Breast Cancer Classification using Computer Vision

Chandramouli Yalamanchili
DSC680 - T302 Applied Data Science (2215-1)
https://chandu85.github.io/data-science

Which Domain?

I chose to work on image classification or computer vision as my 2nd project, especially how it is currently being used in medical domain. For this project I chose the use case of detecting the breast cancer using CNN model.

As I am looking to build a machine learning model to identify the breast cancer based on the histology images. I am going to review the work that is done already in the medical field using the image classification and computer vision with the goal of understanding different types of model that are being used and the advancements that are achieved in medical domain using data science.

Below are several references I am planning to refer as part this project:

1. F. Milletari, N. Navab and S. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 2016, pp. 565-571, doi: 10.1109/3DV.2016.79.
    - https://arxiv.org/pdf/1606.04797.pdf
    - I will review this reference to understand the use of CNNs to solve the problems from both computer vision and the medical image analysis fields.
2. Junfeng Gao, Yong Yang, Pan Lin, Dong Sun Park, "Computer Vision in Healthcare Applications", Journal of Healthcare Engineering, vol. 2018, Article ID 5157020, 4 pages, 2018. https://doi.org/10.1155/2018/5157020
    - https://www.hindawi.com/journals/jhe/2018/5157020/
    - This paper talks about how computer vision can help in different aspects of medical domain, I will review this paper and try to understand the contributions of computer vision in health care domain.
3. Nadim Mahmud, Jonah Cohen, Kleovoulos Tsourides, Tyler M. Berzin, Computer vision and augmented reality in gastrointestinal endoscopy, Gastroenterology Report, Volume 3, Issue 3, August 2015, Pages 179–184, https://doi.org/10.1093/gastro/gov027
    - https://academic.oup.com/gastro/article/3/3/179/613495
    - This reference talks about the Augmented Reality (AR) using the computer vision and how it can help with the endoscopy set-up. I will review this to understand how computer vision helps in building AR and there by helps with endoscopy.
4. Esteva, A., Chou, K., Yeung, S. et al. Deep learning-enabled medical computer vision. npj Digit. Med. 4, 5 (2021). https://doi.org/10.1038/s41746-020-00376-2
    - https://www.nature.com/articles/s41746-020-00376-2
    - I will review this article to understand the benefits of medical imaging benefits in different domains of health care. I will review this article to also understand the hurdles in using these technologies in real world deployments.

5. J. Thevenot, M. B. López and A. Hadid, "A Survey on Computer Vision for Assistive Medical Diagnosis From Faces," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1497-1511, Sept. 2018, doi: 10.1109/JBHI.2017.2754861.
    - https://www.researchgate.net/publication/320250581_A_Survey_on_Computer_Vision_for_Assistive_Medical_Diagnosis_From_Faces
    - This survey gives an overview of the various approaches to assess facial symptoms and to eventually provide further help to the practitioners. I will review this to understand the different approaches discussed in this article.
6. Chaohui Wang, Nikos Komodakis, Nikos Paragios. Markov Random Field Modeling, Inference & Learning in Computer Vision & Image Understanding: A Survey. Computer Vision and Image Understanding, Elsevier, 2013, 117 (11), pp.1610-1627. ff10.1016/j.cviu.2013.07.004ff. ffhal-00858390v2f
    - https://hal.archives-ouvertes.fr/hal-00858390/document
    - This article discusses about using Markov Random Fields (MRFs) models and graph-based models in computer vision. I would like to review to understand how these models work and see if they offer any advantage compared to CNN models.
7. Tim F. Cootes and Christopher J. Taylor "Statistical models of appearance for medical image analysis and computer vision", Proc. SPIE 4322, Medical Imaging 2001: Image Processing, (3 July 2001); https://doi.org/10.1117/12.431093
    - https://www.spiedigitallibrary.org/conference-proceedings-of-spie/4322/0000/Statistical-models-of-appearance-for-medical-image-analysis-and-computer/10.1117/12.431093.pdf
    - This article discusses a different approach of using statistical models to come up with the target image that can be used to solve a variety of different problems. I will review to understand what was done and how it helped in solving different problems in medical domain.
8. Sultana, F., Sufian, A., & Dutta, P. (2018, November). Advancements in image classification using convolutional neural network. In 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) (pp. 122-129). IEEE.
    - https://arxiv.org/pdf/1905.03288.pdf
    - This article discusses different components of CNN. This paper also explains different CNN architectures for image classification. I would like to review this paper and gain more broader understanding about CNN models and recent advancements happened to the CNN models.
9. O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., ... & Walsh, J. (2019, April). Deep learning vs. traditional computer vision. In Science and Information Conference (pp. 128-144). Springer, Cham.
    - https://arxiv.org/pdf/1910.13796.pdf
    - This paper discusses how computer vision was handled before Deep Learning has emerged and how traditional computer vision techniques are different than deep learning techniques and how we can combine both. I want to review this paper to understand traditional computer vision techniques and understand how to combine those within deep learning solution to gain more benefits using computer vision.
10. Data-flair-training python projects. Project in Python – Breast Cancer Classification with Deep Learning.
    - https://data-flair.training/blogs/project-in-python-breast-cancer-classification/

