MSBD 6000B Project 3 - Breast Cancer Classification

Arnor Ingi Sigurdsson*,[1], Chan Ngae Chau*,[2], Lin He*,[3], Zhan Yunzhen*,[4] and Gao Cong*,[5]

*Hong Kong University of Science and Technology, Hong Kong

ABSTRACT This project is based on the paper: Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification.

KEYWORDS deep learning; image recognition; neural networks; whole slide tissue;

Introduction

Deep learning models have seen a surge in popularity after recent successes in computer vision tasks [4, 6, 7]. The models in question are usually evaluated using the ImageNet database, which is a large visual database designed for use in visual object recognition research [1]. With these advancements in mind, it is no wonder that various fields are interested in the application of computer vision. One drawback with using ImageNet however is the small scale of the images. ImageNet pictures are composed of pixels in the thousands, while the images used for disease screening can be in the millions (commonly called gigapixel images). Currently, enough computational power is not commercially available to feed these gigapixel images directly into the computer vision models. Various approaches have been proposed to solve this issue [2, 8, 5]. In this project, we attempt a batch based approach in order to classify high resolution breast cancer dicom images [3].

Methods

Data

The data is composed of a total of 410 breast cancer images. The supplied data is split 330, 40 and 40 in a training, validation and test sets respectively (a proportion of roughly 80%, 10% and 10%). The training and validation sets are normally used for training and hyperparameter estimation, while the test set is to be used for generalization error estimation. The images are split in a roughly 3:1 proportion between non-cancer and cancer dicom images. An example of an image used as input for the classification model can be seen in figure 1.

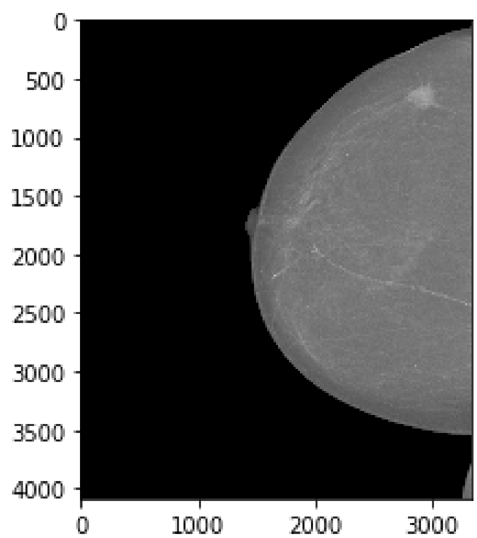

Figure 1 Example of an image used as input to the model used for breast cancer classification. The image shown is of a cancerous.

Preprocessing

In order to ensure that all features (in the case of images, pixels) are on the same scale, the images were normalized

Manuscript compiled: Friday 15th December, 2017

[1]20483569, aisigurdsson@connect.ust.hk

[1]20411891, ncchan@connect.ust.hk

[1]20476097, hlinam@connect.ust.hk

[1]20472766, yzhanae@connect.ust.hk

[1]20385614, cgaoad@connect.ust.hk

to a scale between -1 and 1. This is to avoid the learning rate affecting features differently depending on their scale during backpropagation. The images in this project have only one channel (due to them being grayscale), therefore the mean and standard deviation was calculated for that channel and used to normalize the images prior to training.

Model

As mentioned, the images are too large in order for them to be stored in a commercial computer's memory during training. Additionally, downsampling the image means that important detail needed for the actual classification might be lost. A patch based approach seemed to be a good approach to tackle both of the problems outlined above and was therefore attempted in this project. During the training process, patches of 80×80 pixel sizes are extracted from a given image. Training consists of two steps: An EM (expectation maximization) step and a logistic regression step. The EM Step is to determine the patch-level label (i.e. whether a given patch in the image is discriminative or not). The Logistic Regression Step is to determine the image-level label (i.e. whether an image contains cancer or not) from patch-level label. The process is as follows:

1. EM Step
   (a) All patches are labeled as 1 (indicating if a patch is discriminative or not).
   (b) Gaussian smoothing is applied over all patches to detect spatial relationships within a given patch.
   (c) Train a convolutional neural network (CNN) for few epochs to give prediction probability of all patches.
   (d) Sort all the probabilities in ascending order.
   (e) For a defined percentile, all patches with probabilities greater than threshold are labeled with 1. Others are labeled with 0.
   (f) The patches with label 1 are trained with the CNN again. Step (c) is repeated until converge.
2. Logistic Regression Step
   (a) With all the patch-level labels and image-level labels, a logistic regression model is created.
   (b) Cross-validation is used to obtain the best model parameters.

Results

Discussion

Dataset

Before experiments, it was found that the images in both validation dataset and test dataset are actually from the training dataset. Keeping this structure introduces a certain bias into future predictions, as observations in the test and validation sets have been used for training. Therefore the correct practice is to remove these images from the training set in order for the three sets to be mutually exclusive.

Limitations

The group believes themselves to have been mostly successful in implementing the patch based approach detailed in [3]. The main limitation of the project however is that due to the large images size, it is very time consuming to train and test the model implemented. This resulted in a lot of downtime which caused software development to progress at a slow pace.

Literature Cited

[1] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database, 2009.

[2] Krzysztof J. Geras, Stacey Wolfson, S. Gene Kim, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *CoRR*, abs/1703.07047, 2017.

[3] Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. *CoRR*, abs/1504.07947, 2015.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks, 2012.

[5] Li Shen. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *CoRR*, abs/1708.09427, 2017.

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2015.

[8] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. *CoRR*, abs/1612.05968, 2016.