Enhancing Eye Disease Detection with MViT and SWD Loss

A BS thesis end semester report submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Data Science and Engineering

by

Alli Khadga Jyoth

Roll No.: 19024

Under the Supervision of

Dr. Akshay Agarwal

Department of Data Science and Engineering

Indian Institute of Science Education and Research Bhopal

Bhopal - 462066, India

April, 2023

Abstract

According to WHO, approximately 2.2 billion people worldwide suffer from eye disease. And the number is expected to rise with the increase in life expectancy, age-related problems, and screen-watching time. Several studies have been done to develop a deep-learning framework for the automated detection of eye diseases. In our study, we focus on using Knowledge Distillation to transfer the knowledge from a bigger, more complex teacher model (Resnet50) to a smaller student model(Resnet18). We also propose to use a Masked Vision Transformer autoencoder as a co-teacher for the distillation. The MViT encoder, trained to generate the latent space representations for masked images, is used as a co-teacher to make the student learn input image representations. We make use of the Cosine Similarity loss function, which compares the latent representations obtained from the student with the encoder's latent representations. We also propose Sample-Wise Distillation (SWD) loss, which is a per-sample weighting scheme for KL Divergence and Cross-entropy with target labels. The SWD loss takes into account the sample difficulty as measured by the student and teacher prediction probabilities and the similarity of student and teacher output distributions. The Resnet18 model trained with a co-teacher and normal distillation loss achieves the best distillation accuracy of 84.4%, and training with a co-teacher combined with SWD loss achieves 83.9% accuracy, which is a significant increase in performance compared to finetuned Resnet18 model achieving 76.6% accuracy and approaching the teacher Resnet50 accuracy of 87.5%.

Contents

List of Tables

Chapter 1

Introduction

The prevalence of eye diseases has motivated researchers to develop automated methods for early detection and diagnosis [1]. This study focuses on classifying eye diseases using deep learning techniques on the publicly available Ocular Disease Intelligent Recognition (ODIR) [2] dataset. The ODIR dataset contains 6,854 ophthalmic images, divided into eight classes: normal, cataract, glaucoma, macular degeneration, Diabetes, Hypertension, Pathological Myopia, and Other diseases/abnormalities [2].

Deep neural networks have demonstrated remarkable performance in many tasks, including image classification, object detection, and natural language processing. These models are frequently too expensive computationally to be used on devices with limited resources. This issue has been addressed using model acceleration and compression techniques like knowledge distillation.

Knowledge distillation is a method of transferring knowledge from a larger, more complex model (called teacher) to a much smaller model (called student). Knowledge distillation aims to improve the model performance of the student model while reducing the model complexity. Instead of directly mimicking the teacher's output (logits), knowledge distillation involves the student model to learn the output distribution of the teacher. This is done by minimizing the Kullback-Leibler (KL) divergence between the student and teacher output distributions (1.1).

$$KL(P||Q) = \sum_{x} P(x) \log(\frac{P(x)}{Q(x)})$$

$$\mathcal{L}_{KD} = T^2 D_{KL}\Big(\sigma\Big(\frac{z_t(X)}{T}\Big)\Big|\sigma\Big(\frac{z_s(X)}{T}\Big)\Big), \tag{1.1}$$

However, most existing methods use uniform weights for all samples in the distillation objective; this may not be optimal as some samples may be more difficult than others and may require more attention from the student model. In contrast, several sample-wise weighting schemes exist for the Cross entropy loss,

which assigns different weights to different samples based on their importance or difficulty (e.g., focal loss). Therefore we propose a novel per-sample weighting scheme for KL Divergence loss, which adapts the weight of each sample according to the perceived difficulty by the student and teacher. Our scheme aims to improve the performance and robustness of the student model on challenging samples and datasets.

Despite its effectiveness, the student model can only mimic the distribution of teachers' output but not completely learn the input data representations learned by the teacher. So to address this problem, we propose a novel method of using a Masked Vision Transformer as a co-teacher, which only focuses on teaching the input data representations to the student. This allows the student to learn more effectively and gain a deeper understanding of the input images on which it is trained.

