Multi-task Deep Network for Ophthalmology Screening on Fundus Images

Lijing Song
Department of Computer Science
Stanford University
SUNet ID: lisasong
lisasong@stanford.edu

Bozhao Liu
Department of Computer Science
Stanford University
SUNet ID: bozhao91
bozhao91@stanford.edu

Abstract

We developed a multi-task deep learning system that can detect diabetic ophthalmical disease and glaucoma in the healthcare area. Exponential uneven weight binary cross entropy is designed as our loss function to solve the imbalanced data issue with much fewer ophthalmical images compared to healthy images. We trained our baseline model, AlexNet [9], ResNet [8], and DenseNet [6], and achieved state-of-the-art F1 scores of 97.46 and 91.67 on glaucoma and diabetic retinopathy on our private dataset.

- GitHub link: `https://github.com/supersmm/230project`
- Category: Healthcare, Computer Vision

1 Introduction

Glaucoma and diabetic ophthalmical disease are the two major leading causes of irreversible blindness among eye diseases. By the year 2040 it is projected there will be approximately 112 million glaucoma affected individuals worldwide [13] and around 500 million people have diabetes, with 34.6% have diabetic retinopathy and 7% have vision-threatening diabetic retinopathy [7]. Early screening is essential for early treatment to preserve vision and maintain life quality.

However, diagnosis based on fundus images made by human professionals can be error-prone and slow. Fundus images of diabetic ophthalmical disease are relatively easy to be diagnosed by human. Dark spots on fundus images can usually been seen as signals of diabetic ophthalmical disease. However, diagnoses based on images of glaucoma require more time and expertise, especially in the early stage. Late-stage glaucoma can be diagnosed by observing the yellow macular (often called disc), and diagnosing a single case could take several minutes for an expert. Moreover, human fundus' complexity with other symptoms such as retinal detachment makes quick, accurate, and economic diagnosis less approachable.

In recent years, deep learning has shown clinically acceptable diagnostic performance in detecting ophthalmical diseases. Applying deep learning for initial diagnoses can not only reduce the cost, but is also more efficient and accurate. In this paper, we developed a multi-task deep learning system for two ophthalmical tasks: glaucoma and diabetic retinopathy.

2 Related Work

For diabetic retinopathy, researchers in the healthcare area have switched from hand-crafted features to convolutional neural network to apply on the entire fundus pictures since the rise of deep learning. For example, Li et al. [10] has recruited 21 trained ophthalmologists to classify 48,116 fundus images and trained a convolutional neural network to extract clinical features for classification. Yet recently detection of red lesion has been explored to see if that can boost the performance. Orlando et al. [11] developed a red lesion detection system with hand-crafted features and convolutional neural network to detect red lesions in fundus pictures for diabetic retinopathy screening.

Same as diabetic ophthalmical disease, the initial applications of deep learning on Glaucoma diagnosis is using glocal fundus images as input to network for classificaction. However, recent research has developed a method of image segmentation of the yellow macular (disc) and/or cup segmentation. Fu et al. proposed a neural network architecture [2] for glaucoma screening with disc awareness and another neural network [3] for multi-task screening with disc and cup segmentation on a very limited dataset.

Architecture and performance of recent deep learning applications on diabetic retinopathy and glaucoma classification are summarized in Tbl. 1 and Tbl. 2. Note that those datasets vary a lot in quality and quantity and that performance may not be quite comparable between models, as some datasets are private and not available for public use.

Table 1: Summary of deep learning systems for diabetic retinopathy using fundus photos

Model	Year	Model	Dataset	AUC	Recall(%)	Precision(%)
Abramoff et al. [1]	2016	AlexNet/VGG	Messidor-2	0.98	96.8	87.0
Gulshan et al. [5]	2016	Inception-v3 [12]	Messidor-2	0.99	87	98.5
Gargeya et al. [4]	2017	(customized)	Messidor-2	0.94	N/A	N/A
Ting et al. [14]	2017	(customized)	SiDRP 14-15, etc.	0.936	90.5	91.6
Orlando et al. [11]	2018	(customized)	Messidor-2	0.935	N/A	97.2

Table 2: Summary of deep learning systems for glaucoma using fundus photos

Model	Year	Model	Dataset	AUC	Sensitivity(%)	Specificity(%)
Ting et al. [14]	2017	VGG-19	SiDRP 14-15	0.942	96.4	93.2
Li et al. [10]	2018	Inception-v3 [12]	LabelMe [1]	0.986	95.6	92.0
Fu et al. [3]	2018	M-Net [2]	ORIGA [3]	0.878	N/A	N/A
Fu et al. [2]	2018	DENet [4]	SCES	0.8316	70.7	N/A

3 Data

Through academic connections, we collected the following private data set from Rjukan Synssenter Optometri:

