Proposta de projeto final

Histórico do assunto

O câncer de mama é o segundo tipo mais frequente no mundo e o mais frequente na população feminina brasileira. Segundo estimativas do INCA, em 2012 e 2013, deverão ocorrer 52.680 casos novos de câncer de mama, com um risco estimado de 52 casos por 100 mil mulheres. Com exceção da Região Norte, onde o câncer do colo do útero lidera a incidência, o câncer de mama é o câncer mais incidente do Brasil. Sendo considerado, em geral, um câncer de bom prognóstico quando diagnosticado e tratado precocemente, as taxas de mortalidade por câncer da mama continuam elevadas no Brasil. Provavelmente, essas taxas de mortalidade mantêm-se elevadas porque a doença ainda é diagnosticada em estádios avançados. A sobrevida média após cinco anos na população de países desenvolvidos é aproximadamente 85%. Entretanto, nos países em desenvolvimento, a sobrevida fica em torno de 60%.[1]

O câncer de mama pode ser detectado em fases iniciais, em grande parte dos casos, aumentando assim as chances de tratamento e cura. Mamografia é uma radiografia das mamas feita por um equipamento de raios X chamado mamógrafo, capaz de identificar alterações suspeitas. [2] Se confirmada a suspeita da doença, um histograma é realizado.

A mamografia de rastreamento implica também em certos riscos que precisam ser conhecidos:

1) Resultados incorretos:

- Suspeita de câncer de mama, que requer outros exames, sem que se confirme a doença. Esse alarme falso (resultado falso positivo) gera ansiedade e estresse.
- Câncer existente, mas resultado normal (resultado falso negativo). Esse erro gera falsa segurança à mulher.

2) Sobrediagnóstico e sobretratamento: ser diagnosticada e tratada, com cirurgia (retirada parcial ou total da mama) quimioterapia e radioterapia de um câncer que não ameaçaria a vida. Isso ocorre em virtude do crescimento lento de certos tipos de câncer de mama. [3]

Descrição do problema

A modelagem do problema será constituída de:

Entrada: Imagens .PNG, 50 x 50 pixels de histogramas mamários, com seus devidos rótulos: 1 para a presença de câncer e 0 para a ausência do mesmo. Os rótulos das imagens estão presentes no nome da imagem e podem ser extraídos para formar uma lista.

Saída: A saída esperada será dada pela probabilidade de se conter câncer na imagem, com o valor variando de 0 a 1, depois será escolhido um limite para que os valores de probabilidade se transformem em previsões (se maior que 0,5, valor 1, senão valor 0, por exemplo)

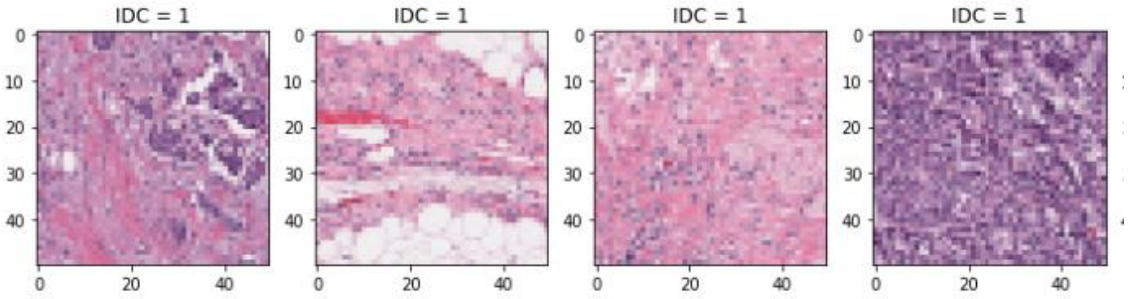
Tarefa de aprendizagem: Classificação binária.

Conjunto de dados de entrada

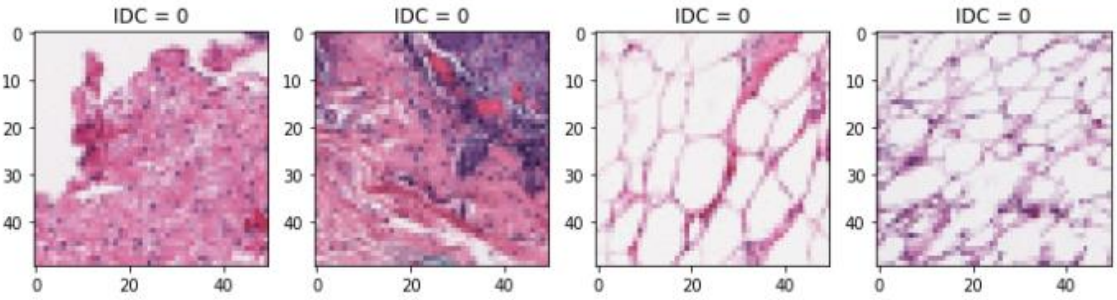
Os dados utilizados nesse trabalho serão imagens histológicas de mamas, com diagnósticos positivos e negativos para câncer. O dataset pode ser encontrado em https://www.kaggle.com/paultimothymooney/predict-idc-in-breast-cancer-histology-images/data.

Esse dataset possui mais de 200 mil imagens (.PNG) 50x50 pixels, tanto com resultados positivos e negativos. Essas imagens serão utilizadas para treinamento da rede neural profunda.

