PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

Ferramenta de auxilio a analise oncológica de mamografias por machine learning

Guilherme Freitas de Araujo

Proposta Projeto Final I

CENTRO TÉCNICO CIENTÍFICO - CTC

DEPARTAMENTO DE INFORMÁTICA

Curso de Graduação em Engenharia da Computação

Rio de Janeiro, Junho de 2018

Guilherme Freitas de Araujo

Ferramenta de auxilio a analise oncológica de mamografias por machine learning

Proposta de Projeto de Conclusão de Curso, apresentado ao curso de **Engenharia de Computação** da PUC-Rio como requisito parcial para a obtenção do titulo de Engenheiro de Computação.

Orientadora:

Prof. Marley Maria B. R. Vellasco

Coorientador:

Prof. Italo de Oliveira Matias

Rio de Janeiro

Junho de 2018

Resumo

Araujo, Guilherme; Vellasco, Marley; Matias, Italo. Ferramenta de auxilio a analise oncológica de mamografias por machine learning. Rio de Janeiro, 2018. 10p. Proposta de Projeto de Conclusão de Curso - Departamento de Informática. Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho tem como objetivo desenvolver uma ferramenta de auxilio ao diagnostico medicinal utilizando machine learning para analise de tumores em imagens mamograficas.

Palavras-chave

Machine learning; Inteligência artificial; Aprendizagem profunda;

Abstract

Araujo, Guilherme; Vellasco, Marley; Matias, Italo. Application to aid the oncological analysis of mammograms by machine learning. Rio de Janeiro, 2018. 10p. Capstone Project Report – Department of Informatics. Pontificial Catholic University of Rio de Janeiro.

This Project aims the development of a tool to assist medical diagnosis using machine learning for the analysis of tumors in mammographic images.

Keywords

Machine Learning; Artificial Intelligence; Deep Learning;

Sumário

1 Introdução

Câncer de mama é o segundo tipo de câncer com mais ocorrência em mulheres nos Estados Unidos da America (EUA) [1] e estima-se que por volta de 268,000 novos casos serão diagnosticados no ano de 2018 nos EUA e que mais ou menos 41,000 pessoas morrerão devido a essa patologia nessa mesma época [1]. A análise clinica de câncer de mama normalmente envolve o uso de diagnostico por imagem e a mamografia é umas das técnicas utilizadas na detecção precoce do câncer de mama [2]. Mamografia é um exame de rastreio por imagem utilizando raios-X que permite visualmente analisar o tecido mamário [2]. Recomenda-se que mulheres acima dos 40 anos, ou consideradas dentro do grupo de alto risco de câncer mamário, façam exames de mamografia anualmente, pois a detecção precoce desse câncer permite um tratamento eficiente para cura do mesmo [3]. Diferentes características de tumores, tecidos e o erro humano podem causar diagnósticos errados dentro da área de oncologia [4].

O objetivo desse projeto é desenvolver uma ferramenta de auxilio ao profissional de medicina na analise de câncer de mama em mulheres. Essa ferramenta utilizará redes convolucionais com aprendizagem profunda para classificar uma ou mais mamografias dentro da categoria de avaliação BI-RADS do American College of Radiology (ACR) [2]. Essa ferramenta facilitaria e aceleraria o processo de exame das imagens de raios-X resultantes da mamografia para detecção e categorização de tumores.

2 Situação Atual

Não existe nenhum exame ou grupo de exames que possa assegurar que a mulher não tenha câncer de mama, os exames clínicos avaliam características teciduais diferentes que serão utilizadas pelo profissional da área para interpretação [2]. Há diversos componentes passivos de analise pelo profissional durante um exame de imagem da mama, são eles: nódulos, calcificações, distorção arquitetural, assimetria, linfonodo intramamário, lesão de pele e ducto único dilatado [2]. A análise de todos esses componentes segue uma orientação complexa e resulta em uma categorização baseada em probabilidade com diferentes condutas que podem causar estresse desnecessário para o paciente ou a prorrogação de um tratamento que já deveria ter sido realizado. Estima-se que uma a cada cinco mamografias resultam em falsos negativos [1] e que 7% a 9% de mulheres que realizam mamografias anualmente receberam um resultado falso-positivo recomendando a etapa de biopsia [5].

Categorização de imagens por meio de redes convolucionais é um processo que tem evoluído muito ao longo dos anos e possui diversas ferramentas para facilitar o desenvolvimento de soluções robustas e assertivas relacionadas a diferentes áreas. O estudo que obteve a menor taxa de erro no desafio de classificação que utiliza o CIPHAR-10 (Canadian Institute For Advanced Research) foi o AutoAugment [6], que usou de técnicas de transformações das imagens da base de dados para aumentar a precisão de seu algoritmo.

