Convolutional Neural Networks and the Analysis of Cancer Imagery

Chris Pearce
Stanford University
cpj.pearce@gmail.com

Abstract

This paper investigates the opportunities for applying deep learning networks to tumour classification. It finds that simple networks can be found to deliver reasonable performance, comparable with mid-range performers on the same dataset. Model saturation is a serious problem which can be resolved by a combination of limiting the number of parameters in the model, include ensuring that training data is balanced between positive and negative observations, low learning rates, and iteratively biasing the input data towards examples that the model has mis-classified after previous training epochs.

1. Introduction

When diagnosing the severity of a cancerous tumor, one core diagnostic method is for a pathologist to assign a severity score based on counts of the rate of mitoses. High rates of progression tend to be associated with worse outcomes for patients [1], and are important for clinicians in determining the intensity of a course of treatment for patients.

Historically these counts have been done using visual inspection of slides prepared from tissue biopsies. Researchers have now begun applying deep learning methods to cancer diagnostics. Advantages of this approach include reproducibility, and the ability to analyse entire slides in detail instead of focusing narrowly on specific areas of interest.

Historic work has been successful at identifying the presence of mitoses in small images with pre-selected areas of interest chosen by trained researchers such as that of Ciresan *et al.* [2]. However, recent work by has sought to generalize these results across the analysis of entire slides instead of specific pre-selected areas. Rubadue *et al.* [2] have demonstrated that deep learning algorithms can achieve human level pathologist levels of accuracy using convolution neural networks such as GoogLe Net and ResNet.

This paper tests the application of deep learning methods to the Tumor Proliferation Assessment Challenge 2016 dataset. In this challenge, participants were asked to correctly classify the location of mitoses in slide images as part of a larger challenge. This paper approaches the problem by trying to identify whether individual sub-samples of slides contain mitoses.

This paper also (unsuccessfully) tries to extend the base model to develop a general adversarial network and a fully connected network for the purposes of generating a map showing the probability that a tumour is present in any given location in a slide image.

2. Related Work

Medical imaging is a task to which various researchers have applied deep learning methods in recent years. Ciresan *et al.* [2] presented an early example of this application as far back as 2013. However, the use of handcrafted features has still been a common research approach as recently as two years ago [12][13]

Recent research has applied a variety of different model types to the problems of cancer cell detection. Arevalo et al. [9] achieved success rates of up to 82% classification accuracy with simple three layer networks applied to histopathological diagnosis. Working on a much larger data set, Paeng et al [10] managed to achieve state of the art mitosis identification accuracy rates with a ResNet based model.

Researchers have recently started to work with much larger datasets and have developed sophisticated algorithims that can be deployed in multiple environments. For example, Esteva et al. [14] have recently trained a network to detect skin cancer lesions to dermatologist level accuracy using a dataset with over 125,000 observations, with the resulting model capable of being deployed on a cellphone.

Other recent research has focussed on using images at different resolutions to identify fine and coarse details that would be indicative of the presence of a cancer cell [15].

This paper trains on a much smaller dataset that some of the most recent work however, and so relies more on techniques such as data augmentation to generate sufficient variance in the dataset, drawing on the work of Paeng et al. [10].

3. Data Sources and Preparation

3.1. Dataset

Data for the analysis has been sourced from the Tumour Proliferation Assessment Challenge 2016 datasets. The raw data set comprises slide images of cancer biopsies labelled by clinical pathologists. The base training dataset contains images taken from 73 mitosis biopsy slides, split into around 650 images. Of these, around 530 are labelled with locations for mitotic cells. The images are supplied principally as 2,000 x 2,000 pixel images in TIFF format, with labels supplied in complementary csv files.

3.2. Labelling Data

Tumours are very small relative to the size of the overall slide. One tumour may be between 30 and 60 pixels across, and an image may contain multiple mitoses, with a typical slide containing between 2 and 5 mitoses.

