Segmentation Guided Regression Network for Breast Cancer Cellularity

Yixuan Wang[1] *, Li Yu[1], and Shengwei Wang[1]

School of Electronic Information and Communications, Huazhong University of Science and Technology (HUST), Wuhan, China
{yixuanwang,hustlyu,kadinwang}@hust.edu.cn

Abstract. Evaluation and diagnosis of breast cancer will be more and more vital in medical field. A general solution to breast cancer cellularity is to modify output of a state-of-the-art classification backbone to prediction a score between 0 and 1. However, this solution does not take clinical meaning of cancer cellularity which defined as proportion of cancer cells over image patches into consideration. In this paper, a segmentation guided regression network is proposed for breast cancer cellularity, adding more semantic detailed features for regression task. Consequently, the proposed method can not only take advantage of global context features from classification backbone, but also position feature and texture feature from segmentation network. A powerful segmentation network with 0.8438 mean Intersection-over-Union is obtained on extremely class imbalanced datasets. The proposed method with Resnet101 as regression backbone gets PK value of 0.9260 and L1 loss of 0.0719.

Keywords: Cancer Cellularity · Image Segmentation · Non-linear Regression.

1 Introduction

Breast cancer has become a general and vital problem for women around the world [1, 2]. To carry out medical diagnosis and inspection on breast cancer, clinician usually utilize tissue slice as golden ground truth to judge cancer state and what treatment can be executed on the patients and the cancer cellularity is defined as percentage area of cancer cell within a tissue section [3]. Specifically, medical doctors obtain whole slide images (WSI) of breast histological sections by the microscope firstly. Then hematoxylin and eosin are stained on pathological slides. Finally, a caner cellularity or score is given out according to observation on growing state and complicated cell structures of cancer cell within each image patches cropped from a tissue section [4]. The cancer cellularity can be seen as a score between 0 and 1 to describe how severe the tumor condition is. To help patients get diagnosis results in time, automated methods for breast

* Yixuan WangCurrently working toward the Master degree in the School of Electric Information and Communications, HuaZhong University of Science and Technology.

cancer cellularity are in great demand.

Breast cancer cellularity challenge can be formulated as a non-linear regression problem: using stained image patches cropped from WSI, a cancer score ranging from 0 to 1 is predicted to represent severity of breast cancer. Previous methods utilize traditional segmentation method to get categories of each pixel and machine learning method for regression [5]. Recently, adopting a classification network and modify the last layer of fully connected layer from 1000 categories to a single output together with sigmoid activation function might be a basic solution. Nonetheless, those solution are not end-to-end or do not take the realistic physical significance of cancer cellularity into account. And the features extracted by classification backbone lose high level detailed semantic information. CNNs features are translation invariant and scale invariant to get robust performance in various image instances within a specific category in image classification task, thus features contain global abstract context information and detailed structure or position spatial information might be lost. However, for image segmentation task, appearance and spatial geographic information will be preserved to output a high-resolution map with precise features on each pixel. A segmentation guided regression network is proposed to fuse segmentation features and classification features for regression in this paper.

The designed network is composed of segmentation part which derives from DeepLabv3+ model and regression part which stems from typical classification models like Resnet101. The devised network is trained in two steps sequentially: segmentation network and the whole network for regression. In segmentation task, class imbalance problem is figured out by weighted cross entropy loss and focal loss [15] techniques. The designed model utilizes encoder weights of DeepLabv3+ as initial parameters of segmentation module. And the whole network is updated with L1 loss for regression. The proposed cancer cellularity model is compared with multiple state-of-the-art classification backbones such as Resnet101, Resnet152[19], Resnext[20], SENet[21] and NASNet[22] on regression task. While all typical backbone with stride of convolution stage1 changed to 2 for computationally efficiency. Experimental results show that the proposed network based on Resnet101 performs better than merely single regression networks with typical CNNs as backbones.

2 Related Works

Deep convolutional neural networks (CNNs) have made great progress in medical image processing [6, 7] recently. CNNs are composed of a series of basic components such as two-dimension convolution operation, batch normalization layers, activation function and spatial pooling layers [12]. CNNs have been widely utilized in various vision task such as image classification, image segmentation, object detection and many other high-level tasks in computer vision due to its strong ability for feature extraction.

