WEAKLY SUPERVISED FLUID FILLED REGION LOCALIZATION IN RETINAL OCT SCANS

ShahRukh Athar [⋆] *Abhishek Vahadane*[†] *Ameya Joshi*[†] *Tathagato Rai Dastidar*[†]

[⋆]Skolkovo Institute of Science and Technology, Moscow, Russia
[†]SigTuple Technologies, Bengaluru, India

ABSTRACT

Retinal Optical Coherence Tomography (OCT) scans are an important diagnostic tool for ophthalmologists. These scans provide a cross-sectional view of the retina for ophthalmologists to detect abnormalities. A common type of abnormality found in these scans is a Fluid Filled Region (FFR). In this paper, we present a method to simultaneously classify and localize FFRs within retinal OCT scans using a specialized Convolutional Neural Network (CNN). The training data is weakly labeled, with only an indication of whether a scan contains FFRs or not. We compare different architectures to see which ones give us the best localization and classification metrics. We have found that architectures using Dense Blocks and Scaled Exponential Unit (SeLU) activations give us the best localizations with a Mean Average Precision (mAP) of 0.75 on true positive images and a classification accuracy of 94.8%.

***Index Terms*—** Optical Coherence Tomography, Eye, Computer Aided Detection and Diagnosis

1. INTRODUCTION

Optical Coherence Tomography (OCT) is an important advancement in the field of retinal ophthalmology. These scans give a microscopic, cross-sectional view of the retina thus giving the doctor a visualization of major retinal layers and possible abnormalities within them. A common type of abnormality found is a Fluid Filled Region (FFR). It consists of a localized expansion of the retinal extracellular space associated with the intracellular space in the macular area [11]. Cysts and subretinal fluid (SRF) are two instances of FFR. In this paper, we refer to a scan with FFRs as an abnormal scan. Our aim is to partially automate the process of visual inspection of scans by training a CNN to not only distinguish between normal and abnormal scans, but also to localize the FFRs in any scan the network predicts as abnormal. Manual localization of FFRs in an OCT scan is a tedious, subjective and error prone process. We solve the challenging problem of FFR localization with weakly supervised techniques only, using OCT frames labeled as abnormal or normal. Our work is motivated by the recent work [12] which showed that a network localizes discriminative regions by using a Global Average Pooling (GAP) [9] layer right before the softmax layer. In this paper, we propose a specialized CNN architecture which consists of convolutional layers with learnable downsampling, followed by dense blocks [4] for simultaneous classification and localization. We also compare the performance of different CNN architectures using a GAP with respect to classification and localization. We conclude that networks using the SeLU activations and Dense Blocks [4] give the best localizations and classification accuracies.

The paper is organized as follows: In Section 2, we elaborate on relevant related work. Section 3 discusses the technical details of the architectures. Experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2. RELATED WORK

Significant research has happened in recent years on generating weak localizations using CNNs. [12] shows that the GAP layer encourages the network to learn discriminative regions within an image. This was extended by [3] which generated weak segmentation masks of pulmonary nodules on lung computed tomography (CT) scans.

Use of CNNs for analysis of retinal OCT scans has been studied. [10] uses a 2D patch-based CNN to classify OCT scan patches as containing one of a set of abnormalities, or being normal. It uses a per-voxel segmentation label as input. [1] presents a method to classify Retinal OCT volumes as either normal or having Age-Related Macular Degeneration (AMD) using a two stage training process. It first trains a CNN to classify frames of the OCT volume using the target label of the 3D volume to which they belong. Next, these pretrained weights are used to initialize the lower layers of a much deeper network and predicts the label of an entire 3D volume. It is observed that the network roughly learns the locations of the AMD related abnormalities in the OCT volume, but no quantification is provided. [8] trains a CNN on 2.6 million OCT scans to classify them as being normal or having AMD. It also visually shows the localization of abnormalities, but provide no quantification of the localization.

