Breast Cancer Detection in Mammograms using Deep Learning Techniques

Description, Objectives, Ethics & Resources

University of St Andrews - School of Computer Science

Supervisor: Dr David Harris-Birtill

Adam Jaamour

5th June, 2020

1 Description

Breast cancer is one of the most common forms of cancer amongst women in the UK, with statistics indicating that 1 in 7 females will be diagnosed with breast cancer in their lifetime. Indeed, 55,200 new breast cancer cases are reported every year in the UK, of which an average of 11,400 lead to death (20% mortality rate) [1]. However, many other types of cancer types exist, ranging from lung and prostate cancers to colon and bladder cancers to name a few [2].

The goal of this project is to design a deep learning pipeline that can learn how to detect cases of breast cancer in mammograms (X-ray pictures of breasts) by detecting the presence of tumours, whether they are benign (non-cancerous) or malignant (cancerous), and their location in the mammogram. The deep learning algorithm will learn the underlying patterns of a large dataset of mammograms to carry out the aforementioned tasks. The motivation behind this project is to allow a technique for early and accurate breast cancer detection to prevent unnecessary treatments in cases of false positives and to prevent late treatments due to false negatives. Ultimately, the target of this project is to combine it with deep learning algorithms developed across other projects supervised by Dr David Harris-Birtill (past and present). This will allow a general artificial intelligence system capable of detecting multiple forms of cancer with higher accuracies than radiologist diagnostics.

Parts of the work undertaken during this project will be conducted as a group comprised of two other members. The tasks in common involve the data cleaning and pre-processing, the results output and the implementation of a basic deep learning pipeline. The reasoning for these common tasks is to allow each group member to further explore deep learning techniques on their own using the common basic pipeline as a starting point and to compare final results at the end. Section 2 covers in more detail which tasks will be conducted personally and in a group.

2 Objectives

Objectives are divided into primary and secondary objectives, where primary objectives correspond to essential milestones (organised as a linear timeline to follow during the 11-week duration of the project), while secondary objectives complement these essential milestones.

Primary objectives (estimated dissertation timeline)

1. The first primary objective conducted individually consists of writing a complete and extensive literature review by the end of the 3rd week (some papers may be shared amongst group members), which will:
    - cover the background of deep learning techniques applied to the field of cancer detection (and disease detection using medical imagery);
    - review state of the art deep learning methods applied to the detection of breast cancer in mammograms, which will be used to guide the research towards promising areas and govern the choice of techniques to implement and explore;

   - identify results achieved using different methods (models, optimisation techniques, etc.).

2. The second objective relies on achieving a prototype implementation of a deep learning pipeline by the end of the 6th week in a group, including:

   - selecting a basic machine learning model that is proven to work on breast cancer detection based on the explored literature;
   - implementing a basic machine learning/deep learning pipeline using Python machine learning software (e.g. Tensorflow, Keras) capable of running on the GPUs using a subset the cleaned and pre-processed data into the selected model;

3. Next, the third primary objective consists of individually exploring more advanced/optimised deep learning algorithms in a fully functional pipeline by the end of the 8th week, consisting of:

   - extending and optimising the basic pipeline created in the third primary objecting for the task of breast cancer detection;
   - designing a full deep learning pipeline capable of running on the entire dataset in batches using efficient memory allocation optimisation methods due to the large size of the dataset (see Section 4 and the limited amount of RAM available).

4. Finally, the fourth primary objective involves the results' evaluation and finishing the dissertation write-up by the end of the 11th week (parts will be written during the previous objectives as well):

   - evaluating the final deep learning pipeline's results by employing common evaluation metrics and output formats used in the basic pipeline and in other papers. This will allow direct comparisons of the final results with:
     - the basic pipeline's results developed in common (from the second objective);
     - the results achieved by the two other group members;
     - the results from papers researching breast cancer detection identified in the literature review.
   - individually writing the dissertation paper;
   - polishing the pipeline.

Secondary objectives These secondary objectives serve as additions to the primary objectives that are hard to achieve, but that will be attempted if time permits it.

- Implement segmentation (the task of localising the tumour in the mammogram) in the pipeline, on top of the task of cancer detection.

- Explore optimisation methods to reduce the training time required, such GPU acceleration.

- Examine the trade-off between speed and accuracy when using different algorithms, model hyperparameters and optimisation techniques.
- Achieve a high-level quality of code by following professional coding guidelines (e.g. PEP8 coding guidelines for Python [3]), along with detailed documentation and useful comments throughout the code. This includes following software-engineering practices of version controlling the code with git and covering vital code sections with a testing suite.

3 Ethics

The deep learning model will require real-life human data to learn its underlying patterns in order to detect cases of breast cancer in mammograms and to evaluate its performance. Therefore, the main ethical concern when using this sensitive medical data is whether it can be traced back to the original patient. As a result, fully anonymised, open-source and public datasets are used (described in Section 4). After filling out the Preliminary Ethics Self-Assessment Form, it is determined that a full Ethics Application needs to be submitted and approved by the Ethics Committee.

4 Resources

This project will make use of three open-source anonymised public datasets:

- "Digital Database for Screening Mammography (DDSM)" dataset [4].
- "Curated Breast Imaging Subset of DDSM" dataset [5], available online from The Cancer Imaging Archive [6]. The dataset contains a total of 10,239 images gathered from 1,566 patients across 6,775 studies [5]. This dataset is a subset of previously mentioned DDSM dataset, containing only cases with benign and malignant tumours (no normal cases).
- "mini-MIAS" database, a smaller dataset of mammograms containing 322 images in Portable Gray Map (PGM) format with associated ground truth data [7].

Datasets used in published papers, such as the "Breast Cancer Wisconsin (Diagnostic) Data Set", were considered, but not chosen as they were not cleaned and processed for machine learning tasks, and features were already extracted from the images, reducing the flexibility for deep learning to learn features from scratch.

Due to the complex nature of the dataset and the project's aims to explore deep learning techniques using dedicated machine learning software (e.g. Tensorflow, Keras), powerful computing resources will be required in the form of Graphical Processing Units (GPU) provided by the School of Computer Science and remotely accessed via SSH. A lab machine equipped with a GPU running on CentOS has already been assigned for the duration of the project. In case of further computing power or memory being required, Dr David Harris-Birtill's (dissertation supervisor) machine (CASE) could be used, providing a total of 256 Gb of RAM and 2 GPUs (NVIDIA GeForce GTX 1080Ti).

5 Supervision

The supervisor for this project is Dr David Harris-Birtill. Weekly video meetings via Microsoft Teams are planned, as well as additional weekly meetings between team members.

References

[1] Cancer Research UK. Breast cancer statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer, 2020. [Online] Accessed: 2020-05-28.

[2] V Cokkinides, J Albano, A Samuels, ME Ward, and JM Thum. American cancer society: Cancer facts and figures. *Atlanta: American Cancer Society*, 2005.

[3] Pep 8: Style guide for python code. https://www.python.org/dev/peps/pep-0008/. [Online] Accessed: 2020-06-01.

[4] Richard Moore Michael Heath, Kevin Bowyer, Daniel Kopans and W. Philip Kegelmeyer. The Digital Database for Screening Mammography. In *Fifth International Workshop on Digital Mammography*, pages 212–218. Medical Physics Publishing, 2001.

[5] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4, dec 2017.

[6] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, dec 2013.

[7] J Suckling. The mammographic image analysis society digital mammogram database exerpta medica. *International Congress Series*, 1069:375–378, 1994.