2.1 CNNs in Semantic Segmentation

State-of-the-art image segmentation methods based on CNNs [8–11] work well on both natural image and medical image. Typical segmentation networks are composed of two modules: encoder module and decoder module. Encoder module performs convolutional operations and max pooling layers for down sampling to obtain a low resolution, rich semantic feature map. Several spatial pooling layers in encoder module are to increase the receptive field of convolutional kernels to get plentiful semantic information. Decoder module adopts the output feature maps of encoder module as input, and executes up-sampling to gain a high-resolution output with abundant categories information for semantic segmentation. General methods of up-sampling are deconvolution operation and bilinear interpolation. Previous improvements on segmentation networks always concentrate on decoder module. To be specific, UNet brings in skip connection in the way like FPN [13] to get high level, high resolution feature maps for better performance. SegNet [14] reuses indices in the max pooling step of encoder module at corresponding decoding layers for more precise up sampling. DeepLabv3+ improves encoder parts of segmentation architecture. DeepLabv3+ focuses on enrichment of encoder module and introduces atrous spatial pyramid pooling layers to extract robust semantic features. DeepLabv3+ model also shows that the design of encoder module is much more significant than decoder module, thus it is believable that the encoder module of segmentation network is responsible for extracting rich semantic features and the decoder part is just used for up sampling. The output stride of segmentation network refers to the ratio of the resolution of input images to feature maps output by the encoder module. The output stride usually equals 16 in segmentation networks.

2.2 CNNs in Image Classification

The fundamental tasks in computer vision are image classification, object detection and semantic segmentation. High level complex tasks such as scene text recognition, automatic driving and image captioning and so on. The most basic task is image classification and all other task adopts CNN parts of image classification model as base network or backbone. Deep learning method outperforms traditional image processing method in computer vision task mostly due to the strong feature extraction ability of convolutional networks. CNNs frameworks in image classification are composed of two parts: CNNs part for extracting convolutional features and fully connected layers used as classifier. CNNs consist of a sequence of two-dimension convolution operations, batch normalization layers, activation functions and max pooling layers. Fully connected layers are made up of linear connected and activation functions.

The main development of classification model focus on CNNs part. From VGGNet [16] to GoogLenet [17], the development in image classification indicates that the increasing depth and width of convolutional layers within CNNs can improve the feature representation ability of CNNs for attaining global context features. By normalizing the input feature maps for each convolutional stage,

batch normalization operation is presented to regulating each feature map generated by corresponding convolutional kernel within a batch size, that is, normalizing output values along batch size and spatial dimension in CNNs. Batch normalization in fully connected layers operates in batch size and node dimension. The loss function can get convergence stably and the internal covariate shift is solved by adding batch normalization. However, increasing the depth of CNNs without limitation can lead gradient vanishing problem and gradient exploding problem. Resnet shows that the CNNs can be designed deeper by adding skip connections. Residual blocks within Resnet performs element wise summation between input features maps and output convolutional feature maps. The gradient from upstream is propagated to the input feature maps if the convolutional stage of current residual block is unimportant, thus avoiding gradient accumulation. Attention mechanism is applied in the design of CNNs backbone to guide CNNs to concentrate more on learning meaningful features related to specific task. SENet introduces channel-wise attention information by utilizing squeeze-and-excitation module. Non local neural network [18] performs attention operation on convolutional feature map directly on both channel dimension and spatial dimension.

3 Segmentation Guided Regression Network

The overview of the proposed network is presented in Fig. 1. It is made up of two submodules: segmentation module and regression module. The segmentation module is derived from the encoder part of DeepLabv3+ which can obtain high level and semantic feature map with high resolution. The regression module can be any classification backbones and can acquire global context features. Feature maps from two submodules are fused to be one feature map as input to prediction layers. The fused feature maps which contain rich semantic features and texture information on each pixel can offer corresponding message about cancer cell areas on input images, which makes output results of prediction layers more reasonable and trustworthy. Cancer cells always have significant difference from normal cells in textures and edges, so it is extremely helpful if high-level semantic features or appearance details could be taken into account when scoring input images. Consequently, the prediction procedure of the proposed network is more similar to the realistic clinical meaning of cancer cellularity, that is, the percentage areas of cancer cell within an image patch.