Work done while at SigTuple Technologies, Bengaluru

3. METHODOLOGY

Inspired by the work of [12], we train a network with a Global Average Pooling (GAP) [9] layer right after the last convolutional layer to simultaneously perform image classification and weakly supervised localization. A GAP [9] layer applied to an input with spatial dimensions $n \times m$ with k feature maps produces a vector v of size k with each v_i defined as in equation (1). M_i is the i^{th} feature map of the input with activation $a^i_{x,y}$ at position (x, y). The output of the GAP layer is then fed into a fully connected layer to calculate the class activations for the L^{th} class as shown in equation (2). $w_{i,L}$ is weight connecting the i^{th} value of v to the output of class L. The class probabilities can be calculated from these class activations by applying a softmax on them. The class activation maps (CAM) [12] or the discriminative feature localizations for the L^{th} class are then calculated as in equation (3)

$$v_i = \text{Avg}(M_i) = \frac{\sum_{x,y}^{n,m} a^i_{x,y}}{nm} \quad (1)$$

$$A_L = \sum_{i=1}^{k} v_i w_{i,L} \quad (2)$$

$$\text{CAM}_L = \sum_{i=1}^{k} w_{i,L} \times M_i \quad (3)$$

Each CAM_L can be interpreted as the localization of the most discriminative features of an image that make it belong to the class L. Recent work by [3] suggests that using lower convolutional layers as input to the GAP layer produces better localizations at the cost of classification accuracy. To combine the features of the lower convolutional layers without significantly sacrificing classification accuracy, we replace the traditional convolutional blocks with Dense Blocks [4] where every convolutional layer uses the concatenated feature maps of all the convolutional layers in the block which appear before it.

4. EXPERIMENTS AND RESULTS

In this work we compare a number of neural networks with different architectures.

4.1. Data and Preprocessing

Our dataset consists of 408 cystic and 809 normal retinal OCT scan images collected from a local hospital. The frames were randomly selected from scans acquired by a Heidelberg Spectralis OCT machine. The frames were then labeled as examples of cystic edema (FFR) or normal by a panel of three expert ophthalmologists. The dataset was randomly split into training (90%) and validation (10%) sets. A separate test set of 117 images with expert annotated ground truth bounding polygons (not exact segmentation masks) for all visible cysts was used to calculate the localization metrics of each model. Each scan in the dataset was resized to 256×512 and denoised using Non-local Means Denoising [2].

4.2. Localization and Classification Metrics

For each scan in the test set, we calculate the CAM for the cystic class on the original scan and its horizontally flipped version. We then appropriately align the two CAMs and take their average. Next, we binarize the CAM with a threshold which is a fraction of the maximum value of the CAM. We treat these binary masks as the localization predictions of our network. A localization is counted as a true positive if the predicted localization overlaps at least 10% of the ground truth bounding polygon. If the overlap is less than 10%, then it is counted both as a false positive and as a false negative. We then calculate the precision and recall at different thresholds and plot the precision-recall curve. The area under this curve gives us the Average Precision (AP) of a particular network for the cystic class. Since this is a binary classification problem, the average precision becomes the mAP. The F1-Score for each network is then calculated at the best threshold value.

The classification ability of our networks was measured by calculating the accuracy, sensitivity and precision of each network on our test set.

4.3. Models

In an attempt to combine the outputs of the features of the lower level convolutions without significantly impacting the accuracy, we used Dense Blocks [4] in our CNN architecture. We also experimented with the SeLU and ReLU activation functions and the batch normalization [5] technique. In all our networks, we have made extensive use of rectangular 3×5 kernels to ensure the ratio of the receptive field along each dimension is equal to the aspect ratio of the original image. The details of the best performing networks are as follows:

- **SeLU-DenseNet:** This network uses the SeLU nonlinearity after each convolution and uses Dense Blocks. The network consists of a downsampling block (described in Table 1) and feature learning blocks. The downsampling block's purpose is to compress the input to a good lower dimensional representation. It is followed by a feature learning block which consists of four dense blocks [4] each followed by a convolution with 32 output feature maps. As shown in Table 2, the dense block contains 3 convolutions with a growth rate of $k = 32$. All the convolutions in the feature learning block use a 3×5 kernel and "same" padding. This network was trained using the Adam optimizer [6] with a learning rate of 1×10^{-5} and a L2 regularization of 5×10^{-4}.
- **SeLU-ConvNet:** The network consists of a downsampling (Table 1) and a feature learning block with 12

Layer num.	Num. feature maps	Kernel size	Stride size
1	1	2×2	2
2	16	3×5	1
3	16	2×2	2
4	32	3×5	1
5	32	2×2	1
6	32	2×2	2

Table 1: Structure of the downsampling block. All layers are convolutional.