1. 390 fundus images of glaucoma
2. 602 fundus images of diabetic retinopathy
3. 7,362 healthy fundus images

Due to the limited amount of data that's collected in the period of the course, the original idea of splitting the images into 80-10-10% doesn't work well as we won't be able to validate the result using only 29 glaucoma fundus images. Instead, randomly splitting the dataset into 60-20-20% would be more reasonable. The final data distribution is:

- Training set: 4,427 healthy images; 361 diabetes images; 188 glaucoma images
- Validation set: 1,472 healthy images; 120 diabetes images; 58 glaucoma images
- Test set: 1,473 healthy images; 121 diabetes images; 58 glaucoma images

Even though the dataset was distributed into 60-20-20%, the overall data set is still dramatically imbalanced. To even this up, we decided to do data augmentation include horizontal flipping, vertical flipping and 180° rotating only to the diabetic and glaucoma images, not the healthy ones. We've also designed exponential uneven weight binary cross entropy loss function, as detailed in Section 4.2.

3.1 Input and Output

The input data are a fundus image and its label of ophthalmical disease type in an array ([0, 0]-healthy, [0, 1]-diabetic, [1, 0]-glaucoma, [1, 1]-diabetic+glaucoma). Although we didn't receive images that are labeled with [1, 1]-diabetic+glaucoma, we still kept it there so that the framework can be extended to future diabetic+glaucoma instances and more tasks.

Fig. 1 and Fig. 2 are two examples of raw input data, of various sizes and from difference devices. Fig. 3 and Fig. 4 are another two examples of transformed input data, of the same size 224×224.

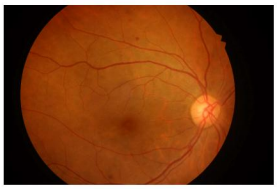

Figure 1: Raw input example: Diabetes [0,1]

Figure 2: Raw input example: Glaucoma [1,0]

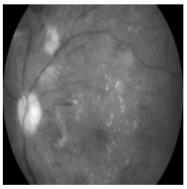

Figure 3: Transformed input example: Diabetes [0,1]

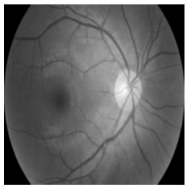

Figure 4: Transformed input example: Glaucoma [1,0]

3.2 Evaluation

Metrics for measuring binary classification can be used for evaluation. This includes F1 score and confusion matrix for individual tasks. We use the average F1 score of two tasks as measurement to select the best model. Since most research use specificity (precision) and sensitivity (recall), we also report them for comparison. Accuracy is also reported although it's not our main metric given the imbalanced data distribution.

4 Approach

We developed a multi-task classification model for glaucoma and diabetic ophthalmical disease detection, using glaucoma-only (vector [1, 0]), diabetic ophthalmology-only (vector [0, 1]), and very rare cases of both diseases (vector [1, 1]) data, versus healthy ([0, 0]) data. Our main methodology is CNN with transfer learning. Due to the limited amount of data, we applied AlexNet [9], DenseNet [6], and a smaller version of ResNet [8] with trainable weights to initiate the layers for fundamental features such as edge and shape detection for faster training.

4.1 Data Cleansing and Augmentation

The following data cleansing and augmentation process has been applied to all input images. Before this data transformation step, original images have various sizes including 2743×1936 and 2376×1584 pixels, etc.

1. Filter
   We added filters to remove humps in fundus images on the upper right corner caused by data collection process with special devices.
2. Downsize
   We resize images of various resolutions to 224×224 pixels. Image downsize also helps industrial applications on high-dimensional photographs collected by different devices.
3. RGB to Greyscale
   We removed the last dimension of every input, which responds to the color channel. Although the model performance decreased by a small degree, we believe it makes the model more robust to different devices, clinics, and data collection processes.
4. Generalization (a.k.a. prenorm)
   Unlike other computer vision problems, fundus images focus more on the pattern instead of the luminance, therefore we generalized the pixel value image to image by the same variance (means the highest pixel value of each image is 255) by the following equation: $p = (p/p_{max}) * 255$, where p_{max} is the value of the pixel in each image that the highest value within this image.
5. Normalization
   In image processing, normalization is a process that changes the range of pixel intensity values. We normalized all the image pixel values to be within [-1, 1] for faster training. In our case, normalization can also ensure consistent contrast to remove contrast variance from multiple ophthalmical devices.
6. Data augmentation on unhealthy images in the train and val sets
   We duplicate unhealthy images and flip the replicas horizontally, vertically, and rotate replicas by 180 degrees only for the unhealthy images in the train and val sets.

After the above data cleansing and augmentation steps, input data become of the same size of 224×224 pixels and in greyscale. Variance among various devices and collection processes can be removed to the largest degree. We also ended up with better balanced data with more unhealthy images after the data augmentation. Examples of input images after this data transformation step are Fig. 3 and Fig. 4 in Section 3.1.

