 imagens, respectivamente;
- 53 imagens de teste no domínio das fotos profissionais em 45 categorias

A base é de alta granularidade, com várias categorias muito parecidas entre si, como mostrado na Figura 3. A diferença entre os domínios pode ser observada na Figura 4.

B. Classificação de Objetos entre domínios

C. Análise dos resultados

Apesar de servir ao propósito, há alguns problemas na base criada:

- o pano de fundo nas imagens do domínio do celular foi uma péssima escolha. O vermelho reflete nos produtos e altera a cor dos produtos, principalmente aqueles mais transparentes.

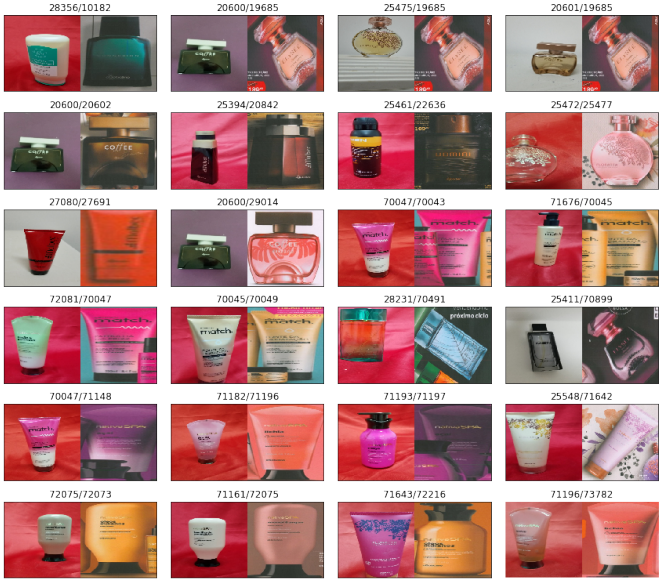

Figura 4: Erros do classificador

Tabela I: Média e desvio padrão da acurácia

	Augmentation	$\overline{\phi}\%$	$\overline{\sigma}\%$
In-domain	Não	99.89	0.05
	Sim	99.94	0.03
Cross-domain	Não	-	-
	Sim	55.32	1.41

- as imagens das revistas se repetem entre revistas, de forma que foi difícil encontrar amostras diferentes de cada produto.
- nem todos os produtos do domínio do celular foram encontrados nas revistas. Em retrospecto, teria sido mais eficiente ter primeiro obtido o maior número de amostras diferentes no domínio das revistas e só fazer vídeos dos produtos encontrados.

V. Discussão e Conclusões

Neste trabalho, implementamos dois algoritmos para rastreamento visual de objetos em vídeo. O algoritmo KCF apresenta bons resultados de acurácia e robustez. Entretanto, mostramos que seu modelo pode melhorar se incorporar incerteza, o que pode ser feito com um filtro de Kalman. Ao *amortecer* o resultado do KCF com um filtro de Kalman, tivemos resultados de robustez entre 60 e 90% melhores, com perdas de acurácia menores que 12%, esse resultado serve de inspiração para melhorias na implementação do rastreador KCF da OpenCV.

Referências

[1] G. V. Horn and P. Perona, "The devil is in the tails: Fine-grained classification in the wild," *CoRR*, vol. abs/1709.01450, 2017.

[2] J. D. L. Fei-Fei, "Where have we been? where are we going?" http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf, 2017, [Online; accessada 28 de Junho de 2018].

[3] E. Brynjolfsson, Y. J. Hu, and M. D. Smith, "The longer tail: The changing shape of amazon's sales distribution curve," *SSRN Electronic Journal*, 2010. [Online]. Available: https://doi.org/10.2139/ssrn.1679991

[4] I. J. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning*, ser. Adaptive computation and machine learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org/

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[6] A. Karparthy, "Building the software 2.0 stack," https://www.figure-eight.com/wp-content/uploads/2018/06/TRAIN_AI_2018_Andrej_Karpathy_Tesla.pdf, 2018, [Online; accessada 28 de Junho de 2018].

[7] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, jun 2011. [Online]. Available: https://doi.org/10.1109/cvpr.2011.5995347

[8] F.-F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories." *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[9] M. Guillaumin and V. Ferrari, "Large-scale knowledge transfer for object localization in ImageNet," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012. [Online]. Available: https://doi.org/10.1109/cvpr.2012.6248055

[10] J. Howard *et al.*, "fastai," https://github.com/fastai/fastai, 2018.