3.1 The Architecture of Segmentation Guided Regression Network

Segmentation Module Segmentation dataset is adopted to train a DeepLabv3+ network which is pretrained on ImageNet classification task with the guidance of transfer learning. The output stride in encoder part of segmentation network equals 16 to get $32\times32\times256$ feature maps for $512\times512\times3$ input images. Output stride is formulated as the ratio of input image resolution to feature map in the last convolutional layer. The segmentation module of the proposed method is identical to encoder part of DeepLabv3+.

Regression Module Regression network can be arbitrary classification backbones. Given the fact that general classification networks need input of 224×224×3 and the standard resolution of image patches in classification datasets is 512×512×3, the stride in first convolutional stage is changed from 1 to 2 in all classification backbones for computational efficiency.

Segmentation Guided Regression Network A DeepLabv3+ model is trained on segmentation dataset at first. Then the encoder part of DeepLabv3+ segmentation network and regression module in the proposed method are trained simultaneously, while weights in first step are applied as initial parameters for segmentation network encoder part and regression network adopts kaiming normal initialization. The output of DeepLabv3+ encoder is concatenated with the outputs of convolution stage2 of regression network in channel dimension, and both output stride equal 16. To obtain better merged feature map, 3×3 and 1×1 convolution layers are applied after concatenation of two feature maps from segmentation module and regression module for smoothing. Global average pooling is operated on 8×8 feature maps with output stride 64. Finally, fully connected layers and sigmoid activation function are used for score regression.

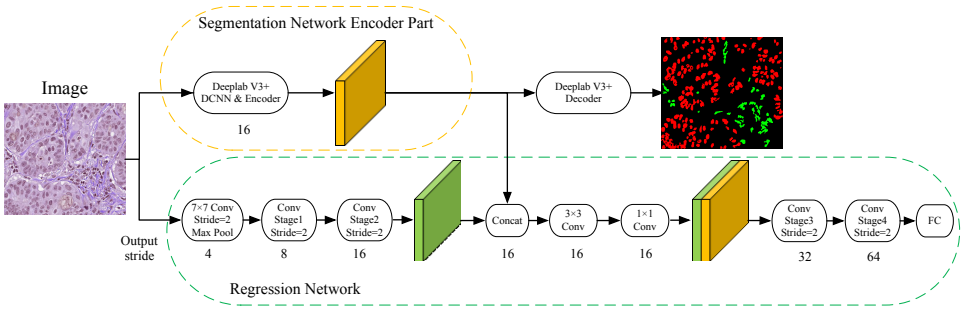

Fig. 1. Network architecture of the proposed segmentation guided regression network for breast cancer cellularity.

3.2 Training

The dataset adopted in this paper is from 2019 MICCAI grand challenge on breast cancer cellularity at http://spiechallenges.cloudapp.net/competitions/14. In this challenge, for regression, there are 2394 training images and 185 validation images with resolution of 512×512 in RGB color space, and regression scores are given for both two batches of images. For additional segmentation dataset, only cell nucleus points coordinate ground truth are given for 154 images with various spatial resolutions. That is, the pixel points that are cell nucleus points are marked as a specific category and other pixels are viewed as background points.

Segmentation Ground Truth Generation Given that only cell nucleus points coordinate ground truth are given for segmentation task, to get cancer cell areas in image patches, ground truth mask should contain information about not only cancer nucleus but also a completed cell area. Therefore, ground truth masks about completed cells are generated with some prior knowledge and traditional algorithms in image processing. Specifically, watershed algorithm is adopted to acquire cell edges and contours, then fill each contour with categories-specific color labels using points coordinate given by experts as center points. Initial image, contours given by watershed algorithm and the generated cell segmentation ground truth are demonstrated in Fig. 2 from left to right, while black/ red/ blue/ green indicates background/ malignant cells/ normal cells/ lymphocyte cells respectively. While center points given by experts are shown as small circles centered on that pixel points for better visualization in the middle column.