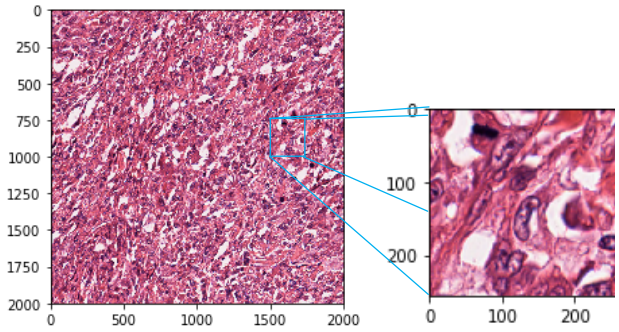


Figure 1: Slide sample, and mitosis section

To train a model to detect cells in these images, the slide is broken down into smaller tiles for training. The two main tile sizes used in this paper are 256 x 256 pixel and 64 x 64 pixel tiles. This splitting of the images can result in tiles where a tumour image breaks across multiple tiles. However, the labels only identify a single point where the tumour is located. Because the model will learn features relating to the areas of the image surrounding the specific labelled point it is important to ensure that adjacent slides are labelled as positive if a mitosis is identified close to the boundary.

To address this problem the data is translated from a single point for each mitosis to a probability map for a wider area around the point marked by the pathologist. This reflects that the mitosis occupies a place in the image larger than just the single point identified by the pathologist.

The probability map is generated based on the Euclidean distance of a pixel from the labeled point. If $\pi(x_i, y_j)$ is a point on the probability map, and X_I, Y_J is the labeled point, then

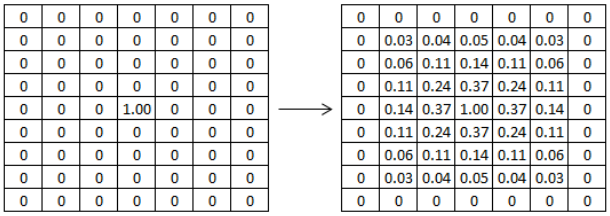


Figure 2: Translating a single point location of a mitosis to a probability map

$$d = \sqrt{(x_i\text{-}X_I)^2 + (y_j\text{-}Y_J)^2}$$
$$\pi(x_i, y_j) = e^{-a \times d}$$

When the image is split into tiles, the probability map is also split and stored alongside the base image. This allows for images to be tagged as containing a mitosis based on whether they meet a minimum probability threshold, with the threshold able to be set dynamically. The probability map also facilitates further exploration of more sophisticated analytical techniques, such as training pix2pix classifiers (Isola et al. [5]) and semantic segmentation classifiers (Shelhamer et al. [6]).

3.3. Slide Deconvolution

A major issue with performing analysis on slide image data is the variability of the image resulting from manual preparation of the slides by lab technicians. Images are stained with haematoxylin and eosin (H&E) to assist pathologists in identifying mitoses, but the manual nature of the process means that the final slides can vary significantly in appearance, leading to false classifications.

Various methods have been developed to automatically separate H&E images from the base slide. The intent is to determine a separation matrix that creates separate H&E layers wherein each pixel in the original image is contributing principally either to the haematoxylin layer or the eosin layer.

Sophisticated methods exist for calculating a separation matrix that can cleanly compute such a deconvolution a slide into H&E layers. Sparse deconvolution (Xu et al. [7]) can achieve a very accurate separation, but in trials the recursive calculation took around 10 minutes to perform on a single slide, so was not feasible for application over a large dataset. Instead, a linear principle components method proposed by Macenko et al [7] was used, which can perform a separation calculation using a single (non-recursive) calculation. Figure 3 illustrates the results of this H&E separation for a specimen slide, along with the associated probability map.