Diversos estudos são feitos relacionando biologia e medicina com inteligência artificial, pois questões dentro dessa área comumente contem informações complexas e numerosas demais para uma analise humana ou baseada em software comum.

A utilização de ferramentas de auxilio ao profissional de medicina pode ter um impacto enorme em custos relacionados à saúde evitando tratamentos desnecessários alem de proporcionar um tratamento mais rápido e assertivo ao paciente.

3 Propostas e Objetivos do trabalho

O projeto tem como objetivo desenvolver uma ferramenta que classifique diferentes imagens oriundas de mamografias no formato DICOM (Digital Imaging and Communications in Medicine) de acordo com a classificação BI-RADS do ACR [2].

Será utilizada uma rede convolucional, ideal para classificação de imagens, pois dispensa boa parte do pré-processamento da base de dados, devido à inteligência do algoritmo de aprender quais filtros devem ser aplicados a imagem antes de passar para a camada de rede neural tradicional para classificação.

Algumas arquiteturas diferentes serão estudadas para se decidir qual deve ser abordada no projeto, utilizando benchmarks de classificação de imagens como CIPHAR-10, CIPHAR-100 e ILSVRC (ImageNet Large Scale Visual Recognition Challenge). Diferentes configurações dos parâmetros da rede e da base de dados utilizada para treinamento serão testadas visando aperfeiçoar a classificação resultante da rede. Dependendo do tempo necessário para se computar o treinamento da rede desenvolvido poderá ser interessante utilizar múltiplas GPUs (Graphic Processing Units) e/ou um cluster para acelerar tal etapa do projeto.

Deseja-se ter ao final do projeto uma aplicação que possa servir para uso clinico real auxiliando profissionais de medicina no dia-a-dia, e ter uma rede robusta que sirva como fonte de estudo para futuro desenvolvimento relacionando oncologia e processamento de imagens com inteligência artificial.

4 Plano de Ação

A primeira fase do projeto foi de estudo teórico da área de câncer de mama e de redes convolucionais para machine learning com imagens. A segunda fase envolveu a decisão de que linguagens, ferramentas e bibliotecas seriam utilizadas no desenvolvimento do projeto.

Python foi linguagem escolhida para o desenvolvimento da rede neural utilizada, pelo fato de ser uma linguagem que possui uma grande quantidade de ferramentas relacionadas a processamento de imagens e redes neurais. A principal biblioteca a ser utilizada será o TensorFlow, um framework para machine learning open source criado pela Google.

O segundo semestre envolve todo o desenvolvimento, treinamento, teste e aperfeiçoamento da rede utilizando a base de dados. Além do desenvolvimento do aplicativo que utilizara a rede para processar as imagens carregadas no programa. E por fim redigir o relatório final e apresentar o projeto para banca.

Cronograma do projeto

5 Referências

[1] SPIEGEL, Rebecca L.; MILLER, Kimberly D.; JEMAL, Ahmedin. **Cancer Statistics, 2018.** Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21442>. Acesso em: 20 jun. 2018.

[2] D'ORSI, C.J.; SICKLES, E.A.; MENDELSON, E.B.; MORRIS, E.A. ACR **BI-RADS® Atlas, Breast Imaging Reporting and Data System**. Reston, VA, American College of Radiology; 2013.

[3] OEFFINGER, K.C.; FONTHAM, E.T.; ETZIONI R. **Breast cancer screening for women at average risk: 2015 guideline update From the American Cancer Society**. *JAMA*. 2015. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21442>. Acesso em: 20 jun. 2018.

[4] GIESS, Catherine; FROST, Elisabeth; BIRDWELL, Robyn. **Difficulties and errors in diagnosis of breast neoplasms**. 2012. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/22824119>. Acesso em: 20 jun. 2018.

[5] HUBBARD R.A; KERLIKOWSKI, K.; FLOWERS, C.I.; YANKASKAS, B.C.; ZHU, W.; MIGLIORETTI, D.L. **Cumulative Probability of False-Positive Recall or Biopsy Recommendation After 10 Years of Screening Mammography: A Cohort Study**. Ann Intern Med. 2011. doi: 10.7326/0003-4819-155-8-201110180-00004. Disponível em: <http://annals.org/aim/article-abstract/474984/cumulative-probability-false-positive-recall-biopsy-recommendation-after-10-years?doi=10.7326%2f0003-4819-155-8-201110180-00004>. Acesso em: 20 jun. 2018.

[6] CUBUK, Ekin D.; ZOPH, Barret; MANE, Dandelion; VASUDEVAN, Vijay; LE, Quoc V. **AutoAugment: Learning Augmentation Policies from Data.** eprint arXiv:1805.09501. 2018. Disponível em: <https://arxiv.org/abs/1805.09501>. Acesso em: 20 jun. 2018.