Classificação de Padrões de Localização de Proteínas Subcelulares Usando Métodos de Segmentação

Exame de Qualificação de Mestrado

Candidata: Juliana Midlej do Espírito Santo
Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Hélio Pedrini

Instituto de Computação
Universidade Estadual de Campinas

18 de Outubro de 2024

Agenda

1. Introdução
2. Fundamentação Teórica
3. Trabalhos Correlatos
4. Materiais e Métodos
5. Plano de Trabalho

Agenda

1. Introdução
2. Fundamentação Teórica
3. Trabalhos Correlatos
4. Materiais e Métodos
5. Plano de Trabalho

Introdução

- As proteínas são as unidades fundamentais das células e executam a maior parte das funções celulares [1];
- A função dos sistemas celulares é predominantemente definida pela estrutura, quantidade, localização espacial e interações das proteínas que compõem o proteoma [2];
- O aumento da quantidade de dados de imagens fluorescentes torna crucial desenvolver modelos computacionais para classificar a distribuição espacial de proteínas em células individuais [2].

Descrição do Problema

- Classificar rótulos de localização de organelas celulares para cada célula em uma imagem.

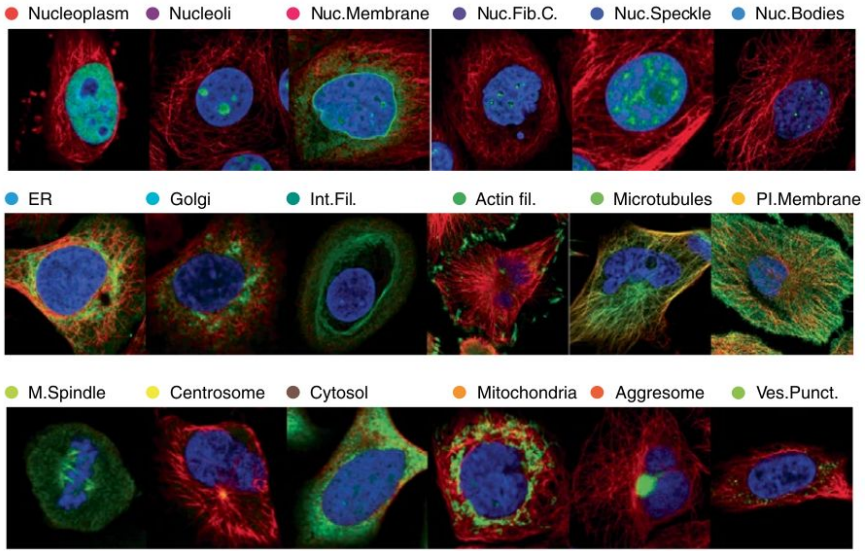

Figura: Rótulos de localização de organelas celulares [2].

Descrição do Problema

Desafios

- Desequilíbrio extremo de classes;
- Classificação multirrótulo;
- Rótulos a nível de imagem para treinamento dos modelos;
- Necessidade de rótulos de células individuais.

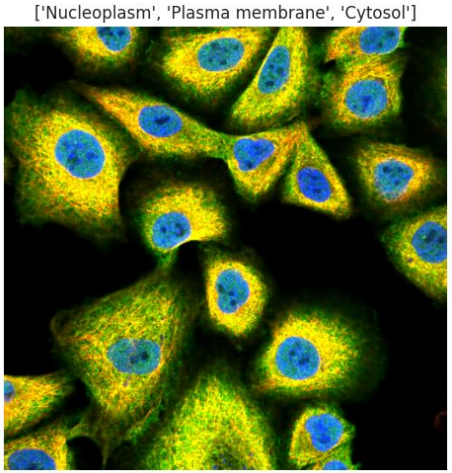

Figura: Classificação multirrótulo de uma imagem.

Objetivos

- Objetivo principal: classificar padrões de localização de proteínas subcelulares em células individuais a partir de imagens de microscopia fluorescente.
- Objetivos específicos:
    1. Revisão bibliográfica e estudo das principais abordagens utilizadas na classificação de proteínas subcelulares;
    2. Aplicar métodos de segmentação com aprendizado fracamente supervisionado para segmentar as células e classificar os padrões de localização de proteínas subcelulares;
    3. Avaliar e comparar o desempenho da metodologia proposta com outras abordagens disponíveis na literatura;
    4. Documentar e publicar os resultados.