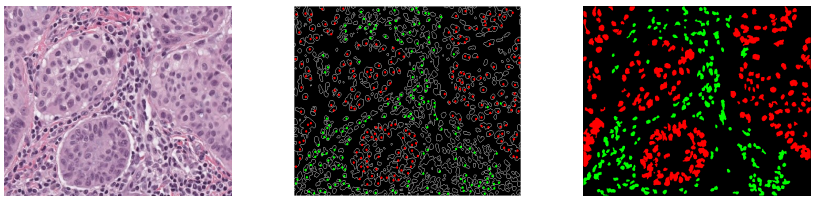

Fig. 2. Generation of segmentation ground truth via a set of labeled pixel points. Initial image is on the left. Contours and edges are produced by watershed algorithm via a list of labels points indicated cell nucleus of each categories. Each contour is filled with center categories by a threshold to obtain segmentation ground truth.

Class Imbalanced in Segmentation Segmentation dataset is grouped into two parts randomly: 146 training images and 8 images for validation. DeepLabv3+ model with output stride 16 is adopted. In training sets, the number of four categories pixel points are: 71338338, 819834, 1848719, 11644755 which indicates background, lymphocyte cells, normal cells, malignant cells respectively. As a result, severe class imbalance problems should be faced with. Weighted cross entropy and focal loss are adopted to figure out this problem. Small weights are distributed to the categories that have large frequency and a log smooth item is employed to balance weights magnitude. However, experiments show that weighted loss is not too effective to perform well on lymphocyte cells for the reason that lymphocyte category is extremely less in training set. Consequently, focal loss technique which obtain perfect performance in object detection filed is added to force training optimizer pay more attention to hard pixel points.

$$w_i = \frac{1}{\log\left(1.10 + f_i\right)} \tag{1}$$

where f_i is the frequency of specific category cell pixels. Eq.(1) shows the method of calculating weights of each categories pixels within a batch size.

Regression Task The encoder weights of well-trained DeepLabv3+ network are initial parameters value of segmentation module in the proposed network. When training the whole network for regression task, the regression module is initialized with kaiming normal. The whole network with Resnet101 as CNNs backbone is updated using L1 loss.

4 Experimental Results

4.1 Training Strategy

Segmentation Training When training DeepLabv3+, various augmentation strategies are exploited, such as: scale with ratios from 0.8 to 1.2 randomly, crop a fixed size 256×256 randomly, flip vertically and horizontally randomly. Through contrast experiments, it can be found that it is better to use pretrained parameters on ImageNet classification task as initial weights of the feature extraction networks of segmentation network. SGD optimizer with an initial learning rate 0.01 and batch size 50 is used and the learning rate is decayed every 3000 epochs for total 7000 epochs.

The Whole Network Training For data augmentation: First rotate images randomly with 0, 90, 180, 270 degrees. Then randomly flip images horizontally and vertically. Finally, shift brightness, contrast, saturation and hue randomly. Adam optimizer is used with start learning rate 0.003 and decay rate 0.1 every 300 epochs for total 900 epochs with batch size 64 on the regression dataset.

4.2 Metrics and Results

Segmentation Results Table 1 and Table 2 shows segmentation MIOU on segmentation validation dataset and training dataset with various γ value in focal loss equation respectively. For the reason that 8 validation images are randomly chosen from segmentation dataset and the number of Lymphocyte pixels is extremely less, so segmentation result on training set is given as a reference. Those process aims to acquire best training strategy on segmentation task, consequently, all data is utilized to obtain an overfitting model to learn high-level complicate semantic features of different kinds of cells. MIOU indicates mean Intersection-over-Union of all categories. PA means pixel accuracy on a specific category and it is formulated as the number of true predictions of that category divides all pixels of that category.

It is obvious that γ equals 1 might be better for class imbalance problem. Table 3, Table 4 and Table 5 demonstrate the results of confusion matrix on validation dataset on different γ values respectively. Confusion matrix is a typical

metric in classification problems. The summation of each row equals the number of ground truth samples of specific category and elements of main diagonal are true positive of each category. Take Table 3 for example: the second row, the first column indicates that 225 pixels which are lymphocyte are predicted as background. It is evident that normal pixels and lymphocyte pixels can be distinguished clearly for the reason that there are no normal pixels being predicted as lymphocyte wrongly and vise versa. The mispredicted pixels for foreground categories are always being predicted as background. It might be that the number of background pixels is extremely large and the model tends to predict any pixels that it cannot determinate correct categories as background to minimize the risk.