Layer num.	Num. feature maps	Kernel size	Input layers
1	32	3×5	Input
2	32	3×5	Input,1
3	32	3×5	Input, 1, 2

Table 2: Structure of the dense block. All layers take as input concatenation of outputs from previous layers in the block. The final output is the concatenation of all layers and input.

convolutional layers each with a 3×5 kernel, 64 output feature maps and "same" padding. All the convolutions in the network are followed by a SeLU nonlinearity. This network was trained using Adam with a learning rate of 1×10^{-5} and a L2 regularization of 5×10^{-4}.

- **ReLU-DenseNet:** The network consists of a downsampling block (Table 1) and a feature learning block with 1 dense block (Table 2) which is followed by a convolution with 32 output feature maps. All the convolutions in the network are followed by batch normalization [5] and a ReLU nonlinearity. This network was trained using Adam with a learning rate of 1×10^{-4} and a L2 regularization of 5×10^{-3}.

- **ReLU-ConvNet:** The network consists of a downsampling block (Table 1) and a feature learning block with 4 convolutional layers each with a 3×5 kernel, 64 feature maps and "same" padding. All the convolutions in the network are followed by batch normalization and a ReLU nonlinearity. This network was trained using Adam with a learning rate of 1×10^{-4} and a L2 regularization of 5×10^{-3}.

Multiple versions of each architecture were tried out, with varying learning rates, L2 regularization, etc. Only the best performing version of each architecture are described above.

Network	mAP	F1 Score
SeLU-DenseNet	**0.75**	**0.86**
SeLU-ConvNet	0.70	0.84
ReLU-DenseNet	0.18	0.71
ReLU-ConvNet	0.09	0.45

Table 3: Localization Metrics on True Positive predictions

Network	mAP	F1 Score
SeLU-DenseNet	**0.63**	**0.82**
SeLU-ConvNet	0.55	0.76
ReLU-DenseNet	0.05	0.25
ReLU-ConvNet	0.04	0.28

Table 4: Localization Metrics on all abnormal predictions

4.4. Results

As can be seen from Tables 3, 4 and 5 **SeLU-DenseNet** and **SeLU-ConvNet** have significantly outperformed **ReLU-DenseNet** and **ReLU-ConvNet** in both classification and localization. **SeLU-DenseNet** performed better than **SeLU-ConvNet** in localizing the FFRs while both have the same classification accuracy on the test set. Tables 3 and 4 suggest that the Dense Blocks [4] seemed to have significantly helped the neural network with SeLU activations to localize better. This seems to be due to the property of SeLU networks converging better with lower amounts of data [7]. Figures 2 and 3 show the localizations produced by the **SeLU-DenseNet** on a few scans from the test dataset.

Despite multiple attempts with different architectures as well as carefully tuning hyper-parameters, we were not able to get the ReLU networks to converge at a reasonable test set accuracy as all of them suffered from severe overfitting.

5. CONCLUSION

Simultaneous localization and classification using weak frame level labels is an active area of research for medical images. In medical images, expert annotations for localization are tedious and time consuming unlike image level class labels. We show that we can simultaneously localize

Network	Acc.%	Sens.%	Prec.%
SeLU-DenseNet	94.8	96.6	**93.4**
SeLU-ConvNet	94.8	**98.3**	92.0
ReLU-DenseNet	45.2	6.7	30.7
ReLU-ConvNet	48.7	55.9	49.2

Table 5: Classification Metrics

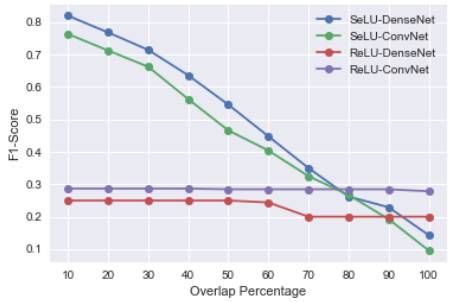
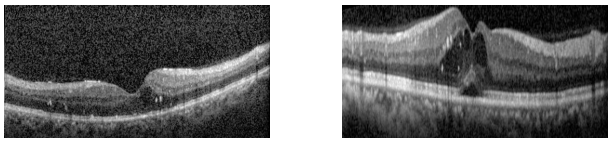
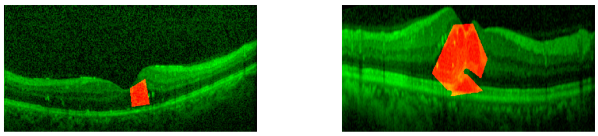


