Deep Learning Applied to the Diagnosis of Thyroid Eye Disease

Team #11: Galen Wong, Rishab Sukumar, Shirly Fang

I. Methods

A. Data Pre-processing

1) Images from the Internet: The dataset we were initially provided with had 799 rectangular images of patients' eyes scraped from the internet. These images comprised patients with Thyroid Eye Disease, other eye diseases, and no ailments. We processed the provided images in two different ways.

To implement transfer learning using models trained on ImageNet we initially cropped the images in the internet dataset to obtain square images for the left and right eyes. We separated each pair of images such that 80% of TED, NO TED, and OTHER images went to the "training" set and 20% went to the "validation" set. Using PyTorch's ImageFolder dataloader, we applied a series of transforms to the images in the training and validation sets. For the training set we implemented a `RandomResizedCrop` to randomly scale each image between size multipliers of 0.8 and 1 while ensuring a 1:1 square crop with an edge size of 224. We also implemented the `RandomHorizontalFlip` and `Normalize` transforms. The input is normalized by rescaling the data to have a zero mean and a standard deviation of one. This normalization is required to make the input data similar to that on which ResNet is trained. For the validation set we implemented a `CenterCrop` to crop each image about its center to an aspect ratio of 1:1 with an edge size of 224.

We later decided to apply our model on rectangular images. If an image's aspect ratio is greater than 2, the preprocessor crops the width on both left and right sides and extracts the middle portion of the image. If the aspect ratio is less than 2, the preprocessor crops the height on both top and bottom sides and extracts the middle portion of the image. This transformation leads to images with 2:1 aspect ratios that are compatible with our model. We still implemented an 80/20 split between images in the training and validation sets. Rather than using the transforms `RandomResizedCrop` and `CenterCrop` for the training and validation sets we used the `Resize` transform to ensure a 2:1 rectangular crop with dimensions 224x448. The remaining transforms remained the same.

2) Patient Images: The patient dataset consists of over 3600 patient images. These images are binary classified and are labeled as either TED or NON_TED. The aspect ratios of the patient dataset are more variable than those in the internet dataset. These ratios generally range from 1:2, 1:1, and 2:1. Our model at this point in time was compatible with 2:1 images with both eyes. The eye placements in the images vary but generally they appear in the middle half of the image. Thus, we chose to continue using the same preprocessor, data transforms and 80/20 split that was used on the internet dataset, mentioned in section I-A.1.

B. Convolutional Neural Network

1) Transfer Learning: Convolutional Neural Network (CNN) is chosen as the architecture to handle the classification task. Considering the small size of the initial data set, we decided to use transfer learning [1]. The experimentation is done on networks that are pre-trained on the ImageNet dataset [2]. Our data are images that capture what a human sees, which is similar to the ImageNet database. Also, ImageNet requires sub-categorical classification. For instance, the dog category contains sub-categories such as "Husky" and "Eskimo" that are specific to the breed. Our dataset contains images of eyes only and we are classifying different "types" of eyes. Given the similarity in the nature of the problem, CNNs trained on ImageNet are chosen to be the basis of our model.

Unlike in other transfer learning tasks where some layers are frozen (parameters being not adjustable through back-propagation) such that those layers are used as feature extractors, we unfreeze all the layers such that we can fine tune the parameter in training. The reasoning is that the task is not a "subtask" of ImageNet classification, since the images that is used in this case are not contained in ImageNet. Therefore, the paramters need to be fine-tuned to solve our specific classification problem.

2) Layer Replacement For Different Input Dimension: All images in ImageNet are square images. Therefore, all the models also take square images (224×224) as inputs, which is incompatible with our dataset, which comprises of mostly rectangular images. Our first idea is to split up the rectangular images into 2 square images as a way of making the data compatible, which is explained in the section I-A.

The second approach is to swap out the incompatible layers within the network. The networks used are comprised of multiple convolutional layers, followed by a fully connected layer. The convolutional layers are input-size independent. Only the fully connected layer is affected by input size. As a work around, we replace the final fully connected layer with a new one that is compatible with the input size 224×448, allowing the model to be applied on rectangular images.