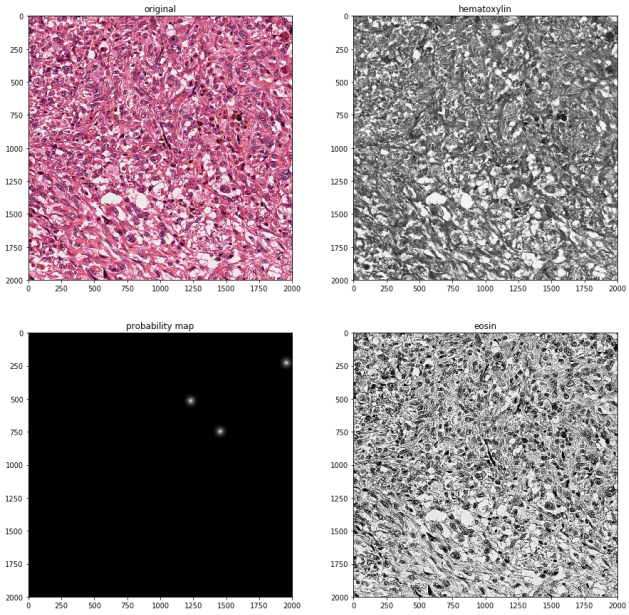


Figure 3: An original slide, and its associated probability map and hematoxylin and eosin deconvolutions

4. Methods

The primary focus of the research was to train a binary classifier to detect whether an individual image section contained a mitosis. Initial training used a simple three-layer convolutional neural network with a binary classifier in the final layer, with images broken into 256 x 256 pixel tiles.

Various internal structures were trialled with different filter sizes, use of batch normalisation and use of regularisation. The primary internal structure of each convolutional layer comprised two convolution filters with batch normalisation and ReLU activations, followed by a max pool layer. The dense layer comprised a single affine layer with dropout, followed by the classifier.

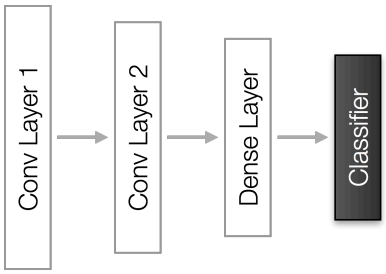


Figure 4: Basic covnet structure

Loss functions used included binary cross-entropy and SVM hinge loss. Variations on the model structure were tested including network depth, filter size and inclusion of regularisation.

Training on the full, unaugmented dataset quickly led to model saturation. A primary cause of this is likely to be the sparse presence of positive results in the dataset. A typical slide contains 2-4 mitoses, so when split into tiles, only around 5% will contain positive observations. As such, the model can achieve high accuracy during training simply by learning to classify all images as negative.

One method attempted to address this was adjusting the loss function so as to increase the penalty for incorrectly classifying a positive image. This "binary rare hinge loss" was specified as;

$$\begin{aligned}\mathcal{L} &= \max\left(0, s_j\text{-}s_{yi} + \Delta\right)\times e^{as_{yi}} \\ &= \max\left(0, s_j\text{-}(1 - s_j) + \Delta\right)\times e^{as_{yi}} \\ &= \max\left(0, 2s_j\text{-}1 + \Delta\right)\times e^{as_{yi}}\end{aligned}$$

This formula simplifies the SVM hinge loss for the binary classification setting, relying on the equivalence that $s_{yi} = (1 - s_j)$ when there are only two classification categories. It also appends the multiplier $e^{as_{yi}}$ to the loss function. For positive observations, this formulation amplifies the loss by e^{a}, whereas for negative observations the multiplier is just 1. This is intended to counterbalance the relative scarcity of positive observations in the dataset by generating a large loss if the model begins saturating and classifying all observations as negative.

A second method used to address saturation was selectively biasing the dataset being fed to the model. This approach is derived from an approach used by Paeng et al. [10], but adopts a process of updating the dataset during training rather than preselecting a specific training dataset to use in tuning the model.