Hipóteses do Trabalho

- É possível classificar padrões de localização de proteínas subcelulares em células individuais usando métodos de segmentação fracamente supervisionados;
- Modelos que empregam segmentação com aprendizado fracamente supervisionado podem obter maior precisão em comparação com abordagens tradicionais e recentes;
- A aplicação de técnicas de aumentação de dados resultará em modelos mais robustos generalizáveis.

Agenda

1. Introdução
2. Fundamentação Teórica
3. Trabalhos Correlatos
4. Materiais e Métodos
5. Plano de Trabalho

Rede Neural Artificial

- Redes Neurais Artificiais (*Artificial Neural Networks*, ou ANNs) são modelos computacionais inspirados na estrutura e funcionamento do cérebro humano [3];
- Perceptron [3]:

$$y = \begin{cases} 0, & \text{se } \sum_j w_j x_j \leq T \\ 1, & \text{se } \sum_j w_j x_j > T \end{cases}$$

- Função sigmoide [4]:

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$z = \sum_{j=1}^{n} w_j x_j + b$$

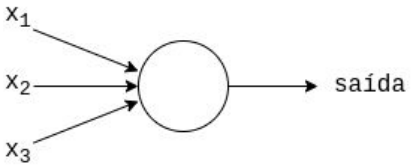

Figura: Exemplo de um perceptron.

Rede Neural Artificial

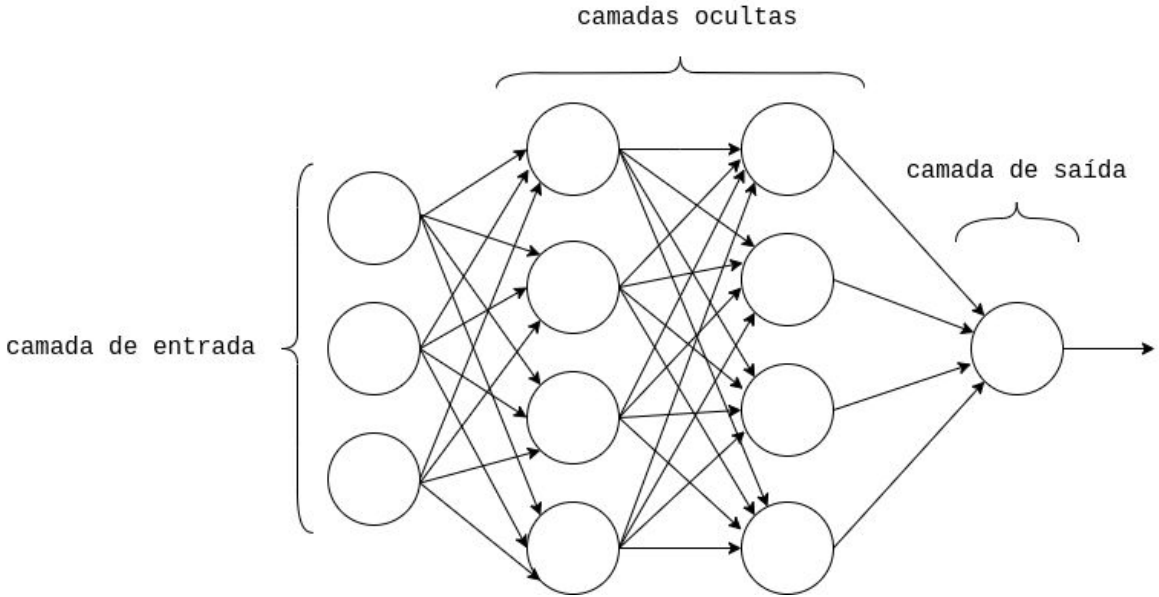

Figura: Exemplo de uma Rede Neural Artificial.

Rede Neural Profunda

- Uma Rede Neural Profunda (*Deep Neural Network*, ou DNN) é uma rede com muitas camadas;
- O aprendizado profundo permite que modelos computacionais aprendam representações de dados com múltiplos níveis de abstração [5];
- Tarefas com imagens obtiveram grandes avanços com o uso de aprendizado profundo [5].

Rede Neural Convolucional Profunda

- *Deep Convolutional Neural Networks*, ou DCNNs;
- Camadas convolucionais aplicam filtros para extração de características locais em diferentes partes da imagem [4];
- Camadas de agrupamento (*pooling*) reduzem o tamanho da entrada e o número de parâmetros [4].