C. Evaluation

Cross Entropy Loss is used as the loss function to evaluate the performance of the model. We also uses the confusion matrix to evaluate the accuracy of the models.

II. Results

The accuracy summary of the models is in table I.

A. Experimentation on Small Dataset

1) Vanilla ResNet18: The first attempt was made with ResNet18, in which the rectangular images of the eye are split into 2 squares images for the left and the right eye. We trained the model with 2 labels (TED vs NON_TED) and 3 labels (TED vs OTHER_DISEASE vs NO_DISEASE) separately. The validation accuracy for 3 labels classification is 84.27%. The 2 labels classification accuracy is 88.35%.

2) Augmented ResNet18 for Non-square Images: A new model is trained to classify rectangular images instead of square images. The hypothesis is that rectangular images contains more information, which helps to train the network better. Also, for some patients, the symptoms of TED might manifest in only one eye. Yet, the other eye with mild or no symptoms is also labeled as TED. This means that our model is learning to label some eyes without any symptoms also as TED, reducing its accuracy.

The model trained with 3 labels attained a validation accuracy of 90.56%. The one with 2 labels attained a validation accuracy of 91.19%. The result matches our hypothesis. The preliminary trial has proven that classification task with CNN is realizable. Next, different techniques are used to improve the validation accuracy.

From the result we can also notice that the validation accuracy for binary classification does better than tertiary classification in both cases. By plotting the confusion matrix (see figure 1 in appendix), we see that the OTHER and NO_DISEASE labels are often mis-classified with each other. Since our goal is to identify TED effectively but not to tell apart whether a person has eye related disease or not, we move on with our experimentation with 2 label classification only.

3) Antialiasing and Contrasting: The MaxPool layers of ResNet introduces aliasing since MaxPool ignored Nyquist sampling theorem [9], making the network shift-variant and prediction will be affected by the shift in the images. We use the anti-aliased version of the ResNet from Zhang [9], in hope to improve accuracy.

We inspected the misclassified images from the validation set and notice that there are a lot of TED examples showing redness around to eyes (see example in figure 6). We could amplify these features by increasing the contrast of the images, which will in turn increase the saturation of the color as well.

We train the model on the 2 label rectangular model and got a better validation accuracy of 93.08%.

4) Snapshot Ensembling: The next approach to improve our model's accuracy included the implementation of Snapshot Ensembling on top of our baseline binary classified model.

The intrinsic nature of the Stochastic Gradient Descent optimizer used by our model is such that we start with a somewhat high learning rate to move closer to a flat local minimum and then the learning rate drops so that the model may converge to a final local minimum in the deep neural network [4]. However, there is no way to tell if the final local minimum gives us the best model for the classification of TED. Snapshot Ensembling is used to get around the problem of being stuck to a single local minimum by converging to multiple local minima and escaping each of them using a large, aggressive learning rate [4]. The model weights at each local minimum are saved as a snapshot. The validation accuracy is obtained by averaging the accuracies across all the saved snapshots.

The model was trained over 3 cycles with 20 epochs each. When trained with 2 labels on rectangular images, this model attained a validation accuracy of 93.08%. Thus, Snapshot Ensembling showed a significant improvement in accuracy for models trained on both square and rectangular images.

5) Cyclic Learning Rate: Snapshot Ensembling improved our model's accuracy by a significant amount. However, considering that ensembling inherently attains the average accuracy across multiple different local minima, we sought to capture the model weights at the local minimum where the best validation accuracy was attained. Thus, we modified the model that used snapshot ensembling to simply save the model with the best validation accuracy over multiple cycles. This model still cycles its learning rate aggressively to converge at multiple local minima. However, only a single, best model is saved.

Similar to the model using Snapshot Ensembling, this model was trained over 3 cycles with 20 epochs each. When trained with 2 labels on rectangular images, this model attained a validation accuracy of 94.97%. The relationship between the accuracy and the number of epochs/cycle run is pictured in figure 2. The cyclic learning rate model showed an improvement over regular snapshot ensembling.

