Proposta de trabalho
Trabalho Supervisionado de Formatura

Segmentação de imagens com redes totalmente convolucionais

Pedro Henrique Barbosa de Almeida
Estudante

Nina S. T. Hirata
Orientadora

Fomentado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) sob o processo nº 2020/02891-3.

Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

São Paulo, 22 de março de 2020

1 Introdução

Os processamentos de imagens, em geral, baseiam-se em combinações de vários tipos de transformações, tarefa realizada pelos operadores de imagens. Vários desses operadores são transformações locais, caracterizadas por uma função local. Por função local, referimo-nos a uma função cuja entrada é em geral uma pequena região da imagem centrada num pixel. Essa função é aplicada pixel a pixel para gerar a imagem transformada. Desta forma, torna-se possível modelar o problema de projetar um operador como um problema de aprendizado dessas funções locais.

As abordagens mais recentes para transformação imagem-para-imagem utilizam modelos de redes totalmente convolucionais (*Fully Convolutional Networks* ou simplesmente FCN, em inglês) [1], que são capazes de processar uma imagem inteira de uma só vez.

O objetivo deste Trabalho Supervisionado de Formatura é estudar e aplicar as FCN em tarefas que usualmente são realizadas via classificação de pixels, como segmentação de vasos da retina, segmentação de textos em imagens de documentos ou a remoção de linhas em partituras de música.

Em particular, estamos interessados em adaptar uma conhecida técnica de combinação de transformações locais, que requer múltiplos passos de treinamento, para o contexto de aprendizado profundo. Desta forma, espera-se que essa combinação de operadores possa ser treinada na forma ponta-a-ponta em apenas um passo, e ainda que a rede gerada seja do tipo FCN, o que permitirá o processamento de todos os pixels de uma só vez. Além disso, outro objetivo é avaliar a performance das redes resultantes, comparando-as com as contrapartes já existentes.

Este projeto é financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) sob o processo nº 2020/02891-3.

2 Objetivos

O objetivo geral desse projeto é estudar e aplicar redes totalmente convolucionais em tarefas de segmentação conhecidas por serem feitas a nível de pixels, tais quais: segmentação de vasos da retina [2], segmentação de texto em imagens de documentos ou remoção de linhas em partituras de música.

A quantidade de dados de treinamento é crucial para obter classificadores com boa performance. Assim, uma situação desafiadora surge quando os dados para o treinamento são limitados. Em particular, estamos interessados em adaptar a abordagem descrita em [3] para o contexto de aprendizado profundo.

De acordo com [3], num cenário com uma quantidade finita de dados de treinamento, a abordagem multinível é um jeito efetivo de melhorar os resultados ao combinar múltiplos classificadores. Especificamente, desejamos desenvolver um procedimento de treinamento fim-a-fim para essa abordagem multinível e avaliar a performance das redes neurais resultantes, comparando com a performance das redes equivalentes que são de nível único.

3 Metodologia

Muitas transformações imagem-a-imagem podem ser caracterizadas em termos de uma função local que recebe um *patch* centrado em um pixel e atribui um valor de saída para esse pixel central. O *patch* é geralmente delimitado por uma janela retangular W. A transformação da imagem toda consiste no deslocamento da janela W pela extensão da imagem e, ainda, por aplicar essa função pixel a pixel. Essa caracterização torna possível inserir o problema de desenhar operadores de imagem no contexto de aprendizado de máquina. Especificamente, o aprendizado dessas funções locais pode ser visto como um problema de treinamento de classificadores [4].

Essa abordagem é largamente explorada na literatura [4][5]. A definição do tamanho do *patch* (ou janela) é um aspecto crítico. Teoricamente, quanto maior o tamanho da janela, melhor. Na prática, deve ser levado em consideração a quantidade de dados de treinamento disponível. Janelas muito pequenas não tem poder discriminativo (alto viés), enquanto que janelas muito grandes não são precisas (alta variância). Dessa forma, a habilidade de discriminação de uma função local será limitada pelo tamanho da janela ótima empiricamente observada. Quaisquer estruturas presentes na imagem maiores que a janela, podem não ser processadas corretamente.

Combinar classificadores especializados em campos receptivos ligeiramente deslocados em respeito ao pixel-alvo é uma forma de aumentar o tamanho efetivo da janela, sem lidar explicitamente com *patches* grandes, como proposto em [3].