The Masked Vision Transformer is a type of neural network that uses vision Transformer architecture and masked autoencoder technique to learn the input image representations [3]. A masked image is fed into the encoder part of the model, and the encoder outputs the latent space representations of the input masked image. Then, the decoder part of the model uses these representations to reconstruct the unmasked image. Since the input images are always masked and the vision transformer can attend to the whole image at once compared to typical CNN, the encoder part of the model can learn a more detailed representation of the input compared to other autoencoder models. These rich encoder representations are used for further downstream tasks.

We aim to make the student model more robust and accurate using the Masked Vision Transformer representations and per-sample weighted Distillation method. Since the accuracy of the model is critical for medical applications such as in the case of retinal Eye disease Classification. We test our methods on the ODIR dataset.

Chapter 2

Background

Eye diseases are becoming more prevalent due to age-related problems and increased life expectancy. According to WHO report, it is estimated that 2.2 billion people globally are affected by some eye disease [4, 5]. In recent years, Deep Learning based methods have achieved tremendous success in medical image analysis [6]. These methods show promising results in automated eye disease detection and classification [7, 8, 9, 10], including detection of Diabetic Retinopathy (DR) [8], age-related macular degeneration [11], and glaucoma [12].

Recent studies have used various pre-trained CNN architectures such as VGG, ResNet, and Efficient Net for eye disease classification [7, 8, 9, 13, 10, 14, 15]. Transfer Learning using the pre-trained models show promising results in accuracy and generalization. For example, a recent study used a VGG16 imagenet pre-trained model, which was finetuned on retinal images to classify the severity of Diabetic Retinopathy, and achieved 74.58% accuracy [16][7].In previous studies researchers in [17] employed the RFMiD dataset to develop a deep learning model named Eye-DeepNet to classify Normal, Diabetic Retinopathy (DR), Media Haze (MH) or Optic disc cupping (ODC) which achieved 76.04% testing accuracy.

Knowledge distillation [18, 19, 20] is a technique for transferring knowledge from a large teacher model to a smaller student model, which can improve the performance and efficiency of the student model. Knowledge distillation has been applied to various tasks in computer vision, including eye disease classification.

One example of using knowledge distillation for eye disease classification is the paper by Chelaramani et al. [21]. The paper proposes a multi-task learning (MTL) framework that combines knowledge distillation with three tasks related to eye disease prediction: coarse-grained disease classification, fine-grained disease sub-classification, and textual diagnosis generation. The paper also introduces a novel MTL-based teacher ensemble method for knowledge distillation, which uses multiple teachers trained on different subsets of tasks to guide the

learning of a single student model. which was published in Biomedical Signal

Another paper by He et al. [22] proposes a self-speculation method based on knowledge distillation for accurate ocular disease classification. The paper argues that existing methods for ocular disease classification do not fully utilize the clinical features that are important for diagnoses, such as optic disc, macula, and blood vessels. Therefore, the paper proposes a method that uses a teacher model to generate pseudo-clinical features for each fundus image and then uses a student model to learn from both the original image and the pseudo-clinical features.

In our study, we propose to use Knowledge Distillation to transfer the knowledge from two co-teachers, the Masked Vision Transformer is responsible for teaching effective input data representations and the regular CNN teacher is responsible for teaching the output logit distributions to the student model. We use the ODIR dataset to detect the presence of eye diseases. The Ocular Disease Intelligent Recognition (ODIR) dataset contains images of the left and right fundus from 5000 patients. Each image is classified into 8 labels. The dataset is divided into $\approx$ 7000 images in the training set and 1000 images in the testing set.