- I will be using this article as reference in building my deep learning CNN model.
- https://www.kaggle.com/paultimothymooney/breast-histopathology-images/. This is where I will be getting the input dataset for my project as well.

11. Breast cancer statistics 2020 - https://www.nationalbreastcancer.org/wp-content/uploads/2020-Breast-Cancer-Stats.pdf

Which Data?

I am planning to use the below dataset from Kaggle with IDC (Invasive Ductal Carcinoma) cancer histology images to train the model as part of this project and we are going to for this project.

Dataset Link - https://www.kaggle.com/paultimothymooney/breast-histopathology-images/

- This dataset is the collection of 277,524 patch images, these patch images were extracted from 162 complete breast cancer specimen images that are captured with 40x magnification.
- Out of these 277k sample images collection, 198K images are samples of negative samples indicating non breast cancer patches, and 78K images are samples of positive samples diagnosed as impacted by breast cancer.
- IDC (Invasive Ductal Carcinoma) is the type of cancer that these patches are classified against.
- Each patch or image file in this dataset is segregated into several folders based on the complete patient id. Within each folder, the images are again segregated by the class of the patch depending on whether it is classified as IDC or not.
- Each individual image follows a specific file name formatting as well:
  - uxXyYclassC.png is the format being used for file name.
  - 'u' is the patient ID
  - 'X' is the x-coordinate of where the particular patch image was copied from.
  - 'Y' is the y-coordinate of where the particular patch image was copied from.
  - 'C' indicates the classification, where 0 indicates non-IDC and 1 indicates IDC.
  - 10253idx5x1351y1101class0.png is an example patch image file name.

Research Questions? Benefits? Why analyze these data?

- Breast cancer is the most common type of cancer in women, in 2020, 276,480 new cases were estimated to be diagnosed with invasive breast cancer and 48,530 non-invasive breast cancer cases were estimated.
- Invasive ductal carcinoma (IDC) is the most common form of breast cancer.
- As one can imagine, accurately identifying and categorizing breast cancer subtypes is an important clinical task. Using the automated methods in categorizing the breast cancer subtype can be of great help in saving time as well as reducing incorrect diagnosis.
- Through this project I would like to understand how deep learning based computer vision can help in diagnosing a certain type of cancer efficiently and accurately.

Method?

- I am planning to use Python for this project.
- I will initially do some data visualization to understand any trends I can derive out of the data.
- I will then build a deep learning model to evaluate the input images and come up with prediction for the breast cancer.

Potential Issues?

- At this time, I am not anticipating any major roadblocks or issues with respect to data.
- The only issue I might run into is with ingesting the huge data, as the input data is in the form of images, I might run into CPU constraints working with the data.
- One other challenge for me is that I have not spent a lot of time on EDA for image datasets before, so I might have to explore a little bit and gain some knowledge on what I can do with respect to EDA for image datasets.

Concluding Remarks

Breast cancer is one among the foremost common forms of cancer in American women, it's estimated that within the year of 2020, approximately 30% of the new cancer diagnosed women are carcinoma. Of all the carcinoma, the Invasive Ductal Carcinoma (IDC) is the most common subtype. Within the year of 2020, IDC subtype accounted for 85% of total carcinoma cases.

To grade the full mount samples, pathologists typically target the regions which contain the IDC. As a result, one of the critical pre-processing steps is to define the precise regions of IDC within an entire mount slide. Using the automatic process to evaluate each of the patches or the mount samples would be great help in saving time and increasing the accuracy of the diagnosis.

Through this project we will build a Keras based CNN (Convolutional Neural Network) classifier model that would evaluate the patches collected from several whole mount slide images and accurately classify a histology image as benign or malignant.