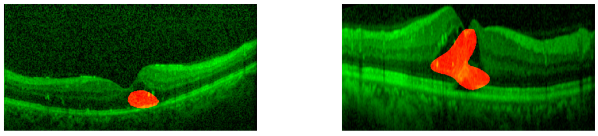
Fig. 1: The variation of the F1-Score with overlap percentage.



(a) Original scan



(b) Ground truth bounding Polygon



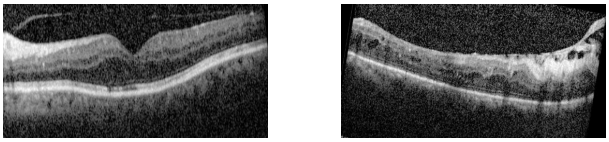
(c) Predicted Localization

Fig. 2: Two examples from the test set where **SeLU-DenseNet** has localized quite well. Here we have encoded the original scans in the green-channel.

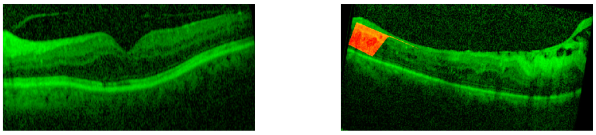
abnormalities in an image as well as classify the same image as abnormal or normal using image level class labels as a solution to the above problem.

Our experiments also show that using SeLU activations provides better convergence than ReLU activations and that Dense blocks gives better localization of FFRs in abnormal OCT scans than standard convolutions. Our method achieves a state of the art results with a true positive mAP score of 0.75 for localization and a precision of 93.4% with a sensitivity of 96.6% on the test dataset of OCT scans.

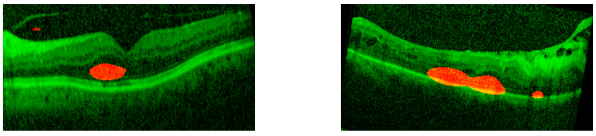
In future work, we hope to extend this method for other pathologies diagnosed with an OCT scan, including epiretinal membranes, hard exudates and RPE changes, spongy-edema, drusen etc., thereby being able to clearly localize and identify a set of clinically relevant set of abnormalities in a retinal OCT. In addition, we also would be analyzing the effect of more data on the above architecture. We also hope to prove that this method has potential applications for other medical imaging modalities.



(a) Original scan



(b) Ground truth bounding Polygon



(c) Predicted Localization

Fig. 3: Two examples from the test set where **SeLU-DenseNet** has not localized well. Interestingly, in both the cases the network has localized another type of abnormality called a spongy-edema.

6. REFERENCES

[1] S. Apostolopoulos et al. Retinet: Automatic AMD identification in OCT volumetric data. *CoRR*, abs/1610.03628, 2016.

[2] A. Buades, B. Coll, and J.-M. Morel. Non-Local Means Denoising. *Image Processing On Line*, 1:208–212, 2011.

[3] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini. Discriminative Localization in CNNs for Weakly-Supervised Segmentation of Pulmonary Nodules. *ArXiv e-prints*, July 2017.

[4] G. Huang et al. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[7] G. Klambauer et al. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017.

[8] C. S. Lee, D. M. Baughman, and A. Y. Lee. Deep learning is effective for classifying normal versus age-related macular degeneration oct images. *Ophthalmology Retina*, 1(4):322 – 327, 2017.

[9] M. Lin et al. Network in network. *CoRR*, abs/1312.4400, 2013.

[10] T. Schlegl et al. *Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks*, pages 437–448. Springer International Publishing, Cham, 2015.

[11] S. Scholl et al. Pathophysiology of macular edema. *Ophthalmologica*, 224(Suppl. 1):8–15, 2010.

[12] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.