In the method applied, after running the model through an initial epoch of training, the model classifier was then applied to the training dataset. All misclassified observations were then grouped into a "negative feed" dataset. This dataset was then augmented with a random sample of correctly classified observations, and distorted using random rotations and applying Gaussian noise.

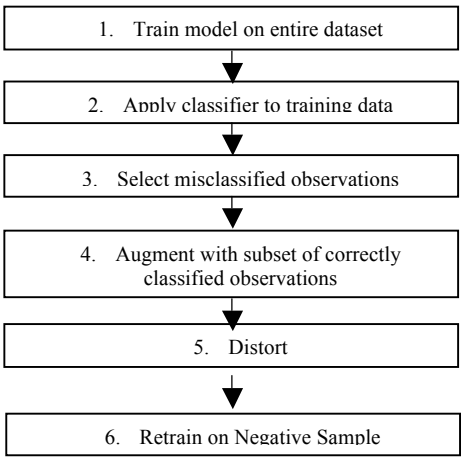


Figure 5: Negative Feed dataset generation

The learning rate was also varied during testing to determine an optimal rate that would not cause premature saturation.

The primary metric used to measure success of the model was simple classification accuracy i.e. is a tile correctly classified as containing a mitosis. This measure has been used by researchers as a base for measuring model accuracy (Arevalo et al. 2015 [9]), and gives a base measure of the functionality of the model.

A second metric used in the literature is the F1 score for binary classification (Paeng et al. 2016 [10]). The F1 score is calculated as;

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

where precision is the number of correct positive predictions divided by the number of predications made, and recall is the number of correct positive predictions made divided by the true number of positive observations in the dataset (Van Rijsbergen, 1979 [11]).

5. Results

5.1. Initial Model Testing

Initial training was performed on the entire training dataset using the basic covnet structure outlined above, along with Inception and Resnet models. Standard SVM hinge and binary cross-entropy losses were used for the classifier layer, and learning rates were set in the range of 1e-2 to 1e-4. The dataset was split into 256 x 256 pixel tiles and all observations were initially used. Models were trained on a combination of H&E separated inputs, and H&E inputs combined with the original RGB slide image (i.e. six layers in total). The Adam optimiser was used to control gradient descent, with its default settings of $\beta 1 = 0.9$ and $\beta 2 = 0.999$ and learning rate decay set to 1e-6.

Initial training confirmed the problem of model saturation. After a single epoch all configurations of the model were converging to classify all images as negative.

As a first response, the "binary rare hinge loss" method was introduced into the model, with the parameter "a" set at values of 2 and 4. This method proved able to slow the rate of saturation to two epochs instead of one, but it could not prevent saturation from occurring entirely.

As a second response, the training dataset was pared back to contain around 2/3 negative images and 1/3 positive to force the model to learn more features relevant to positive classification. The learning rate was also reduced to 1e-5 to 1e-6 reduce the extent of overfitting. These methods again proved successful at slowing the rate of model saturation, but not stopping it altogether.

The final method that was introduced was the "negative feed" method described above. This approach finally resulted in the "basic covnet" design model beginning to generate classifications across both categories. ResNet and Inception architectures still suffered from saturation problems. As such, it was decided to proceed with the basic covnet structure in conjunction with the negative data feed method for more detailed fine tuning.

5.2. Model Fine Tuning

Various model structures were then trialled, varying different features of the model including;

- Training on 256x256 images and on 64x64 images
- Varying filter sizes in the first and second conv layers, including (3x3)(3x3), (8x8)(8x8) and (8x8)(8x4)
- Varying the number of neurons in the affine layer to 512, 1024 and 2056
- Varying the number of filters as 32, 64 and 128

The binary rare hinge loss was tested, but ultimately performed no better than binary cross-entropy or SVM hinge loss. Given the risk that with a more balanced dataset this method might start distorting the data, binary cross-entropy was selected as the preferred method for the model.

The table below illustrates the results that were generated from the model. By far the best performing model is the first in the list, where validation accuracy of 78% was achieved. This performance is approaching that of Arevaloa et al. [9], who reported model performance of 82% with a three layer network on 150 x 150 pixel images.