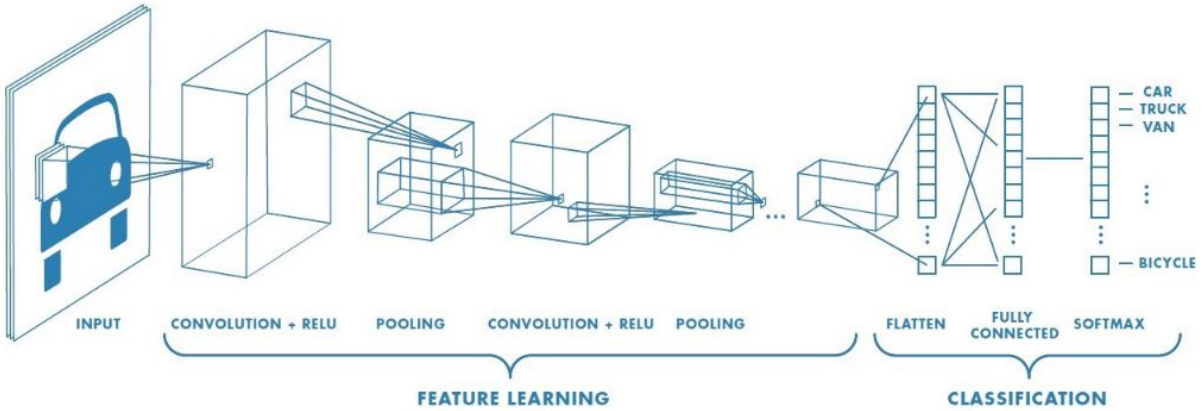

Figura: Exemplo de uma Rede Neural Convolucional Profunda [6].

Aprendizado Fracamente Supervisionado

- O alto custo de rotulagem torna desejável que técnicas de aprendizado de máquina lidem com supervisão fraca. [7].
- Tipos de supervisão fraca:
  - Supervisão incompleta: apenas um subconjunto dos dados de treinamento é rotulado;
  - Supervisão inexata: apenas rótulos de baixa precisão;
  - Supervisão imprecisa: os rótulos fornecidos nem sempre são a verdade absoluta.

Aumentação de Dados

- Aumentar artificialmente o tamanho e a diversidade do conjunto de dados de treinamento [8];
- Usa técnicas de distorção de dados ou superamostragem [9].

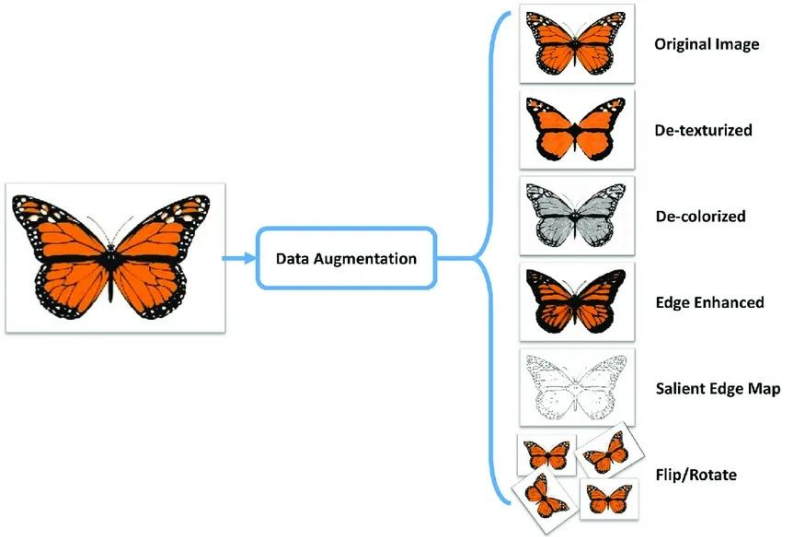

Figura: Aumentação de dados em uma imagem [10].

Segmentação de Imagens Fracamente Supervisionada

- Identificação e a separação de cada instância de um objeto em uma imagem, atribuindo um rótulo a cada pixel pertencente a uma instância específica.
- Podemos destacar as seguintes técnicas:
  - Puzzle-CAM [11];
  - P-NOC [12];
  - C²AM-H [12].

