Proposta submetida ao edital Nº 1/2021 ProPes – Iniciação Científica UFABC

Projeto de Pesquisa

Auxílio ao Diagnóstico em Imagens Médicas com Modelos Interpretáveis

Resumo

Santo André, maio 2021

1 Introdução

Na última década, modelos de redes neurais artificiais (RNAs) se estabeleceram como o estado da arte em aplicações como classificação [KSH12, ALE+16] e segmentação de imagens [HDW+17]. Essa nova geração de redes neurais artificias é caracterizada pela profundidade (existência de múltiplas camadas), número grande de parâmetros ajustáveis (passando de milhões em alguns casos) e uso de bases de treinamento com número grande de exemplos (centenas de milhares ou mesmo milhões). Em estudos realizados com número pequeno de exemplos de treinamento, os sistemas gerados podem sofrer de baixa capacidade de *generalização*. Isto é, quando transportados para outros cenários, o sistemas de classificação não atingem desempenho tão bom quanto no estudo original. Modelos de redes neurais para classificação de imagens são considerados *opacos*: via de regra, não é possível inspecionar a mecânica de decisão do modelo ou compreender quais atributos da imagem de entrada resultaram na saída observada. Essas características: necessidade de grandes quantidades de dados, possibilidade de baixa generalização e falta de transparência, são desafios para a adoção de modelos de redes neurais em aplicações de imagens médicas, em geral marcadas por escassez de dados, necessidade de alta confiabilidade e transparência.

O presente texto propõe .... O restante do texto é organizado da seguinte maneira: na Seção 2, alguns conceitos fundamentais são apresentados.... A Seção 3 detalha os objetivos da proposta atual. A Seção 4 descreve os métodos a serem usados e apresenta o cronograma proposto para a execução do projeto.

2 Conceitos Fundamentais

O uso de inteligência artifical em aplicações de imagens médicas pode impactar significativamente os processos de diagnóstico. Modelos mais recentes de redes neurais têm reportado acurácia compatível com a de radiologistas [WPP+19], embora em certos casos o melhor resultado ainda seja obtido com uma combinação de diagnósticos, de um médico e do algoritmo[WPP+20]. Em uma pesquisa recente [SRM+21], médicos avaliam que avanços em inteligência artificial podem reduzir o tempo gasto em tarefas repetitivas e melhorar a acurácia do diagnóstico por imagens, embora temam uma desvalorização da *expertise* médica e a redução da participação de profissionais no processo de diagnóstico [SRM+21]. Além disso, médicos que realizam diagnóstico por imagens manifestam preocupação com a possibilidade de responsabilização devido a erros cometidos pelos algoritmos [SRM+21]. Para

que as novas ferramentas de diagnóstico auxiliadas por inteligência artifical sejam adotadas, é necessário que os profissionais possam entender o processo de decisão, aumentando sua confiança no processo.

Um traço comum a de modelos complexos, como redes neurais profundas, é a *subespecificação* [DHM+20], caracterizada pela existência de vários modelos diferentes (*i.e.* diferentes conjuntos de parâmetros aprendidos) que alcançam desempenho semelhante em treinamento. Ainda assim, esses modelos podem ter desempenho muito diferente quando levados para campo, resultando em baixa confiabilidade [DHM+20]. É fato conhecido que modelos de imagens médicas podem obter desempenho muito diferente quando implantados em ambientes distintos daquele onde foram obtidos os dados de treinamento [BZOR+18, MSG+20, DHM+20], estudos mais recentes tentam aplicar avaliações multi-instituição para diminuir este problema [MSG+20].

Outro fator de preocupação é que bases públicas de imagens médicas podem apresentar problemas relacionados à baixa qualidade das anotações [OR20] e/ou pela presença de subpopulações desconhecidas (*estratificação oculta*) [ORDCR20] que produzem distorções de treinamento e afetam a capacidade de generalização. Isso reforça a necessidade de modelos transparentes, em que a mecânica de decisão seja observável.

Algumas tentativas de criar modelos com explicações ou interpretáveis para auxílio a diagnóstico por imagens médicas incluem:

- Criar modelos de similaridade entre imagens, mostradas ao profissional de diagnóstico, em vez de gerar um diagnóstico [TH20].
- Gerar mapas de saliência, que mostrariam as regiões da imagem mais relevantes para o diagnóstico [SWP+20].

3 Objetivos e metas

A presente proposta tem dois objetivos principais:

- Realizar um levantamento bibliográfico sobre modelos interpretáveis em aplicações de diagnóstico médico por imagens.
- Avaliar um modelo interpretável de classificação no problema de diagnóstico, utilizando uma base de imagens públicas.

Com esses objetivos, nós pretendemos alcançar os seguintes resultados: em primeiro lugar, contribuir para a formação da bolsista em iniciação científica em uma área aberta de pesquisa e que tem alto potencial de impacto, habilitando-a a prosseguir em estudos mais avançados posteriormente. Em segundo lugar, estabelecer fundamentos para outros estudos futuros. Com o levantamento bibliográfico, esperamos mapear oportunidades de pesquisa na área de modelos interpretáveis, e o estudo de um modelo interpretável ajudará a criar *expertise* em nosso grupo de pesquisa, fornecendo um modelo que sirva de base de comparação para estudos futuros.