A abordagem de treinamento tratada em [3] consiste em primeiro treinar um número de classificadores $\Psi_1, \Psi_2, ..., \Psi_k$, cada um baseado em uma janela W_j distinta, que mapeia patches de entrada x_i em alvos correspondentes y_i. As janelas W_j são escolhidas, cada uma, para coletar *patches* ligeiramente deslocados em respeito ao pixel-alvo. Assim, para cada pixel-alvo p_i, o alvo y_i é comum a todos os classificadores, mas suas entradas são partes ligeiramente distindas da imagem de entrada ao redor do pixel p_i. Assim, num segundo passo, uma nova rodada de treinamento é realizada para aprender como combinar as saídas dos classificadores $\Psi_1, \Psi_2, ..., \Psi_k$. Mais especificamente, supondo que S é uma imagem de entrada e $\Psi_j(S)$ é a imagem S processada por W_j, os dados de treinamento para o segundo passo são da forma $x_i = ([\Psi_1(S)](p_i), [\Psi_2(S)](p_i), ..., [\Psi_k(S)](p_i))$ e y_i.

Para implementar esse método em uma abordagem fim-a-fim (imagem como entrada e imagem como saída), nós adaptaremos uma rede de aprendizado profundo totalmente convolucional [1, 2, 6]. A vantagem de uma rede neural profunda é a possibilidade de realizar a otimização conjunta do pipeline completo. Nesse caso, será desenvolvida a arquitetura de rede que combina os k classificadores. Dessa forma, será possível realizar o treinamento em duas etapas descrito em [3] de uma forma fim-a-fim usando apenas uma rede "única". Além disso, como essa será uma arquitetura totalmente convolucional, isso significa que o modelo será capaz de processar imagens de qualquer tamanho em apenas um único passo *forward*.

Pretendemos usar os frameworks do Keras combinados com Tensorflow ou

mesmo o PyTorch para a implementação. Os possíveis conjunto de dados para a parte experimental desse projeto são o DRIVE [1] e o MUSCIMA++ [2].

4 Plano de trabalho

Para alcançar os objetivos descritos, o projeto será desenvolvido tendo como pontos norteadores:

1. Elaboração da proposta.
2. Criação de site para divulgação de resultados e códigos.
3. Estudar o paper de referência "Multilevel Training of Binary Morphological Operators".
4. Estudar o paper "U-Net: Convolutional Networks for Biomedical Image Segmentation".
5. Estudar PyTorch/Tensorflow ou outros frameworks semelhantes conforme as demandas de implementação dos modelos desenvolvidos.
6. Implementação, treino e avaliação de um dos modelos totalmente convolucionais (possivelmente U-Net).
7. Desenvolvimento da arquitetura de combinação de classificadores.
8. Implementação, treino e teste da arquitetura de combinação de classificadores.
9. Aplicação e avaliação da arquitetura de combinação de classificadores em diversos datasets.
10. Escrita de relatório ou artigo.
11. Escrita da monografia.
12. Elaboração do pôster ou de apresentação.

[1] `https://www.isi.uu.nl/Research/Databases/DRIVE/`
[2] `https://ufal.mff.cuni.cz/muscima`

CRONOGRAMA									
Atividades	**Mês**								
	3	4	5	6	7	8	9	10	11
1.	x								
2.		x							
3.		x	x						
4.		x	x						
5.			x	x	x	x			
6.				x					
7.					x	x			
8.							x		
9.							x		
10.								x	x
11.								x	x
12.								x	x

Com a execução dos passos supracitados, os resultados esperados são:

- Oportunidade de ampliar o conhecimento e treino do estudante na área de visão computacionais e aprendizado de máquina, especialmente, na subárea de redes de aprendizado profundo.
- A habilidade de programar arquiteturas customizadas. Isso se dá, pois muitas pessoas na área usam apenas arquiteturas já implementadas ou modelos simples. Com esse projeto, espera-se disseminar a prática de programar modelos personalizados de mais "baixo nível".
- Publicação de resultados. É esperado também que uma execução bem sucedida desse projeto resulte na submissão de pelo menos um paper, além da monografia e apresentação (ou pôster), obrigatória a todos que realizam um trabalho supervisionado de formatura.

Referências

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[3] Nina S. T. Hirata. Multilevel training of binary morphological operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):707–720, April 2009.

[4] Igor S. Montagner, Nina S. T. Hirata, and R. Hirata Jr. Image operator learning and applications. In *Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, pages 38–50, 2016.

[5] Fabian Tschopp, Julien NP Martel, Srinivas C Turaga, Matthew Cook, and Jan Funke. Efficient convolutional neural networks for pixelwise classification on heterogeneous hardware systems. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1225–1228. IEEE, 2016.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.