Agenda

1. Introdução
2. Fundamentação Teórica
3. Trabalhos Correlatos
4. Materiais e Métodos
5. Plano de Trabalho

Modelos com Padrões Únicos

Boland, Markey e Murphy (1998) [13]

- Imagens representando os padrões de localização de 4 tipos de proteínas e de DNA;
- Descritores numéricos como momentos de Zernike e características de textura de Haralick;
- Classificadores baseados em árvores de decisão e redes neurais.

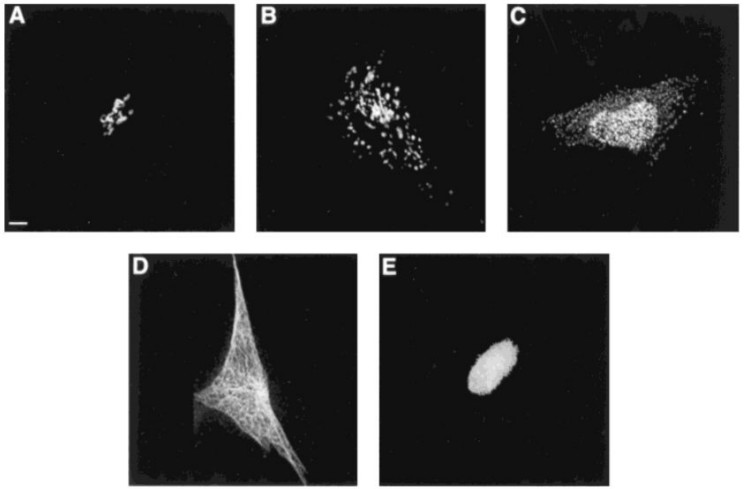

Figura: Imagens mostram células marcadas com anticorpos contra giantina (A), LAMP2 (B), NOP4 (C), tubulina (D) e com a coloração de DNA Hoechst 33258 (E) [13].

Modelos com Padrões Únicos

Kraus et al. (2016) [14]

- DCNNs com aprendizado de múltiplas instâncias (*Multiple Instance Learning*, ou MIL) para segmentar e classificar imagens de microscopia celular;
- **Imagens de microscopia fracamente rotuladas**;
- Bancos de imagens: MNIST; Células de Câncer de Mama; Proteínas de Levedura.

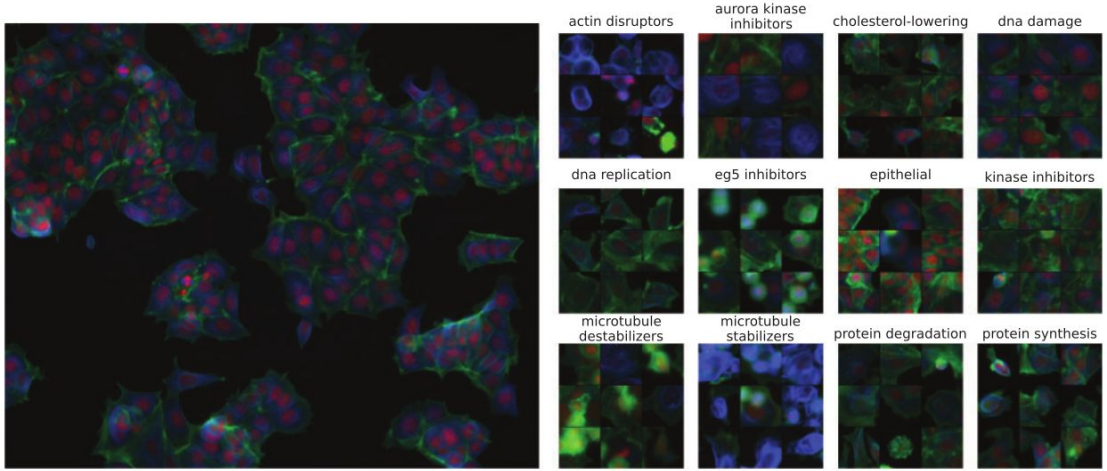

Figura: Células de Câncer de Mama. À esquerda, imagem de amostra em resolução total. À direita, amostras de células segmentadas provenientes de 12 classes [14].