4.2 Loss Function *[Original]*

In general we applied binary Cross Entropy as the loss function. However, during training the network tends to shift towards predicting all images as healthy to improve the accuracy. In order to solve this problem we tried Binary Cross Entropy with uneven weight. Although this method has some effect, it still didn't solve our problem completely.

We designed and applied our original exponential uneven weight binary cross entropy (Exp_{UWBCE}) loss as the loss function. Considering the characteristic of the project where recall is considered more important than precision and the small amount of data where the label is '1', we created our own loss which we call it exponential uneven weight binary cross entropy (Exp_{UWBCE}) loss, formula as follows:

$$loss_{Exp_{UWBCE}} = W_1 * (\exp(-y * \log(\hat{y})) - 1) - W_2 * (1 - y) * \log(1 - \hat{y})$$

In practice we set $W_1 = 1, W_2 = 0.3$. Here we enforce the gradients from data whose labels are 1 to be dramatically bigger than label 0's when the prediction is wrong, therefore pushing the learning to prioritize samples whose labels are 1 to make sure that recall gets improved. When the learning proceeds to a high recall state, the gradients of positive labeled predictions will get close to 1 which is the same as normal Cross Entropy loss.

5 Experiments and Result Analysis

We conducted a series of experiments with multiple architectures and various hyperparameters. The performance of each architecture on the **val** set is summarized in Tbl. 3, Tbl. 4, and Tbl. 5. For simplicity purpose, models with same architectures are only listed with the best performance after hyperparameter search. We set the drop rate to 0.8, and the number of epochs from 50 to 200 depending on the network (10 in the baseline model). Batch size varies between 30 and 100.

When applying AlexNet, DenseNet and ResNet, we started to apply Exp_{UWBCE} loss with the same learning rate as BCE experiments. However as we added exponential to the loss function, the training process showed serious large gradient problem making the model really hard to converge with higher learning rate. Thus we lowered the learning rate significantly. After several iterations of training we lowered the learning rate even more to make sure the model converge to an optimal point.

Table 3: Model performance (in %) on val set with uneven weighted cross entropy loss

	Glaucoma				Diabetic Retinopathy			
Model (lr, wd)*	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Baseline (0.001, 0.0001)	98.89	41.93	36.02	68.74	97.58	79.57	67.69	84.61
AlexNet (0.0003, 0.0005)	99.76	67.96	96.67	78.06	98.85	79.02	90.93	84.56

* Hyperparameters: lr = learning rate, wd = weight decay in Adam.

Table 4: Model performance (in %) on val set with Exp_{UWBCE} loss and higher (1e-4 level) learning rate

	Glaucoma				Diabetic Retinopathy			
Model (lr, wd)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
AlexNet (0.0003, 0.0005)	98.85	87.65	99.16	92.74	99.45	91.00	92.79	91.89
ResNet (0.0001, 0.0005)	98.06	79.19	99.16	88.06	99.64	94.74	94.74	94.74
DenseNet 121 (0.0003, 0.0005)	98.85	89.76	95.00	92.31	98.61	71.60	100.00	83.45

Table 5: Model performance (in %) on val set with Exp_{UWBCE} loss and lower (1e-5 level) learning rate

	Glaucoma				Diabetic Retinopathy			
Model (lr, wd)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
AlexNet (0.00001, 0.0005)	99.64	97.53	97.53	97.53	99.70	94.19	96.05	95.11
ResNet (0.00001, 0.0005)	99.88	99.16	99.16	99.16	99.94	98.28	100.00	99.13
DenseNet 121 (0.00001, 0.0005)*	99.94	100.00	99.17	99.58	99.94	98.31	100.00	99.15

* Our best model

Our best model is DenseNet-121 using Exp_{UWBCE} loss with 1e-5 learning rate and 5e-4 weight decay. On the **test** set, it achieves 97.12% F1 score on glaucoma with 96.72% precision and 97.52% recall, and 92.98% F1 score on diabetic retinopathy with 94.64% precision and 91.38% recall.

1. **Baseline model**
   To establish baseline performance, we developed a convolutional network. It consists of 3 convolution layers followed by batch normalization layers that help stabilize training. Fully connected layers and batch normalization layers are then added to transform the output of convolution layers. In the end 2 fully connected layers and log softmax activations are added in parallel to generate the multi-task outputs. The baseline model uses uneven weighted cross entropy loss.
2. **AlexNet**
   AlexNet [9] outperformed our baseline model although they have similar architectures. It shows that deeper network with more layers, non-saturating ReLU activation function, and weights pretrained from bigger datasets can help improve the performance largely.
3. **ResNet**
   To improve model performance, we developed a residual-based network on top of the ResNet model in Johnson et al. [8]. Our residual-based networks outperformed AlexNet model and baseline in F1 score for both tasks. Residual connection does not only make deep learning models easier to improve based on residuals, but also adds end-to-end learning opportunities since it gives the model flexibility to choose either to keep the information from previous layers or to discard it without information passed. Due to the big number of parameters in ResNet, we shrank the depth of the original ResNet-18 to half for faster training.
4. **DenseNet**
   In our experiment, DenseNet-121 [6] achieves the best results. Its shorter connections between layers close to the input and those close to the output help alleviate the problem of vanishing gradient and strengthen feature propagation, with fewer parameters. We also prepared the project with other DenseNet models like 161, 169, 201, but as DenseNet-121 already performs really well, we decided to not to use the other DenseNet versions in the current phase of the project.

