Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

Treinamento Adversarial para Melhora de Interpretabilidade e Robustez em Redes Neurais Convolucionais Aplicadas a Radiografias de Tórax

Aluno: Lucas Santana da Silva (lss5@cin.ufpe.br)
Orientador: Fernando Maciano de Paula Neto (fernando@cin.ufpe.br)
Área: Aprendizagem de Máquina

Recife, Junho de 2021

Sumário

Resumo

Cada vez mais, vemos que a tecnologia tem desempenhado um papel fundamental no auxílio de profissionais de saúde na detecção de vários tipos de doenças. Com o grande aumento de dados disponíveis para treinamento assim como da potência dos computadores, algoritmos e técnicas mais e mais pautadas em Aprendizagem de Máquina têm sido desenvolvidos constantemente para essa e outras aplicações.

Uma dessas técnicas, as Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Networks*) tem crescido em problemas de saúde, mais especificamente na área de radiografia de tórax. A radiografia de tórax é utilizada pelos médicos como ferramenta de apoio ao diagnóstico, uma vez que as doenças pulmonares são identificáveis por apresentarem alterações na imagem em relação ao que é considerado padrão fisiológico ou padrão de um indivíduo saudável. As vantagens da redes CNN neste tipo de problema é que elas conseguem extrair automaticamente características de padrões de imagens, sem precisar da ajuda de um especialista que indique os padrões a serem coletados. Avaliar o treinamento de redes CNN com exemplos adversariais tem permitido entender melhor o funcionamento dessas redes, ver suas fragilidades e potencialidades.

O objetivo deste trabalho é o de vincular técnicas de treinamento adversarial com técnicas de visualização em CNN para avaliar a melhora na robustez e na interpretabilidade dos resultados, na área de aplicação citada acima.

Possíveis Avaliadores

A seguir, alguns avaliadores possíveis do trabalho:

- Cleber Zanchettin
- Germano Crispim Vasconcelos
- George Darmiton da Cunha Cavalcanti
- Nivan Roberto Ferreira Júnior

Introdução

A área da inteligência artificial tem se debruçado há um bom tempo em resolver problemas que envolvem a classificação e agrupar padrões, assim como também realizar previsões de valores contínuos [16]. Vários modelos computacionais foram desenvolvidos a partir de conceitos estatísticos e matemáticos para lidar com essas questões, mas a base dessas ideias não veio somente dessas áreas do conhecimento. Alguns desses modelos tiveram sua inspiração vinda das ciências biológicas. Esse é o caso das Redes Neurais Artificiais (RNA), cujo funcionamento é inspirado no comportamento neural.

As RNAs são modelos de aprendizagem de máquina capazes de aprender funções complexas através do ajuste de seus parâmetros internos mediante a apresentação de dados à rede. Elas são formadas por conjuntos de decisores lineares chamados de “neurônios” que são organizados em camadas. O crescimento da taxa de acurácia e robustez desses modelos, ao longo do tempo, fez com que eles fossem aplicados nos mais variados problemas [1][2], e fossem ficando mais complexos, permitindo adicionar mais camadas e funções matemáticas diversas, surgindo assim as Redes Neurais Profundas (do inglês *Deep Neural Networks*) que têm permitido reconhecer, classificar e predizer comportamentos muitos complexos envolvendo espaço-temporalidade, principalmente em imagens e em vídeos [3]. Um modelo de rede neural chamado Rede Neural Convolucional (ou CNN, do inglês *Convolutional Neural Network*) teve sua origem a partir do *Neocognitron* [4] e desde então é melhorada e modificada para os mais diversos fins.

Para muitos problemas em que a CNN é capaz de resolver, explicar o seu funcionamento de decisão passa a ser um requisito do usuário. Faz-se necessário entender o por quê de uma dada decisão do algoritmo, além de simplesmente saber qual teria sido essa decisão. Essa necessidade está presente quando lidamos com questões que envolvem a saúde humana [5] ou problemas que impactam a vida das pessoas, como na determinação de uma pena criminal [6] ou na contratação para uma empresa [7]. No entanto, redes neurais artificiais são modelos com uma baixa interpretabilidade, diferentemente de outros como árvores de decisão [8], e tal fator tende a piorar à medida que as redes ficam mais profundas.

Os trabalhos que envolvem o processamento de redes neurais profundas geralmente têm sido propostos para detectar objetos em imagens. Porém, em [9], uma arquitetura de Rede Neural profunda foi proposta para identificar a região da imagem que foi usada para que a rede fizesse sua classificação. Outros trabalhos seguiram melhorando a qualidade dos resultados encontrados [10]. Em resumo, o que esses trabalhos fazem é criar um mapa de ativação de classe (ou CAM, do inglês *Class Activation Maps*) com as regiões de excitação da rede neural durante a execução dos seus filtros. Existem bons resultados publicados utilizando este tipo de abordagem na área de saúde [11] [12], como na detecção da área de fraturas e câncer em radiografias de tórax.

Além disso, em 2013, descobriu-se que modificações intencionais leves e humanamente imperceptíveis nas imagens de entrada poderiam afetar drasticamente a eficiência de redes neurais [13], o que levantou ainda mais questões sobre o funcionamento interno desses algoritmos, tornando os chamados exemplos adversariais (do inglês *Adversarial Examples*, tradução livre) uma valiosa ferramenta nesses estudos e como forma de melhorar a interpretabilidade das redes [14] [15].