Modelos com Padrões Mistos

Autores	Ano	Contribuições	Resultados
Sullivan et al. [15]	2018	Gamificação; Desenvolvimento da ferramenta automatizada de anotação celular Loc-CAT; Combinação das anotações dos jogadores com modelos de aprendizado profundo.	Taxa F_1 de 0,72. Macro F_1 de 0,47.
Xiang et al. [16]	2019	Desenvolvimento da rede neural convolucional assimétrica e em várias escalas (AMC-Net).	Taxa F_1 de 0,82.
Ouyang et al. [17]	2019	Análise dos resultados de uma competição no Kaggle para classificação de padrões de localização de proteínas.	Macro F_1 de 0,59.
Le et al. [2]	2022	Análise dos resultados de uma competição no Kaggle para classificação de padrões de localização de proteínas subcelulares.	mAP de 0,57.

Agenda

1. Introdução
2. Fundamentação Teórica
3. Trabalhos Correlatos
4. Materiais e Métodos
5. Plano de Trabalho

Materiais e Métodos

- Metodologia
- Base de Dados
- Métricas de Avaliação
- Recursos Computacionais

Metodologia

Pré-Processamento

- Redimensionamento das imagens;
- Análise de rótulos incorretos;
- Estratificação multirrótulo para gerar um conjunto de validação;
- Aumentação de dados e aplicação de estratégias de reamostragem;
- Balanceamento multirrótulo.

Metodologia

Seleção de Modelos

- Modelos de classificação e segmentação;
  - Puzzle-CAM;
  - P-NOC;
  - C²AM-H.

Treinamento

- Weighted Loss;
- Focal Loss.

Teste

- Cada amostra do conjunto de teste será submetida ao modelo de classificação e segmentação.

Base de Dados

- A base de dados que será utilizada é do Atlas de Proteínas Humanas (HPA);
- A base faz parte da competição pública de Aprendizado de Máquina para classificação de padrões de localização de proteínas em células humanas[1];
- Adicionalmente ao conjunto de treinamento disponibilizado para a competição, será utilizado o conjunto de amostras da Seção Subcelular HPA (HPAv20) [19].

[1]https://www.kaggle.com/c/hpa-single-cell-image-classification

Base de Dados

- Há 19 rótulos diferentes (18 rótulos para localizações específicas e um rótulo para sinal negativo e não específico);
- As amostras são representadas por quatro filtros: a proteína de interesse (verde), os núcleos (azul), os microtúbulos (vermelho) e os retículos endoplasmáticos (amarelo).

0	Nucleoplasma	7	Complexo de Golgi	14	Mitocôndrias
1	Membrana nuclear	8	Filamentos intermediários	15	Aggresoma
2	Nucléolos	9	Filamentos de actina	16	Citosol
3	Centro fibrilar do nucléolo	10	Microtúbulos	17	Vesículas e padrões citosólicos pontilhados
4	Pontos nucleares	11	Fuso mitótico	18	Negativo
5	Corpos nucleares	12	Centrossomo		
6	Retículo endoplasmático	13	Membrana plasmática		

Base de Dados

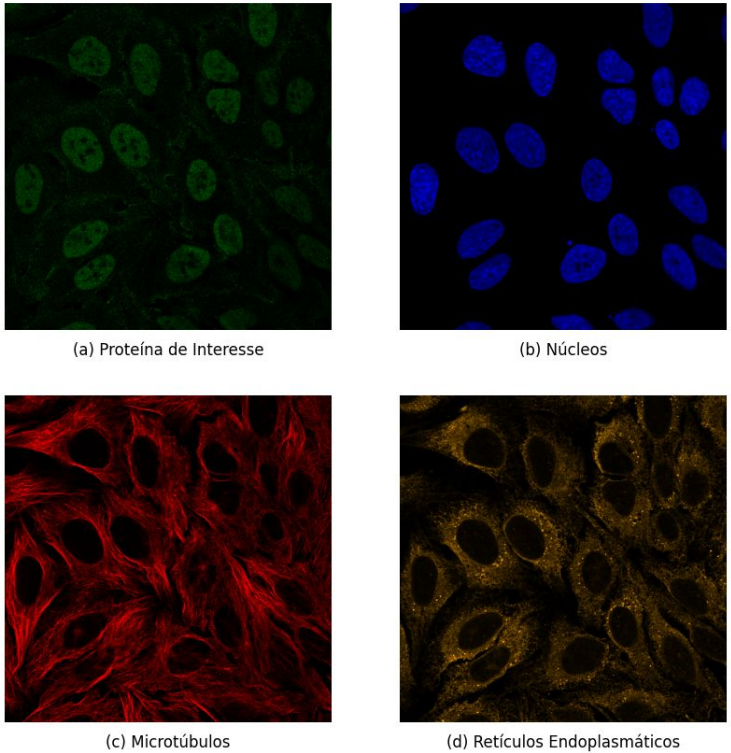

Figura: Exemplo dos filtros de uma amostra.