4 Materiais e Métodos

O projeto empregará bases públicas de imagens médicas, como...

Bases públicas de imagens médicas podem sofrer como problemas como baixa qualidade de rótulos [OR20] e *estratificação oculta* [ORDCR20], em que um subconjunto de exemplos pode apresentar características muito distintas do resto da amostra, distorcendo as medidas de desempenho de modelos. Nós pretendemos estudar as técnicas propostas na literatura [OR20, ORDCR20] para diagnóstico visual e auditoria de modelos, avaliando a viabilidade da aplicação ao nosso projeto.

Para o treinamento dos modelos, serão usados recursos computacionais dos laboratórios da UFABC. A implementação dos classificadores será baseada em bibliotecas de software livre, como *Numpy* [Oli15] e *Tensorflow* [AAB+16].

No início do projeto, uma pesquisa bibliográfica preliminar será conduzida, para construir um panorama do estado da arte em modelos interpretáveis na área de auxílio ao diagnóstico médico por imagens. Serão incluídos nessa revisão artigos mais gerais sobre interpretabilidade em aprendizado de máquina [Lip17, DVK17, GBY+19, MCB20, RCC+21], métodos específicos que resultam em modelos interpretáveis [LZH+20] e artigos que analisam questões como confiança em resultados de modelos de aprendizado de máquina [RU18, PSGH+21].

Como produto desse estudo, será gerado um texto de revisão, ...

Etapa	Mês											
	1	2	3	4	5	6	7	8	9	10	11	12
Revisão bibliográfica inicial	■	■	■	■								
Estudo de bibliotecas de software	■	■										
Implementação de modelos para classificação			■	■	■	■						
Revisão bibliográfica complementar					■	■	■	■	■	■		
Especificação de experimentos				■	■	■						
Escrita de relatório parcial					■	■						
Execução e análise de experimentos						■	■	■	■	■	■	
Refinar técnicas e experimentos									■	■	■	
Escrita do relatório final									■	■	■	■

Figura 1: Cronograma de execução da proposta

Referências

[AAB+16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[ALE+16] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C. Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206, 10 2016.

[BZOR+18] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Beth Percha, Thomas M. Snyder, and Joel T. Dudley. Deep Learning Predicts Hip Fracture using Confounding Patient and Healthcare Variables. *arXiv:1811.03695 [cs]*, November 2018. arXiv: 1811.03695.

[DHM+20] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv:2011.03395 [cs, stat]*, November 2020. arXiv: 2011.03395.

[DVK17] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. arXiv: 1702.08608.

[GBY+19] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretabi-

lity of Machine Learning. *arXiv:1806.00069 [cs, stat]*, February 2019. arXiv: 1806.00069.

[HDW+17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron C. Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[Lip17] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*, March 2017. arXiv: 1606.03490.

[LZH+20] Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and Scalable Optimal Sparse Decision Trees. *arXiv:2006.08690 [cs, stat]*, August 2020. arXiv: 2006.08690.

[MCB20] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *arXiv:2010.09337 [cs, stat]*, October 2020. arXiv: 2010.09337.

[MSG+20] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, January 2020. Number: 7788 Publisher: Nature Publishing Group.

[Oli15] Travis E. Oliphant. *Guide to NumPy*. CreateSpace Independent Publishing Platform, USA, 2nd edition, 2015.

[OR20] Luke Oakden-Rayner. Exploring Large-scale Public Medical Image Datasets. *Academic Radiology*, 27(1):106–112, January 2020. Publisher: Elsevier.

[ORDCR20] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020:151–159, April 2020.

[PSGH+21] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*, January 2021. arXiv: 1802.07810.

[RCC+21] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *arXiv:2103.11251 [cs, stat]*, March 2021. arXiv: 2103.11251.

[RU18] Cynthia Rudin and Berk Ustun. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. *INFORMS Journal on Applied Analytics*, 48(5):449–466, October 2018. Publisher: INFORMS.

[SRM+21] Jane Scheetz, Philip Rothschild, Myra McGuinness, Xavier Hadoux, H. Peter Soyer, Monika Janda, James J. J. Condon, Luke Oakden-Rayner, Lyle J. Palmer, Stuart Keel, and Peter van Wijngaarden. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific Reports*, 11(1):5193, March 2021. Number: 1 Publisher: Nature Publishing Group.

[SWP+20] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S. Gene Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv:2002.07613 [cs, eess, stat]*, February 2020. arXiv: 2002.07613.

[TH20] Johnson Thomas and Tracy Haertling. AIBx, Artificial Intelligence Model to Risk Stratify Thyroid Nodules. *Thyroid: Official Journal of the American Thyroid Association*, 30(6):878–884, June 2020.

[WPP+19] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. T. K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, pages 1–1, 2019.

[WPP+20] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2020.