6) ResNeXt101_32x8d with Cyclic Learning Rate: ResNeXt is another model trained on ImageNet. It improves upon the ResNet model with the addition of a new dimension called "cardinality" [5]. ResNeXt has gotten better results with both 1K-way and 5K-way classification of images in ImageNet (see figure 3). We attempted to improve our model by using transfer learning with ResNeXt101_32x8d, the most accurate ResNeXt model available in PyTorch. We implemented this along with contrasting and a cyclic learning rate as specified in sections II-A.3 and II-A.5.

This model, when trained with 2 labels on rectangular images, attained a validation accuracy of 95.59%. Thus, transfer learning with ResNeXt using contrasting and a cyclic learning rate showed the best accuracy across the methods used on the internet dataset.

B. Experimentation on Patient Dataset

While the internet dataset has less variability between images and most have the eyes as the main focus of the image, the patient dataset does not. Some images start off as full face images while others are already more centered on only the eyes. There are also various magnification levels between images. Most images have the patient looking in the direction of the camera, but for some the patient is looking

to the side. Again, these variables lead to some issues in preprocessing and ultimately led to a more diverse set of processed images than the internet dataset. We believe these characteristics may have led to lesser accuracy from our model.

1) Augmented ResNet18: Our baseline rectangular model with ResNet18 reached a validation accuracy of 90.4%. This is slightly less than the 91.19% achieved on the internet dataset. This may be due to the variability of images and the difficulty of processing them properly.

2) Snapshot Ensemble: Our next approach included the addition of snapshot ensembling on top of the baseline rectangular model. This was able to achieve 91.66% validation accuracy. This is less than the 93.08% achieved on the internet dataset with the same model. This could be because the ensemble members were not as diverse and did not have as differing distributions of prediction errors [4].

3) Anti-aliasing and Contrasting: Next, we included the addition of anti-aliasing and contrasting transformations on top of the baseline rectangular model. This was able to achieve 90.04% validation accuracy. This is about the same accuracy as the baseline model, showing this technique was not effective for this dataset compared to the internet dataset, which was able to improve to 93.08% accuracy.

4) Cyclic Learning Rate: Our next model included the addition of the cyclic learning rate technique on top of the baseline rectangular model. This was able to achieve 90.10% validation accuracy. This is about the same accuracy as the baseline model, showing this technique was not effective for this dataset compared to the internet dataset, which was able to improve to 94.97% accuracy.

5) ResNeXt101_32x8d with Cyclic Learning Rate: Our next model, which utilized transfer learning with ResNeXt using contrasting and a cyclic learning rate was the most successful model on the internet dataset, as stated in section II-A.6. However, on the patient dataset, it was only able to achieve 90.70% validation accuracy.

Overall, our models performed a bit poorer on the patient dataset than on the internet dataset. Our thinking of why this may be the case is highlighted in the previous sections and in section III-A. The Resnet18 model with snapshot ensembling performed best on the patient dataset, achieving an accuracy of 91.66%. This is a different model from the most successful model on the internet dataset, and achieved a lower accuracy. The confusion matrices of these different approaches do not show a bias towards a certain misclassification (see figure 4). Our model did not misclassify TED as NON_TED more than vice-versa. It is observed that TED images that were misclassified as NON_TED tended to have subtle symptoms of TED. East Asian configuration eyes are more likely to be misclassified as NON_TED because they tend to have subtle physical manifestations.

III. Discussion

A. Limitations

The internet dataset that we worked with for the majority of the project consisted of about 800 images. This would be considered small data, in fact even our patient dataset of 3600 images can still be considered small data. Small samples are common in a field like medicine where data can mainly only be collected from clinical interactions with patients [6]. Because our data is limited, this can cause some problems such as outlier handling and overfitting among others.

We were able to notice some overfitting after working with the patient dataset. Our training accuracy inched closer to 100% while validation hovered around 90% across several epochs. The variability in the images required further human data review. However, this was only done to the extent of removing fully cropped out eye images and the resulting processed images were still very diverse. The greater inconsistencies between the images may have led to the model's poorer performance on the patient dataset.

B. Future Improvement

Given the limited time that we had to experiment on the large patient data set, we were not able to achieve the same level of accuracy as the smaller data set. The model can be improved in the following few ways:

1) Data Cleaning: As mentioned above in section I-A, the data is not as well cleaned up compared to the small data set. We were only able to use a simple algorithm to attempt to normalize the data set as much as possible. We would ideally like to use an eye-detection algorithm to detect the best crop which will give us better processing.