Base de Dados

- As imagens do conjunto de dados de treinamento foram anotadas de maneira que um ou múltiplos rótulos foram atribuídos a cada imagem ao avaliar os padrões de localização de todas as células;
- Devido à heterogeneidade celular, os rótulos atribuídos a cada célula em uma imagem podem não ser totalmente precisos;
- Em uma mesma imagem de uma população geneticamente idêntica, células individuais podem apresentar diferentes padrões de localização de proteínas.

Base de Dados

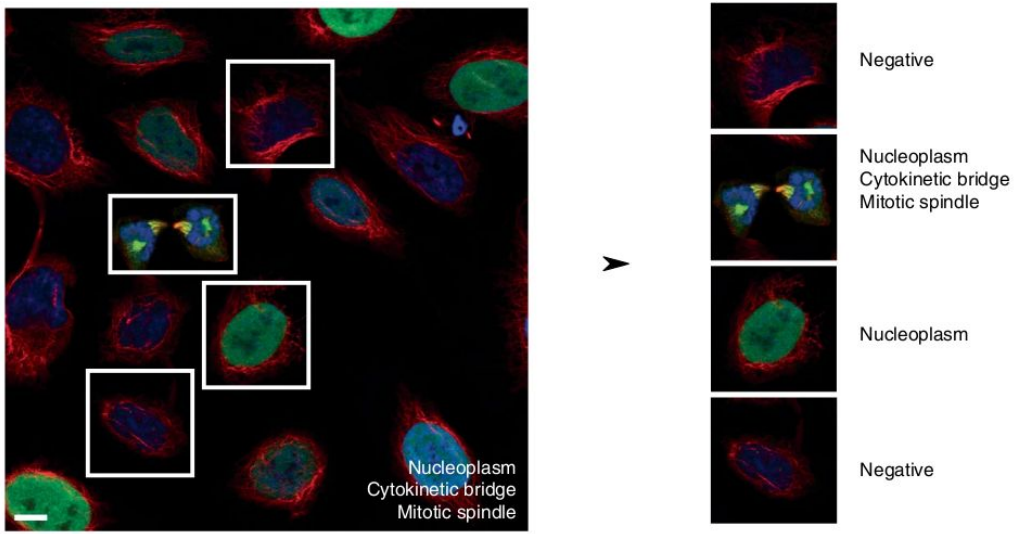

Figura: Exemplo da heterogeneidade celular em uma amostra [2].

Base de Dados

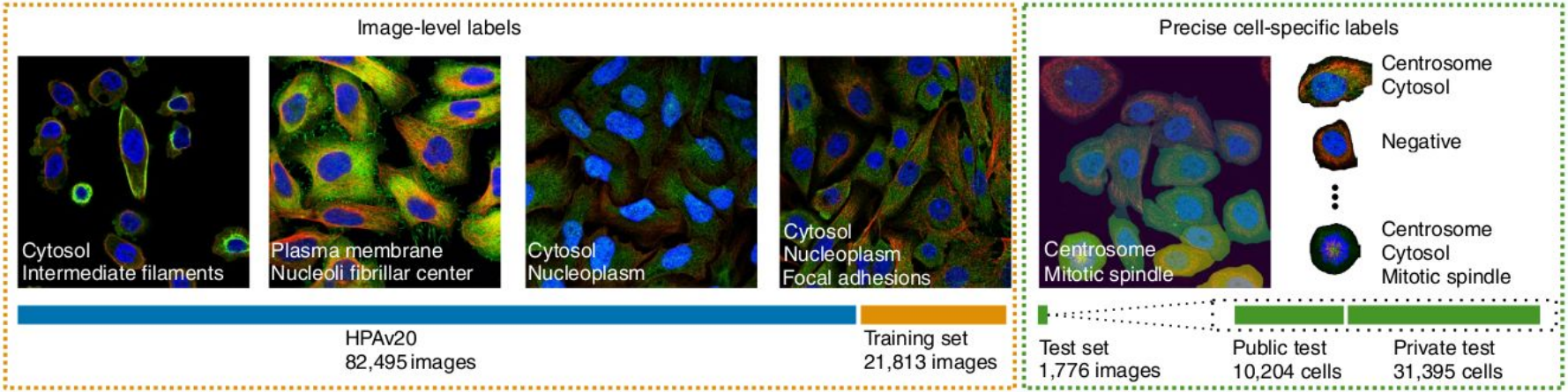

Figura: Divisão das imagens entre treino e teste na base de dados do HPA [2].