Image Size	Filter Size	Affine Size	Filter Number	Validation Accuracy
256x256	(3x3)(3x3)	512	128,64	78%
64x64	(3x3)(3x3)	512	128,64	61%
256x256	(3x3)(3x3)	1024	128,64	55%
256x256	(3x3)(3x3)	2056	128,64	54%
64x64	(8x8)(4x4)	512	128,64	58%
64x64	(8x8)(4x4)	512	64,32	56%

However, other model specifications performed significantly worse than this. A general trend in the data is that models seem to perform worse as the ratio of parameters increases relative to the image size. The small size of the overall dataset may be a cause of this problem, as a dataset with a great many parameters may begin to suffer from problems with co-linearities and a lack of degrees of freedom to fit the model.

5.3. Results and Comparisons

The model generalises reasonably well to the training data set, scoring 72% accuracy. **Figure** 6 shows the confusion matrix for the test dataset. The model shows good accuracy in predicting true negatives, but the true positive prediction rate is not good. The overall F1 score for the model is 0.42, placing it slightly below the middle of the performance table for participants in the 2016 Tumour Proliferation Challenge.

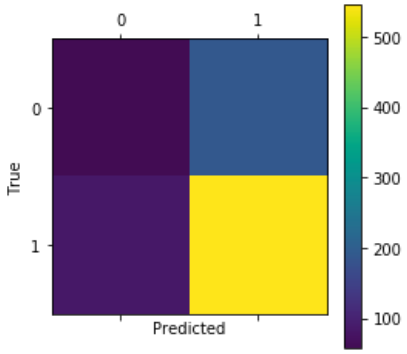
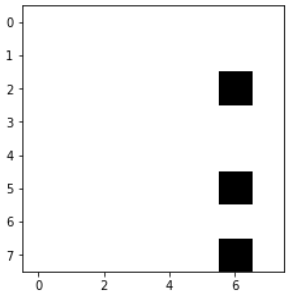


Figure 6: Confusion matrix for test data (1 = negative classification)

When scaling up the overall predictions to an entire slide, the model does make reasonable predictions as to the location of tumours however. This is especially the case given that it is trained in an environment where positive and negative images are roughly balanced in terms of their prevalence, whereas in the test environment negative images are significantly more prevalent.

To visualise this,

Figure **7** shows a predicted probability map versus the ground truth for an observation from the test dataset (in this representation, black squares represent a positive prediction by the model). Two of the three predictions are in the correct location, but it misses one prediction and generates a false positive for a second one.



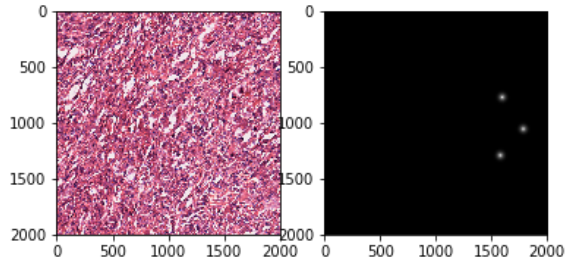


Figure 7: Probability Map versus Ground Truth Data – Test dataset

In the second example from the validation dataset, the model correctly identifies that one of the mitoses splits across two tiles in the dataset, and identifies it as being present in both the images (see the two adjacent black squares). However, it again misses one mitosis, and has two false positive predictions.