Neste projeto, o aluno estará envolvido no estudo e desenvolvimento de modelos de interpretação de Redes Neurais Convolucionais usando exemplos adversariais de diferentes tipos em imagens durante o treinamento do modelo. Trata-se de um assunto muito útil para diversas aplicações desenvolvidas nos dias atuais e promissor para o desenvolvimento de novos algoritmos. Ele fará uma comparação qualitativa de técnicas existentes, em termo de performance, assim como verificará se as formas de ataque adversarial afetam o desempenho da interpretação.

Objetivos

Realizar um estudo comparativo das redes neurais convolucional aplicada a problemas de classificação de imagens de radiografia utilizando técnicas interpretativas de resultados e exemplos adversariais no treinamento com diferentes tipos de ataques adversariais, avaliando os resultados encontrados com o uso de bases de dados reais, considerando medidas de acerto e métricas para interpretabilidade.

Metodologia

A metodologia deste trabalho será dividida em 4 etapas, sendo elas: Revisão bibliográfica e formação, Implementação da CNN interpretativa, Análise e comparação dos resultados e, por fim, a escrita e apresentação do trabalho.

Na etapa de revisão bibliográfica e formação, o aluno aprenderá ativamente a partir do estudo tanto da teoria e das bases matemáticas quanto da implementação prática de redes neurais artificiais através de artigos científicos, utilizando bibliotecas de código e frameworks na linguagem de programação Python, como o PyTorch. Em seguida, o estudante continuará a formação mediante estudo de redes neurais convolucionais, métodos de interpretabilidade para esses modelos e uso de exemplos adversariais.

Nesta etapa, o aluno implementará e executará os modelos de CNN interpretativas fazendo uso de métodos de treinamento adversarial. Neste projeto, o aluno poderá validar a hipótese de que a rede neural convolucional interpretativa, submetida a um conjunto de diferentes algoritmos adversariais em seu treinamento, torna-se mais robusta e interpretável sem prejudicar sua acurácia. Trata-se de uma verificação importante e poderá ter achados científicos relevantes para a área.

O estudante então analisará, na próxima etapa, os resultados dos modelos executados utilizando medidas e testes estatísticos. Ele poderá propor alterações nos modelos que suprimam as limitações das técnicas existentes e aumente a vantagem do uso dessas redes. É possível que o aluno possa identificar as falhas dos modelos e propor melhorias, sob supervisão do orientador.

Por fim, o aluno escreverá a monografia referente ao trabalho de conclusão de curso e preparará os slides para a defesa do projeto, modificando a monografia para possíveis correções que sejam apontadas durante a defesa.

Cronograma

O cronograma de trabalho do estudante seguirá conforme a tabela abaixo, apresentada abaixo, considerando que as atividades estão separadas em quatro grupos, a saber: Revisão bibliográfica e formação (I); Implementação da CNN interpretativa (II); Análise e comparação dos resultados (III); Escrita e apresentação do trabalho (IV).

Atividades	Meses		
	Jun	Jul	Ago
I			
II			
III			
IV			

Referências Bibliográficas

[1] BRAGA, A.P.; CARVALHO, A.C.P.L.F. & LUDERMIR, T.B. Redes Neurais Artificiais: Teoria e Aplicações. 2000.

[2] PAULA NETO, F.M., et al. Extreme learning machine for real time recognition of brazilian sign language. 2015 IEEE International Conference on Systems, Man, and Cybernetics. Hong Kong, China. 2015.

[3] VOULODIMOS, A., et al. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience. Vol. 2018. 2018.

[4] FUKUSHIMA, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, Vol. 36, n. 4, p.193-202. 1980.

[5] RAJPURKAR, P., et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv: 1711.05225. 2017.

[6] WASHINGTON, A. How to Argue with an Algorithm: Lessons from the COMPAS ProPublica Debate. Accepted for publication. The Colorado Technology Law Journal. Vol. 17, N. 1. 2019.

[7] Amazon scrapped ‘sexist AI’ tool. BBC News, 10 de out. de 2018. Disponível em: <https://www.bbc.com/news/technology-45809919>. Acesso em: 09 de jul. de 2020.

[8] QUINLAN, J.R. Induction of decision trees. Mach Learn. Vol. 1, p. 81–106. 1986.

[9] ZHOU, B., et al. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] ZHANG, Q.; WU, Y. N. & ZHU, S.C. Interpretable convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[11] DUNNMON, J.A., et al. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology. Vol. 290, n. 2, p. 537-544. 2019.

[12] YAMASHITA, R., et al. Convolutional neural networks: an overview and application in radiology. Insights into imaging. Vol. 9, n. 4, p. 611-629. 2018.

[13] SZEGEDY, C., et al. Intriguing properties of neural networks. arXiv: 1312.6199. 2013.

[14] DONG, Y., et al. Towards interpretable deep neural networks by leveraging adversarial examples. arXiv preprint arXiv:1708.05493. 2017.

[15] TAO, G., et al. Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples. NeurIPS. Montreal, Canadá. 2018.

[16] RUSSEL, S.; NORVIG, P. Inteligência artificial. Editora Campus, [S.l.], p.26, 2004.