Base de Dados

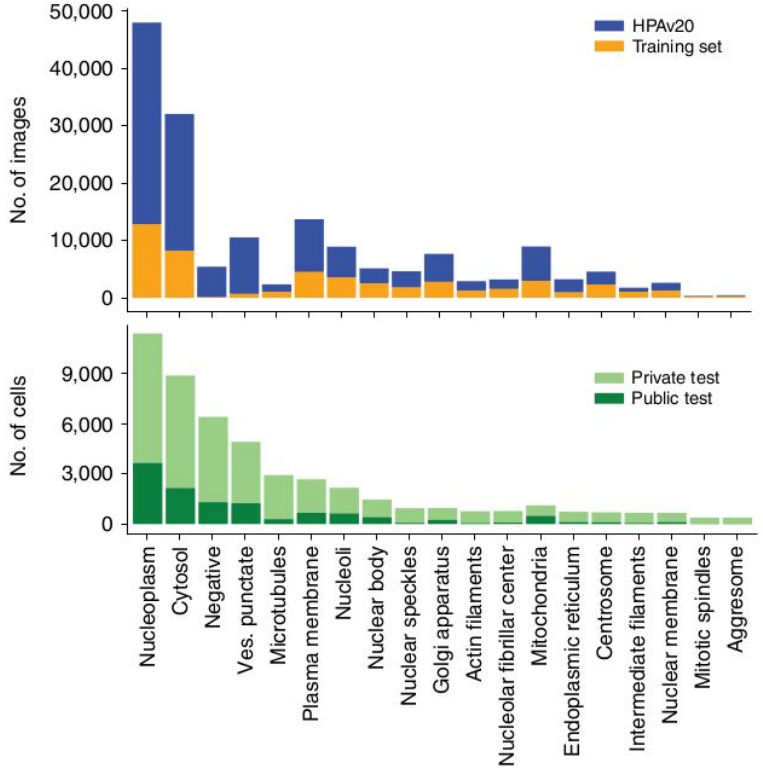

Figura: Número de imagens e células por classe nos conjuntos de treinamento e teste [2].

Métricas de Avaliação

- P: o número de casos positivos;
- N: o número de casos negativos;
- VP: o número de verdadeiros positivos;
- VN: o número de verdadeiros negativos;
- FP: o número de falsos positivos;
- FN: o número de falsos negativos.

Métricas de Avaliação

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{P} + \text{N}}$$

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

$$F_1 = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

Métricas de Avaliação

$$\text{IoU} = \frac{\text{Área da Interseção}}{\text{Área da União}}$$

$$\text{AP}c = \sum_{n} \left(R_n - R_{n-1}\right) P_n$$

- *Pn* e *Rn* são, respectivamente, a precisão e a revocação no ponto *n* da curva de precisão-revocação;
- *c* representa uma classe específica.

$$\text{mAP@60} = \frac{1}{|C|} \sum_{c \in C} \text{AP}_c$$

Recursos Computacionais

Bibliotecas

- NumPy;
- Scikit-learn;
- PyTorch;
- Matplotlib.

Ambientes

- Supercomputador Santos Dumont;
- CENAPAD.

Agenda

1. Introdução
2. Fundamentação Teórica
3. Trabalhos Correlatos
4. Materiais e Métodos
5. Plano de Trabalho

Plano de Trabalho

1. Obtenção dos créditos obrigatórios em disciplinas;
2. Participação no Programa de Estágio Docente (PED);
3. Seleção de uma base de dados;
4. Revisão bibliográfica;
5. Exame de Qualificação do Mestrado (EQM);
6. Pré-processamento da base de dados;
7. Seleção dos modelos de classificação e segmentação;
8. Treinamento dos modelos;
9. Realização de testes e análise dos resultados;
10. Documentação e publicação dos resultados;
11. Escrita do documento da dissertação;
12. Defesa da Dissertação de Mestrado.