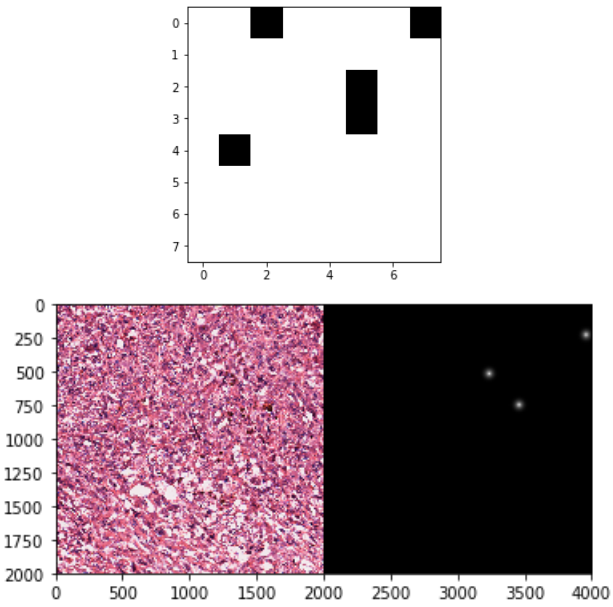


Figure 8: Probability Map versus Ground Truth Data – Mitosis split across two tiles

6. Conclusions

Cell slide data presents a difficult classification task. The sparsity of positive data creates significant challenges in creating the right training environment for the model. This can be overcome by using techniques including training with low learning rates, using negative feed data generation and potentially using models with smaller numbers of parameters.

The simple convolutional network trained in this paper performs well on predicting true negative observations, even give the fact that it is trained in an environment

where negative observations are relatively scarce.

7. Appendix 1 – Further Work

Outside of the base model, two methods were attempted for generating a probability map directly from the input image. One model structure was a fully connected network, attempting to make pixel level classifications, inspired by the work of Shellhamer et al. [6];

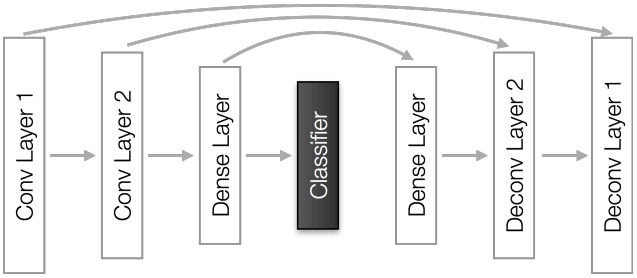


Figure 9: Fully Connected Network Schematic

In this model, each of the convolution layers and the dense layer are connected to a deconvolution layer that ultimately scales up to an image the same size as the original. This model was trained using the preprocessed slide image as an input and the probability map as an target with a categorical cross-entropy loss function without regularisation. End to end training was attempted, along with training the classifier network first, then locking the classifier weights and training the deconvolution network.

This model was adapted from the Keras-FCN Github repository (github.com/jihongju/keras-fcn). However, the model was shortened to contain only two convolutional layers instead of the original model's seven layer structure. In part, the intent of making this change was because of the small size of the training dataset, and concerns that a significantly larger model would create problems with over-parameterisation leading to further problems with model saturation.

In a second model attempted, a GAN structure taken from the affinelayer.com implementation of the pix2pix network was applied to the H&E slide data as an input, and to the probability map as an output [16]. The model contains a generator and two discriminator networks (one to classify fooling images and one to classify real images).

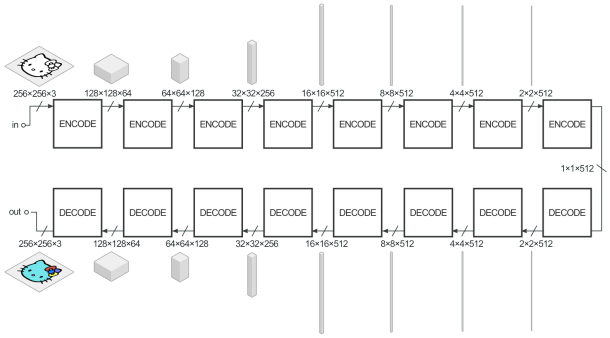


Figure 10: pix2pix Model Network (reproduced from affinelayer.com)

Neither method successfully generated probability maps. Possibly due to the complexity of the model and the sparsity of the data, the pix2pix network learned to generate some small amounts of white data in a circle, but failed to correctly locate the data in the correct section of the probability map. The FCN model simply failed to train. Given the difficulties getting a simple convolutional network to train on this dataset, it is possibly not surprising that this result occurred. These approaches still seem like they could hold some theoretical promise however, and could be a fruitful area for further exploration.