2) Prevent Overfitting: The model on the large data set exhibits overfitting. We did not have enough time at the end to optimize our model to reduce the overfitting (See figure 5). If possible, we would like to increase the accuracy even more by introducing generalizing techniques such as but no limited to: (1) data augmentation with AutoAugment [7], (2) introducing dropout to the model.

3) Enable Extra Feature Input: Several symptoms of TED include eyelid retraction, lid lag, exophthalmos, etc [8]. If the model could be trained with symptoms as extra features alongside the image, the accuracy can be further improved, since some symptoms such as dry eyes might not be visible through images of the eye.

C. Potential Impact

Our best binary model can be used as a preliminary screening tool to rule out other diseases or non disease when thyroid eye disease is being considered. Our 90%+ validation accuracy shows our model is a viable tool to support physicians' impressions that thyroid eye disease plays a role in the patient's physical eye manifestations. We believe it is especially useful for patient interactions with first-contact clinicians that are inexperienced with thyroid eye disease.

IV. Acknowledgement

We would like to thank Dr Karlin for providing the dataset and providing feedback on our project. Special thanks to Professor Scalzo for providing the hardware and infrastructure support enabling us to access patient data through UCLA network remotely.

REFERENCES

[1] Tan, Chuanqi, et al. 'A Survey on Deep Transfer Learning'. ArXiv:1808.01974 [Cs, Stat], Aug. 2018. arXiv.org, http://arxiv.org/abs/1808.01974.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[3] He, Kaiming, et al. 'Deep Residual Learning for Image Recognition'. ArXiv:1512.03385 [Cs], Dec. 2015. arXiv.org, http://arxiv.org/abs/1512.03385.

[4] Gao Huang, et al. 'Snapshot Ensembles: Train 1, get M for free'. arXiv:1704.00109 [cs.LG], Apr. 2017. arXiv.org, https://arxiv.org/abs/1704.00109.

[5] Xie, Saining, et al. 'Aggregated Residual Transformations for Deep Neural Networks'. ArXiv:1611.05431 [Cs], Apr. 2017. arXiv.org, http://arxiv.org/abs/1611.05431.

[6] EduPristine. Problems of Small Data and How to Handle Them. EduPristine. Feb. 2016. https://www.edupristine.com/blog/managing-small-data.

[7] Cubuk, Ekin D., et al. 'AutoAugment: Learning Augmentation Policies from Data'. ArXiv:1805.09501 [Cs, Stat], Apr. 2019. arXiv.org, http://arxiv.org/abs/1805.09501.

[8] Weiler, D. L. (2016). Thyroid eye disease: a review. Clinical and Experimental Optometry, 100(1), 20–25. doi: 10.1111/cxo.12472

[9] Zhang, Richard. 'Making Convolutional Networks Shift-Invariant Again'. ArXiv:1904.11486 [Cs], June 2019. arXiv.org, http://arxiv.org/abs/1904.11486.