Atividades	1º ano						2º ano					
	1	2	3	4	5	6	1	2	3	4	5	6
1	•	•	•	•	•	•						
2	•	•	•	•	•	•						
3		•	•									
4		•	•	•								
5				•								
6				•	•	•						
7					•	•						
8							•	•	•			
9										•		
10							•	•	•	•		
11										•	•	
12												•

Próximos Passos

- Busca de amostras de classes minoritárias;
- Treinamento de um modelo *baseline*;
- Análise de rótulos incorretos;
- Treinamento de modelos com uso de métodos que utilizam mapas de ativação de classe (CAMs).

Referências

[1] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walter, J. Wilson, and T. Hunt. *Biologia Molecular da Célula*. Artmed Editora, 2017.

[2] T. Le, C. F. Winsnes, U. Axelsson, H. Xu, J. M. Kaimal, D. Mahdessian, S. Dai, I. S. Makarov, V. Ostankovich, Y. Xu, E. Benhamou, C. Henkel, R. A. Solovyev, N. Banić, V. Bošnjak, A. Bošnjak, A. Miličević, W. Ouyang, and E. Lundberg. Analysis of the human protein atlas weakly supervised single-cell classification competition. *Nature Methods*, 19(10):1221–1229, 2022.

[3] F. Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[4] M. A. Nielsen. *Neural networks and deep learning*, volume 25. Determination Press, 2015.

[5] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[6] C. C. Chatterjee. Basics of the classic CNN, 2019. Disponível em: https://towardsdatascience.com/basics-of-the-classic-cnn-a3dce1225add. Acesso em: 12/09/2024.

[7] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

[8] J. Wang and L. Perez. The effectiveness of data augmentation in image classification using deep learning, 2017. Disponível em: https://arxiv.org/abs/1712.04621. Acesso em: 12/09/2024.

Referências

[9] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[10] S. Singh. What is data augmentation? Techniques, examples & benefits, 2022. Disponível em: https://www.labellerr.com/blog/what-is-data-augmentation-techniques-examples-benefits. Acesso em: 30/09/2024.

[11] S. Jo and I.-J. Yu. Puzzle-CAM: Improved localization via matching partial and full features. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 639–643. IEEE, 2021.

[12] L. David, H. Pedrini, and Z. Dias. P-NOC: Adversarial training of CAM generating networks for robust weakly supervised semantic segmentation priors. *Journal of Visual Communication and Image Representation*, page 104187, 2024.

[13] M. V. Boland, M. K. Markey, and R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry: The Journal of the International Society for Analytical Cytology*, 33(3):366–375, 1998.

[14] O. Z. Kraus, J. L. Ba, and B. J. Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

[15] D. P. Sullivan, C. F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, K. Smith, B. Revaz, B. Finnbogason, A. Szantner, and E. Lundberg. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature Biotechnology*, 36(9):820–828, 2018.

Referências

[16] S. Xiang, Q. Liang, Y. Hu, P. Tang, G. Coppola, D. Zhang, and W. Sun. AMC-Net: Asymmetric and multi-scale convolutional neural network for multi-label HPA classification. *Computer Methods and Programs in Biomedicine*, 178:275–287, 2019.

[17] W. Ouyang, C. F. Winsnes, M. Hjelmare, A. J. Cesnik, L. Åkesson, H. Xu, D. P. Sullivan, S. Dai, J. Lan, P. Jinmo, S. M. Galib, C. Henkel, K. Hwang, D. Poplavskiy, B. Tunguz, R. D. Wolfinger, Y. Gu, C. Li, J. Xie, D. Buslov, S. Fironov, A. Kiselev, D. Panchenko, X. Cao, R. Wei, Y. Wu, X. Zhu, K.-L. Tseng, Z. Gao, C. Ju, X. Yi, H. Zheng, C. Kappel, and E. Lundberg. Analysis of the human protein atlas image classification competition. *Nature Methods*, 16(12):1254–1261, 2019.

[18] S. Aggarwal, S. Gupta, D. Gupta, Y. Gulzar, S. Juneja, A. A. Alwan, and A. Nauman. An artificial intelligence-based stacked ensemble approach for prediction of protein sub-cellular localization in confocal microscopy images. *Sustainability*, 15(2):1695, 2023.

[19] P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, and E. Lundberg. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, 2017.

Classificação de Padrões de Localização de Proteínas Subcelulares Usando Métodos de Segmentação

Exame de Qualificação de Mestrado

Candidata: Juliana Midlej do Espírito Santo
Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Hélio Pedrini

Instituto de Computação
Universidade Estadual de Campinas

18 de Outubro de 2024