6 Conclusions and Future Work

In this project we applied convolutional neural network to classify fundus images to healthy, glaucoma, and/or diabetic retinopathy within a multi-task framework. Here are a few things we learned through the process:

1. Our original exponential uneven weight binary cross entropy loss improved the performance to a very large degree with respect to recall, which we consider to be the most important evaluation metric, along with precision and accuracy.
2. After lots of iterations over model architectures and hyperparameter search, DenseNet is selected as our best model with industry-leading performance.
3. To train the models well with our original exponential uneven weight binary cross entropy loss, we had to set the learning rate cautiously. Without small learning rate and weight decay, the large gradients from data points with 1 labels will make too big updates and overshoot the global minimum.
4. We outperformed similar research papers mainly because of the limited size and diversity of our private dataset. It's possible to achieve 100% recall on 58 glaucoma images in the val set, but almost impossible on thousands of glaucoma images.

Despite the good model performance, our work is limited to the short course time frame, small and imbalanced dataset, and computing power. For further improvement, future work may include but are not limited to the following items:

- Heatmap Localization
  For better understanding of the model and assistance in human professionals' diagnosis, attention layers for highlighting key features that maximize the activations in a heatmap can be added.
- Image Segmentation
  To achieve better performance, image segmentation can also be added to segment disc and/or cup and use the segmented areas for further classification, with manually labeled data.
- Retrieve more data
  We have already arranged meetings with an eye clinic in Oslo and Rjukan Synssenter Optometri for retrieving more images with different labels as long as OCT images to expand the scope of the course project.

Contributions

Both authors contributed equally to the project. Lijing built the pipeline for training and evaluation with imbalanced data loader and baseline model for multi-task learning and data augmentation. Bozhao collected data set, created the filters to process raw inputs, conducted the transfer learning from AlexNet, ResNet, and DenseNet as well as the pipeline of training these models, tried out Exponential Uneven Weight Binary Cross Entropy loss. Both authors contributed to data cleansing, research over existing projects, the model architecture development, hyperparameter tuning, result analysis, discussion over the new loss function characteristics and report writing.

Acknowledgments

Thanks to teaching assistant Aarti Bagul and all the CS230 teaching staff for guidance on this project. Thanks to Pr. Per Olof Lundmark and Dr. Knut Luraas for providing the fundus images and consulting over diabetes and glaucoma diagnoses.

References

[1] Michael David Abràmoff et al. "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning". In: *Investigative ophthalmology & visual science* 57.13 (2016), pp. 5200–5206.

[2] Huazhu Fu et al. "Disc-aware ensemble network for glaucoma screening from fundus image". In: *IEEE transactions on medical imaging* 37.11 (2018), pp. 2493–2501.

[3] Huazhu Fu et al. "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation". In: *IEEE transactions on medical imaging* 37.7 (2018), pp. 1597–1605.

[4] Rishab Gargeya and Theodore Leng. "Automated identification of diabetic retinopathy using deep learning". In: *Ophthalmology* 124.7 (2017), pp. 962–969.

[5] Varun Gulshan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22 (2016), pp. 2402–2410.

[6] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[7] A Hutchinson et al. "Effectiveness of screening and monitoring tests for diabetic retinopathy–a systematic review". In: *Diabetic medicine* 17.7 (2000), pp. 495–506.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European conference on computer vision*. Springer. 2016, pp. 694–711.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[10] Zhixi Li et al. "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs". In: *Ophthalmology* 125.8 (2018), pp. 1199–1206.

[11] José Ignacio Orlando et al. "An ensemble deep learning based approach for red lesion detection in fundus images". In: *Computer methods and programs in biomedicine* 153 (2018), pp. 115–127.

[12] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[13] Yih-Chung Tham et al. "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis". In: *Ophthalmology* 121.11 (2014), pp. 2081–2090.

[14] Daniel Shu Wei Ting et al. "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes". In: *Jama* 318.22 (2017), pp. 2211–2223.

[15] Zhuo Zhang et al. "Origa-light: An online retinal fundus image database for glaucoma analysis and research". In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 2010, pp. 3065–3068.