8. Appendix 2 – Key Repositories Used

HistomicsTK
LargeSlide
Keras
TensorFlow
Keras-FCN

Acknowledgements

I would like to thank the staff and tutors of the course for their excellent work in compiling an excellent class. It has been thoroughly enjoyable. I would especially like to thank Shayne Longpre for providing feedback on the work during its preparatory satges.

References

[1] M. Veta, P. J. van Diest, M. Jiwa, S. Al-Janabi, and J. P. Pluim, "Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method" PloS one, vol. 11, no. 8, p.e0161286, 2016.

[2] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, 2013, pp. 411–418

[3] C. Rubadue, D. Suster, and D. Wang. “Deep Learning Assessment of Tumor Proliferation in Breast Cancer Histological Images”, arXiv:1610.03467, https://arxiv.org/pdf/1610.03467.pdf
[4] A Radford, R Jozefowicz and I Sutskever. Learning to Generate Reviews and Discovering Sentiment, arXiv:1704.01444, https://arxiv.org/pdf/1704.01444.pdf
[5] P. Isola, J. Zhu, T. Zhou and A. Efros. “Image-to-Image Translation with Conditional Adversarial Networks”, arXiv:1611.07004v1, 2016.
[6] E. Shelhamer, J. Long and T. Darrell. “Fully Convolutional Networks for Semantic Segmentation”, arXiv:1605.06211v1, 2016
[7] J. Xu, L. Xiang, G. Wang, S. Ganesan, M. Feldman, N.N. Shih, H. Gilmore, A. Madabhushi, "Sparse Non-negative Matrix Factorization (SNMF) based color unmixing for breast histopathological image analysis" IEEE Computer Graphics and Applications, vol.46,no.1,pp.20-9, 2015
[8] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, G. Xiaojun, C. Schmitt, and N. E. Thomas. “A method for normalizing histology slides for quantitative analysis” IEEE ISBI, 2009. dx.doi.org/10.1109/ISBI.2009.5193250
[9] J. Arevaloa, F. A. Gonzáleza, R. Ramos-Pollánb, J. L. Oliveirac and M. A. Guevara Lopez. “Representation learning for mammography mass lesion classification with convolutional neural networks” Computer Methods and Programs in Biomedicine, vol. 127 (2016) pp. 248–257
[10] K. Paeng, S. Hwang1, S. Park, M. Kim and S. Kim. “A Unified Framework for Tumor Proliferation Score Prediction in Breast Histopathology” arXiv:1612.07180v1, 2016
[11] Van Rijsbergen, C. J. “Information Retrieval (2nd ed.)” Butterworth, 1979
[12] X. Liu, J. Tang, “Mass classification in mammograms using selected geometry and texture features, and a newSVM-based feature selection method, Syst. J. IEEE 8 (3) (2014) 910–920, http://dx.doi.org/10.1109/JSYST.2013.2286539.
[13] M. Dong, X. Lu, Y. Ma, Y. Guo, Y. Ma, K. Wang, “An efficient approach for automated mass segmentation and classification in mammograms”, J. Digit. Imaging 28 (5) (2015)
[14] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun. “Dermatologist-level classification of skin cancer with deep neural networks”, Nature 542, 115–118 (Feb 2017)
[15] K. J. Gerasa, S. Wolfsonc, S. G. Kimc, L. Moyc, and K. Cho. “High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks”, arXiv:1703.07047v1, 2017
[16] C. Hesse. “Tensorflow port of Image-to-Image Translation with Conditional Adversarial Nets” https://github.com/affinelayer/pix2pix-tensorflow