Chapter 3

Research Gaps

Eye diseases are becoming more prominent in recent times. Various studies try to automate the detection of eye diseases from fundus images. But most, if not all, of the studies focus on developing deep learning models for detecting one or a few of the popular diseases like Diabetic Retinopathy, cataracts, etc. But they ignore other diseases like Hypertension, Media Haze, etc., due to either lack of data or other reasons. In our study, we aimed to develop a generalized Deep Learning model to detect various types of abnormalities that might be present. We focus on classifying as many diseases as possible by employing the ODIR dataset.

Masked Autoencoders are efficient representation learners, and masked vision transformers (MViT) are more effective learners. But existing work only focuses on using the representations learned by the MViTs for downstream tasks such as image classification. But none focuses on transferring the learned representations to other models for improved performance. This work focuses on transferring the MViT encoder latent representations to the student model.

With Medical images such as Eye Disease datasets, the input data is much more complex compared to other datasets. Hence, the MViT is perfect for learning the input data latent representations since the Masked Vision transformer works by masking the input image and then reconstructs the masked portion of the image by using the latent space representations from the encoder. This suggests that the encoder part of the MViT is able to effectively learn the spatial and other fine-grained details of the input. These complex input representations are well-suited for downstream tasks such as image classification. But training an MViT and deploying the MViT is expensive in terms of time and space. So we use Knowledge distillation which is able to reduce the complexity of the model without much loss in performance.

And with Knowledge Distillation, we use KL Divergence Loss, which compares the divergence of a probability distribution to a target distribution. The

KL Loss between student and teacher output logits makes the student logit distribution similar to the teacher logit distribution. But the distillation weight is uniform for all the samples; this is not optimal. Since there may be samples that have higher difficulty compared to other samples, then one might want the model to focus more on the difficult samples rather than the easy samples. This is not possible with existing KL Divergence, so we develop a novel Sample wise weighting scheme for KL Divergence loss which takes into account the prediction probabilities of both teacher and student for a particular sample and assigns the divergence weight accordingly.

Chapter 4

Objectives

The objectives of the study can be summarized as follows:

- **Data Acquisition and Processing:** For the study, we need to collect the retinal fundus images belonging to different categories. The ODIR is available with open-source licenses and can be downloaded for free. The dataset contains images that belong to multiple classes, and the dataset is imbalanced in nature. These are solved in the preprocessing and Data Augmentation steps along the process.

- **Finetuning of Models:** For finetuning, we use an imagenet pretrained VGG16, Resnet50 and Densenet and ShuffleNet models. The models are trained on the dataset with a low learning rate. A low learning rate allows the model to learn without overfitting or getting into vanishing gradient problems.

- **Knowledge Distillaton** We proposed to use knowledge distillation for the efficient transfer of Knowledge from a teacher model to a much smaller student model. It involves a capable teacher model to be present, which in our case is the finetuned pretrained Model. And for the student model, we are using Resnet18, which is half the size of the Resnet50 model. And along with the Resnet50 teacher model, we propose to use a co-teacher MViT for the efficient transfer of input representations.

  - **Maksed Vision Transformer** We proposed to use Masked Vision Transformer as a co-teacher. The MViT is trained on the imagenet dataset and is finetuned further on our dataset to improve input data representations. After finetuning the encoder and decoder blocks of MViT, we discard the decoder block of MViT, whose task is to reconstruct the input image from the masked image representations pro-

duced by the encoder block. We use the encoder block of the MViT as our latent feature generator for our input data.

– **Sample-wise weighted Loss** We propose to use a new sample-wise weighting scheme for KL divergence and Cross-entropy loss functions. The weighting scheme considers the student prediction confidence and Teacher prediction confidence to assign weights to both the Student-hard target Cross entropy loss and Student-Teacher KL divergence loss. The loss functions and the weighting scheme are defined in further sections.

Chapter 5

Methodology

The proposed Artificial Intelligence framework for the study is divided into 3 sections: data acquisition, data augmentation and preprocessing, and last, training the models. The first section talks about the features of the data and the conversion of multi-labeled data to single-labeled data. It also talks about the selection of classes. The second section details the data imbalance present in the data and how data augmentation with rotation, flipping, etc., is used to overcome the problem. It also gives more details on the preprocessing of the data such as cropping and resizing. In the third section, we discuss the training of the models. We detail the model architecture used to detect different diseases from the retinal fundus images.