Exemplo de imagens com resultado positivo



Exemplo de imagens com resultado negativo



Descrição da Solução

A solução para o problema é utilizar o método de transfer learning na rede VGG16 com pesos do dataset imagenet e treinar uma Rede Neural Convolucional, para que, quando forem imputadas imagens de histogramas mamários, a rede identifique se estas possuem câncer ou não. Algumas variáveis serão estudadas, são elas: As camadas que serão congeladas da rede, para constatar qual é a última camada genérica o suficiente para o problema e a quantidade de épocas de treinamento.

Modelo de referência (benchmark)

O modelo de referência que será utilizado será o artigo [4], nesse estudo, o pesquisador utilizou técnicas de processamento de imagem e obteve uma tabela de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. Com esses dados é possível calcular o F1 score desse estudo e posteriormente compara-lo com o obtido por esse trabalho final. Além desse modelo, o trabalho será baseado no repositório apresentado durante a Udacity live do pesquisador Fabio Perez, disponível em https://github.com/fabioperez/udacity-live-presentations/blob/master/Apresentacao_2/Fine-tuning%20com%20Keras.ipynb.

Métricas de avaliação

A métrica de avaliação escolhida é o F1 score. Nesse caso, pelo fato de o dataset já ser desbalanceado quanto a incidências de câncer, essa métrica já seria uma boa escolha para avaliação do modelo. Além disso, a ocorrência de falsos positivos e falsos negativos é extremamente indesejada, o que ajuda a afirmar a escolha da métrica.

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

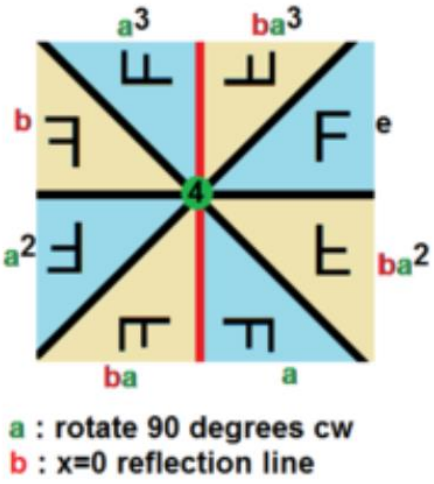
Design do projeto

O projeto será composto das seguintes etapas:

- Extração dos dados: os dados estão em um formato .PNG e contém as informações descritas na seção “Conjunto de dados de entrada”.

- Tratamento das imagens: será promovido o tratamento necessário: normalização da intensidade de cor dos pixels e outras técnicas que forem pertinentes.

- Data augmentation: com técnicas de aumento no número de imagens usando permutação, é possível aumentar o número dos conjuntos de validação e treino,caso seja necessário, bem como diminuir o viés causado pelo posicionamento da imagem. Por exemplo, se mais mamas esquerdas nesse dataset possuírem cancêr, o modelo ficará enviesado, diagnosticando mamas esquerdas com maior probabilidade de câncer. Além da rotação da imagem em graus diferentes, são realizadas as permutações, como mostrado na imagem abaixo.



- Dividir em treino, validação e teste (mantendo as proporções): como o dataset está desbalanceado, é importante dividí-lo em treino e validação em proporções iguais. Primeiramente será testado se, igualando o número de imagens positivas e negativas para câncer no treinamento, obtém-se bons resultados, o que é esperado, pois a técnica de transfer learning requer menos imagens se comparada ao treinamento de uma rede do zero. A divisão entre treino, validação e teste será aproximadamente 70%, 15% e 15% respectivamente.

- Treino da rede: em primeira instância, o ideal seria treinar uma rede convolucional do zero, mas como pode demorar muito tempo para a rede ser treinada, será utilizada a

técnica de transfer learning, onde utiliza-se uma rede já treinada (o keras do python possui algumas) e treina-se apenas as camadas mais externas e a de classificação, com o objetivo de otimizar menos pesos da rede. Essa tática funciona pois os pesos mais internos da rede geralemente são formas simples, como bordas, frequencias, etc. Essas formas estão presente em todas as imagens. As camadas mais externas possuem formas mais complexas, adequadas às imagens de treinamento, que serão removidas e retreinadas.

- Diagnóstico: A etapa final é produzir o diagnóstico, compara-lo com os diagnósticos reais e verificar o desempenho do modelo.

Referências

[1] Porto, M. A. T. ; Teixeira, L. A. ; Silva, R. C. F.; Aspectos Históricos do Controle do Câncer de Mama no Brasil; Revista Brasileira de Cancerologia 2013; 59(3): 331-339

[2] http://www.inca.gov.br/wcm/outubro-rosa/2015/deteccao-precoce.asp

[3] http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/deteccao_precoce+

[4] Silva, T. C.; Boos, C. F.; Medeiros, D. S.; Lobato, E. M.; de Azevedo, F. M.; Detecção automática de tumores em mamografias utilizando técnicas de processamento digital de imagem; XXIV Congresso Brasileiro de Engenharia Biomédica, 2014. Link: https://wiki.sj.ifsc.edu.br/wiki/images/2/2c/DiegoMedeiros-ArtigoOrientados1.pdf