APPENDIX

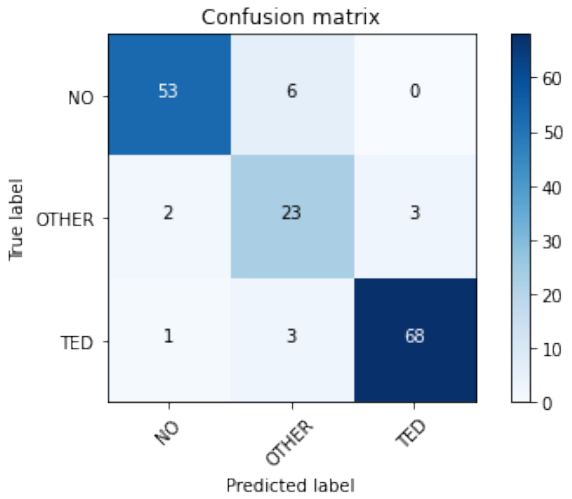

Fig. 1: Confusion matrix for 3 labels rectangular model

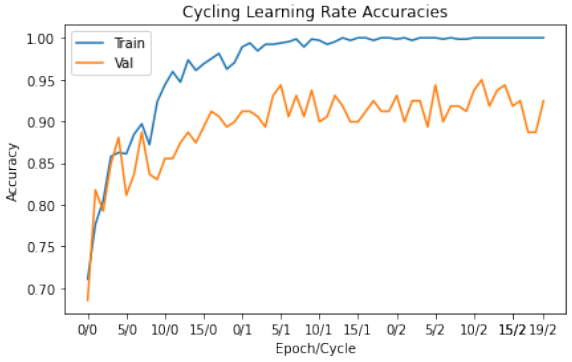

Fig. 2: Accuracy vs Epoch/Cycle using Cycling Learning Rate with internet dataset

	setting	5K-way classification		1K-way classification	
		top-1	top-5	top-1	top-5
ResNet-50	1 × 64d	45.5	19.4	27.1	8.2
ResNeXt-50	32 × 4d	**42.3**	**16.8**	**24.4**	**6.6**
ResNet-101	1 × 64d	42.4	16.9	24.2	6.8
ResNeXt-101	32 × 4d	**40.1**	**15.1**	**22.2**	**5.7**

Fig. 3: ImageNet Classification Results: Resnet vs ResNeXt [5]

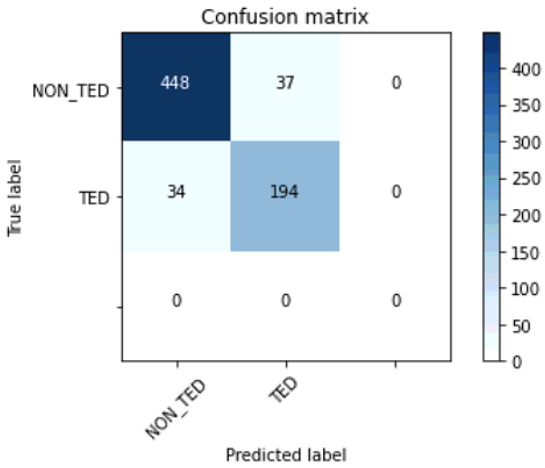

Fig. 4: Confusion matrix for 2 labels rectangular model on patient dataset

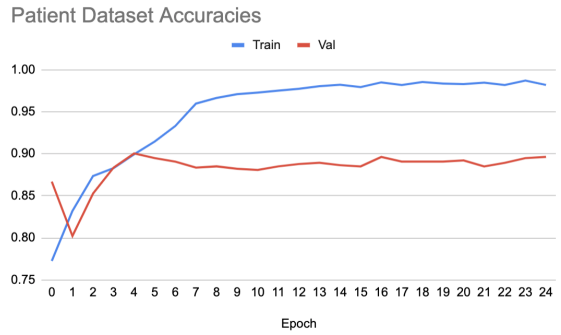

Fig. 5: Accuracies vs Epoch on Large Patient dataset with ResNet

Models	Internet Dataset	Patient Dataset
Augmented ResNet18	91.19%	90.40%
With Snapshot	93.08%	**91.66%**
With Anti-aliased & Contrasted	93.08%	90.04%
With Cycle-LR ResNet18	94.97%	90.10%
Cycle-LR Contrasted Augmented ResNeXt101_32x8d	**95.59%**	90.70%

TABLE I: Validation Accuracy for different models on binary classification

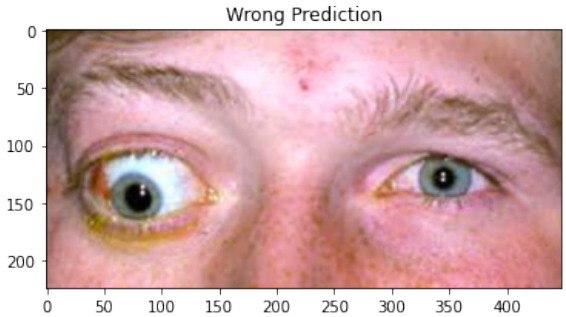

Fig. 6: A misclassified TED example. Notice the redness in the eye and the yellowness around in the eye