5.1 Data Description

In our study, we use an open-source Ocular Disease Intelligent Recognition (ODIR) Dataset and a multi-labeled Retinal Fundus Multi-Disease Image Dataset (RFMiD).

5.1.1 ODIR dataset

ODIR-2019 is a public dataset containing the structured ophthalmic database of 5,000 patients with age, color fundus photographs from the left and right eyes, and doctors' diagnostic keywords from doctors. This dataset is meant to represent a "real-life" set of patient information collected by Shanggong Medical Technology Co., Ltd. from different hospitals/medical centers in China. In these institutions, fundus images are captured by various cameras in the market, such as Canon, Zeiss, and Kowa, resulting in varied image resolutions.

The dataset contains an annotation file where the images are tagged with the corrected words which define each class. Specialists score patients with 8

different labels that identify the following pathologies: N, D, G, C, A, H, M, and O. The different categories are:

- Normal (N),
- Diabetes (D),
- Glaucoma (G),
- Cataract (C),
- Age related Macular Degeneration (A),
- Hypertension (H),
- Pathological Myopia (M),
- Other diseases/abnormalities (O)

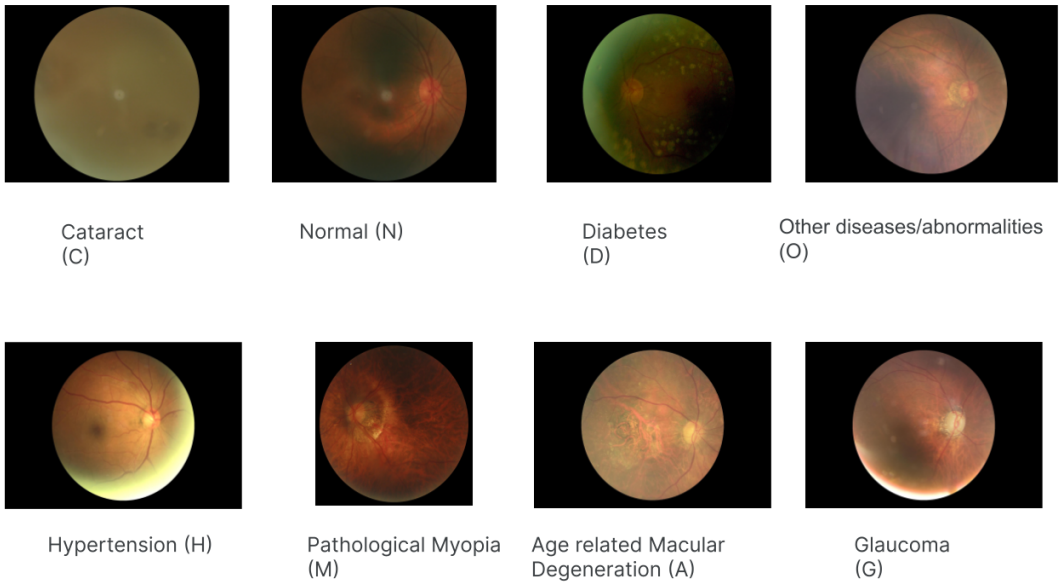


Figure 5.1: ODIR data samples

Each patient may be marked with one or more labels indicating that they may suffer from more than one disease. The dataset also contains basic information such as the patient's sex, age, the labels of their pathology, and some keywords that best define them. We have a total of 7000 images, 3500 left eye and 3500 right eye images for the training set and 1000 images for the validation set. The validation set doesn't have ground truth labels. The data distribution of the ODIR dataset can be seen in fig 5.2. Like the RFMiD data, we add the images of the RFMiD dataset to the "O" class of the ODIR dataset. The images added belong to the uncommon classes between the two datasets. The image distribution after adding the RFMiD images is seen in fig 5.3