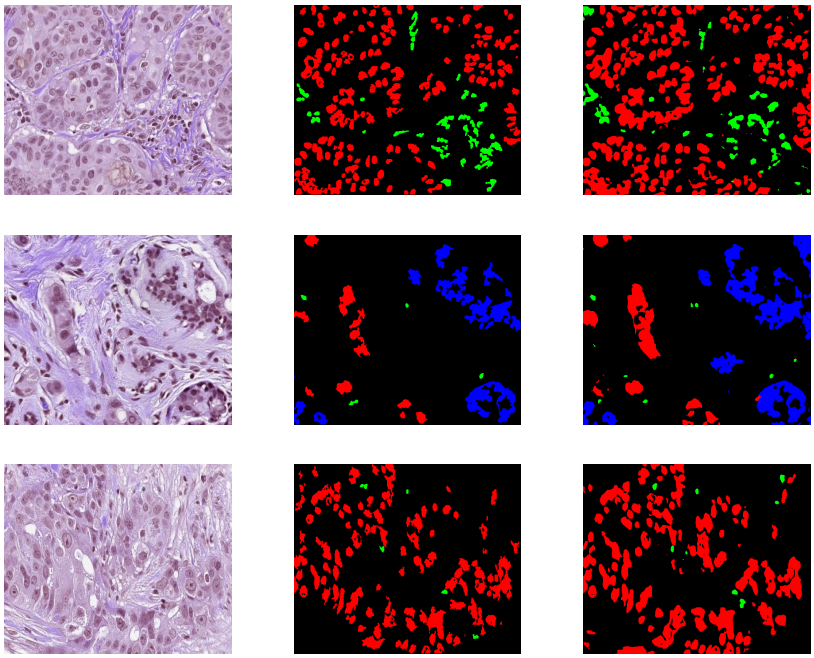

Fig. 3. Segmentation result with γ equals 1 in focal loss. Initial images are on the first column. Ground truths generated by watershed algorithm are on the second column and predictions by segmentation network are on the last column.

Regression Results Resnet101, Resnet152, Resnext, SEnet, NASNet and the proposed regression network based on Resnet101 are all trained with L1 loss for non-linear regression. PK value for predicting probability is formulated in [23]. The experiment results express that the proposed segmentation guided regression network gets more excellent performance than any other regression network

Table 1. Segmentation results of different γ in focal loss on validation set.

γ	MIOU	Background PA	Lymphocyte PA	Normal PA	Malignant PA
1	0.6449	0.9467	0.6432	0.9257	0.8608
2	0.6461	0.9421	0.5298	0.9430	0.8824
5	0.6233	0.9344	0.6009	0.9047	0.8902

Table 2. Segmentation results of different γ in focal loss on training set.

γ	MIOU	Background PA	Lymphocyte PA	Normal PA	Malignant PA
1	0.8438	0.9625	0.9663	0.9725	0.9681
2	0.8254	0.9588	0.9631	0.9743	0.9674
5	0.6310	0.9017	0.8044	0.9322	0.8449

Table 3. Confusion matrix of γ=1 in focal loss on validation set.

pred / gt	Background	Lymphocyte	Normal	Malignant
Background	1840327	7883	17059	58244
Lymphocyte	226	10575	0	2
Normal	1632	0	51512	0
Malignant	8483	51	0	258342

Table 4. Confusion matrix of γ=2 in focal loss on validation set.

pred / gt	Background	Lymphocyte	Normal	Malignant
Background	1812079	4916	26636	79882
Lymphocyte	4828	5723	0	252
Normal	1890	0	50114	1140
Malignant	31020	356	0	235500

Table 5. Confusion matrix of γ=5 in focal loss on validation set.

pred / gt	Background	Lymphocyte	Normal	Malignant
Background	1797263	7707	30056	88487
Lymphocyte	4278	6492	0	33
Normal	3304	0	48079	1761
Malignant	28313	1000	0	237563

based on typical classification networks either on PK value or L1 loss. The experimental results also indicate that all state-of-the-art classification backbone get very close performance with regression module merely. The classification backbone can extract global abstract features without detailed texture features. Consequently, the regression layers cannot distinguish which category each pixels belong to and cannot get area of cancer cell in image patches which generate accurate prediction score for cancer cellularity. The proposed segmentation guided regression network utilizes both semantic features from segmentation module and global context features from regression module to obtain rich features for better regression performance.

