Multi-Task Deep Network for Ophthalmology Screeing on Fundus Images

Video Link: https://youtu.be/pyd4yWYcX2g

Lijing Song (lisasong), Bozhao Liu (bozhao91)

Introduction

Early screening is essential for early treatment of ophthalmology to preserve vision and maintain life quality. However, diagnosis based on fundus images made by human professionals can be error-prone and slow.

Applying deep learning for initial diagnoses can not only reduce the cost, but is also more efficient and accurate. In this project, we developed a multi-task deep learning system for two ophthalmical tasks: glaucoma and diabetic retinopathy, which are the two major leading causes of irreversible blindness among eye diseases.

Data

Through academic connections, we collected the 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

We split them into 60-20-20% for train, val, and test.

- Multi-task labels:[0, 0]-healthy, [0, 1]-diabetic, [1, 0]-glaucoma, [1, 1]-diabetic+glaucoma
- Raw input examples:

Figure 1: *Diabetes [0,1]*

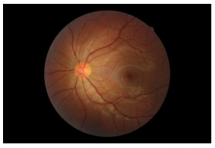

Figure 2: *Glaucoma [1,0]*

- Transformed input examples:

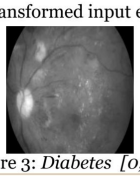

Figure 3: *Diabetes [0,1]*

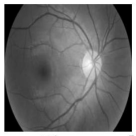

Figure 4: *Glaucoma [1,0]*

Approach

1. Data cleansing
   Filter -> Downsize -> Greyscale -> Prenorm -> Normalization
2. Data augmentation
   Only on unhealthy images
3. Metric
   average F1 score of the two tasks to select the best model.
4. Original loss function: Exponential uneven weight binary cross entropy

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

5. Set small learning rate
   Because our gradients are relatively large.

[Intuition]: We enforce the gradients from data whose labels are 1's to be dramatically bigger than label 0's when the prediction is wrong.
We prioritize samples whose labels are 1 to make sure that recall gets improved. When learning proceeds to a high recall state, the gradients of positive labeled predictions will get close to 1, same as normal Cross Entropy loss.

Experiments

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 our custom loss. 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.

Conclusions

- Our original loss improved the performance greatly with respect to recall.
- 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.
- We outperformed similar research papers mainly because of the limited size and diversity of our private dataset.

Future Work

1. 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.
2. 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.
3. Retrieve more data

Acknowledgments

We would like to thank the CS230 teaching staff and TA Aarti for guidance.