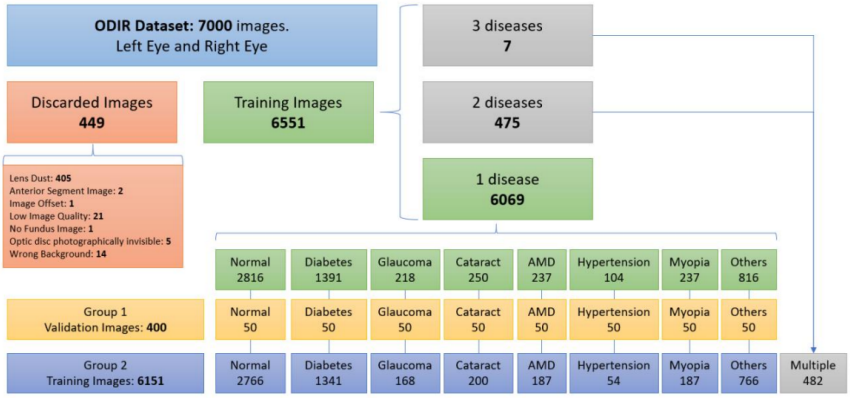


Figure 5.2: ODIR Original data Distribution

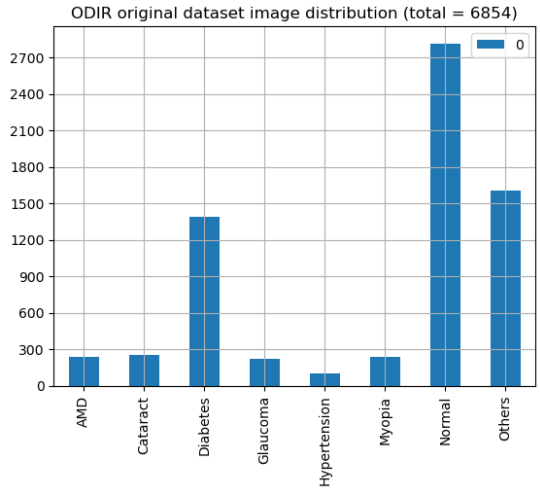


Figure 5.3: ODIR data Distribution before Augmentation

5.2 Dataset augmentation and pre-processing

After acquiring the data and initial processing of the dataset by removing the multi-label images and removing classes with very few images, we have a workable dataset. We then add the images from the RFMiD dataset into the "Other" class of the ODIR dataset to make the dataset richer, now after seeing the data distribution of the dataset from the figures 5.3, we see the class imbalance issue persists in the dataset. The class imbalance, if not taken care will lead to the misclassification of models.

The class imbalance issue is avoided by using data augmentation techniques. The data augmentation methods were selected based on different variations of fundus images that are present in the real world. For our study, we include different geometric transformations, i.e., right 8° rotation, left 8° rotation, etc. The different augmentations applied can be seen in figure 5.4. After the augmentations were applied, the total number of images increased significantly, and the new data distribution is seen from figs 5.5a. But since the augmentations