Table 6. Evaluation of different network on regression validation dataset.

Model	Resnet101	Resnet152	Resnext	SEnet	NASnet	The Proposed
PK	0.9229	0.9206	0.9228	0.9232	0.9237	0.9260
L1 distance	0.0779	0.07231	0.0778	0.0735	0.0748	0.0719

5 Conclusion

In this paper, a segmentation guided regression network was proposed to fuse segmentation feature and classification feature to extract more detailed abstract semantic features for regression task. Segmentation features are usually with high resolution and well-preserved geometric information to get better understanding of local pixels, while classification features contain global context. The final prediction score is based on the fused feature map which contains high level semantic features and global context information. The main idea behind the proposed network is to combine segmentation features and classification features to gain higher performance on scoring task.

The segmentation guided regression network is trained by two steps sequentially. First, a completed DeepLabv3+ model is trained with multi-categories cross entropy and focal loss on segmentation dataset with backbone weights initialized on ImageNet classification task. Next, the whole network is trained with L1 loss for regression. And the segmentation module of network is initialized by encoder part of DeepLabv3+ in the first step. Resnet101 is used as base network in network with kaiming normal initialization. Experimental results indicate that the proposed segmentation guided regression network performs better than using a regression network merely.

References

1. Siegel R L, Miller K D, Jemal A: Cancer statistics. CA: a cancer journal for clinicians, 65(1), 5-29 (2015)

2. Symmans, W. Fraser, et al: Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. Journal of Clinical Oncology 25.28, 4414-4422 (2007)
3. Thompson, A. M., and S. L. Moulder-Thompson: Neoadjuvant treatment of breast cancer. Annals of Oncology 23, x231-x236 (2012)
4. Hermanek, Paul, and Christian Wittekind: Residual tumor (R) classification and prognosis. Seminars in surgical oncology (1994)
5. Peikari M, Salama S, NofechMozes S, et al: Automatic cellularity assessment from posttreated breast surgical specimens. Cytometry Part A, 91(11), 1078-1087 (2017)
6. Z. Wang, C. Liu, D. Cheng: Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network. IEEE Transactions on Medical Imaging (2018)
7. Sachin Mehta, Ezgi Mercan, Jamen Bartlett: Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. CoRR, abs/1806.01313 (2018)
8. Jonathan Long, Evan Shelhamer, Trevor Darrell: Fully Convolutional Networks for Semantic Segmentation. The IEEE Conference on Computer Vision and Pattern Recognition (2015)
9. Olaf Ronneberger, Philipp Fischer, Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation. CoRR abs/1505.04597 (2015).
10. Chen, Liang-Chieh, Zhu: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611, (2018)
11. Liang-Chieh Chen, George Papandreou, Florian Schroff: Rethinking Atrous Convolution for Semantic Image Segmentation. CoRR, abs/1706.0558 (2017)
12. Sergey Ioffe, Christian Szegedy: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. CoRR, abs/1502.03167 (2015)
13. Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
14. Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39, no. 12, 2481-2495 (2017)
15. Lin, Tsung-Yi, Goyal: Focal loss for dense object detection. IEEE transactions on pattern analysis and machine intelligence. (2018)
16. Simonyan, Karen, and Andrew Zisserman: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Szegedy, Christian, et al: Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. (2015)
18. Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He: Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794-7803 (2018)
19. He, Kaiming, Zhang, Xiangyu: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
20. Xie, Saining and Girshick, Ross and Dollár, Piotr and Tu, Zhuowen and He, Kaiming: Aggregated residual transformations for deep neural networks. Computer Vision and Pattern Recognition, (2017)
21. Hu Jie, Shen Li: Squeeze-and-excitation networks. arXiv prepr: 1709.01507(2017)
22. Zoph, Barret and Vasudevan, Vijay and Shlens, Jonathon: Learning transferable architectures for scalable image recognition. arXiv preprint arXiv:1707.07012. (2017)
23. Smith W D, Dutton R C, Smith N T: A measure of assosication for assessing prediction accuracy that is a generalization of nonparameteric ROC area. Statistics in Medicine,15(11):1199 (1996)