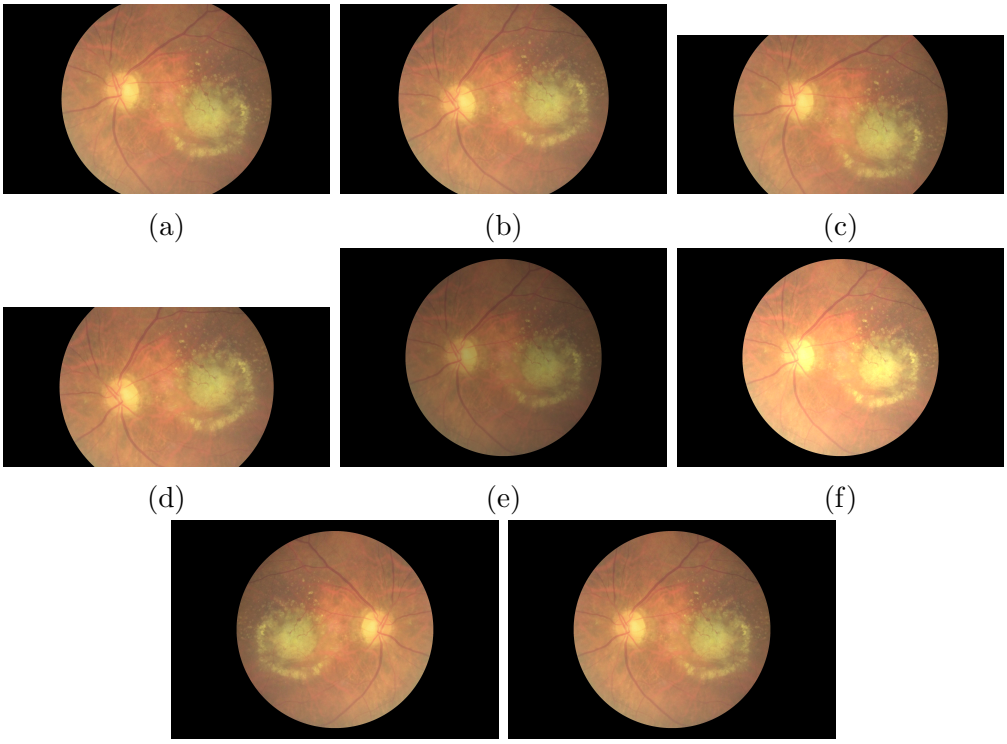


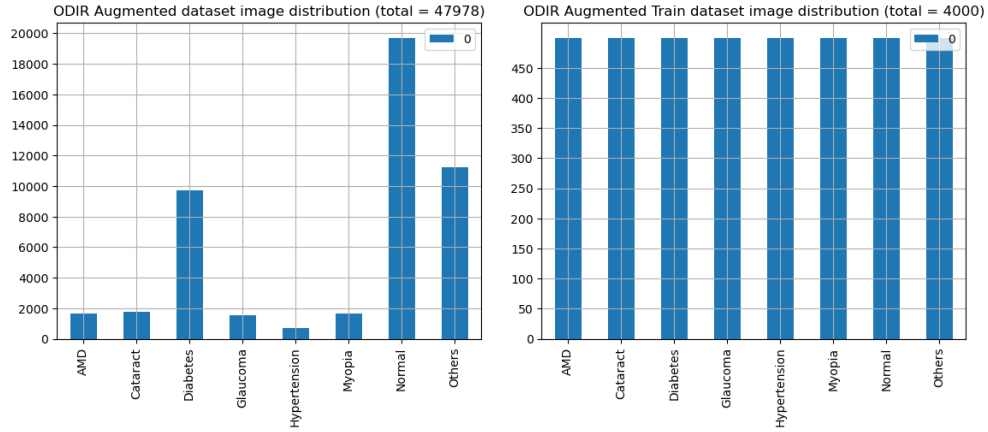
Figure 5.4: (**a**). rotation 8° right, (**b**). rotation 8° Left, (**c**). rotation 15° Right, (**d**). rotation 15° right, (**e**). Brightness −30%, (**f**). Brightness +30%, (**g**). Horizontal Flip, (**h**). Original Image,

only scale the image distribution, the class imbalance is still present. To solve this, we resample some subset of images from the augmented images. For the ODIR dataset, we sample 500 images for each class in the training set and 200 images per class for the testing set. The distribution of images in training set after resampling can be seen in figure 5.5b. These resampled images are further processed to remove the black pixels present at the edges.

5.3 Masked Vision Transformer as Co-Teacher

we propose a knowledge distillation method that uses a masked vision transformer (MViT) as a co-teacher to guide a student model for image classification. MViT is a type of vision transformer (ViT) that uses masked autoencoders (MAEs) as a self-supervised learning technique. MAEs are models that learn to fill in the missing parts of input images by using latent features from an encoder. MViT can learn rich and complex features from visual data without using any labels and can transfer them to the student model efficiently.

We use MViT as a co-teacher because it can provide supervision on the input representation, which is complementary to the output logit distribution



(a) ODIR data Distribution after augmentation

(b) ODIR data Distribution after augmentation and resampling

Figure 5.5: ODIR data Distribution

supervision provided by another teacher model. We use ResNet-50 as the other teacher model, which is a regular fine-tuned model trained on a labeled dataset using supervised learning. We use a smaller and simpler Resnet18 as the student model, which we aim to improve by distilling knowledge from both teachers.

MViTs are based on ViTs, which are models that use self-attention to process visual data as a sequence of patches. ViTs have shown impressive results on image classification and other vision tasks, but they require a large amount of labeled data for pre-training. To overcome this limitation, MAEs are proposed as a data-efficient self-supervised learning method for ViTs. MAEs randomly mask out some patches of the input image and use an encoder-decoder architecture to reconstruct them. The encoder takes only the visible patches as input and produces a latent representation that captures the global context of the image. The decoder takes the latent representation and the mask tokens as input and generates the missing patches. The model is trained by minimizing the reconstruction loss between the original and the generated patches.

Our method consists of three steps: pre-training, fine-tuning, and distillation. In the pre-training step, we finetune an imagenet pre-trained MViT model on our dataset using MAEs. We randomly mask out some patches of the input image and use an encoder-decoder architecture to reconstruct them. The encoder produces a latent representation that captures the global context of the image. The decoder generates the missing patches based on the latent representation and the mask tokens. We train the MViT model by minimizing the reconstruction loss between the original and the generated patches.

In the fine-tuning step, we train an imagenet pretrained ResNet-50 model on a labeled dataset using supervised learning. This model serves as the first teacher

for knowledge distillation. In the distillation step, we train a student model on the same labeled dataset using two loss functions: one from each teacher. We use two strategies to transfer knowledge from the teachers to the student: output alignment and representation alignment.

Output alignment aims to make the student's output logit distribution similar to the ResNet-50's output logit distribution. We use KL divergence as the loss function for this objective. KL divergence measures the difference between two probability distributions and encourages the student to mimic the ResNet-50's predictions.

Representation alignment (RA) aims to make the student's input representation similar to the MViT's input representation. We use cosine similarity as the loss function for this objective. Cosine similarity measures the angle between two vectors and encourages the student to learn features that are aligned with the MViT's features. To compute this loss, we take the last convolutional layer outputs from the student model and pool them to match the size of the MViT's encoder outputs. Then we compare them with cosine similarity.

By combining these two loss functions, we can leverage both teachers' knowledge and improve the student's performance. Our method can reduce the complexity of the model without much loss in accuracy and can also benefit from MViT's self-supervised learning ability.

5.4 Sample-wise weighted Distillation Loss

Sample-wise weighted distillation loss assigns different weights to different samples based on their difficulty and importance for the student model. This way, the student model can focus more on the samples that are harder to learn or more relevant to the task and less on the samples that are easier to learn or less relevant to the task. Sample-wise weighted distillation loss can also reduce the noise and bias in the teacher model's predictions and make the student model more robust and accurate. The typical Distillation objective is written as,

$$\mathcal{L}_{distill} = \lambda \cdot H_{CE}(y, S(X)) + (1-\lambda) \cdot D_{KL}\left(\sigma\left(\frac{z_s(X)}{T}\right) | \sigma\left(\frac{z_t(X)}{T}\right)\right) \quad (5.1)$$

where H_{CE} represents the cross entropy between the true label y and the student network prediction $S(X)$ for a given input X, D_{KL} is the KL divergence between the teacher and student predictions softened using the temperature parameter T, $z(X)$ is the network output before the softmax layer (logits), and $\sigma(.)$ indicates the softmax function. The term λ in the above equation is a hyperparameter that controls the contribution from cross entropy and Divergence loss.

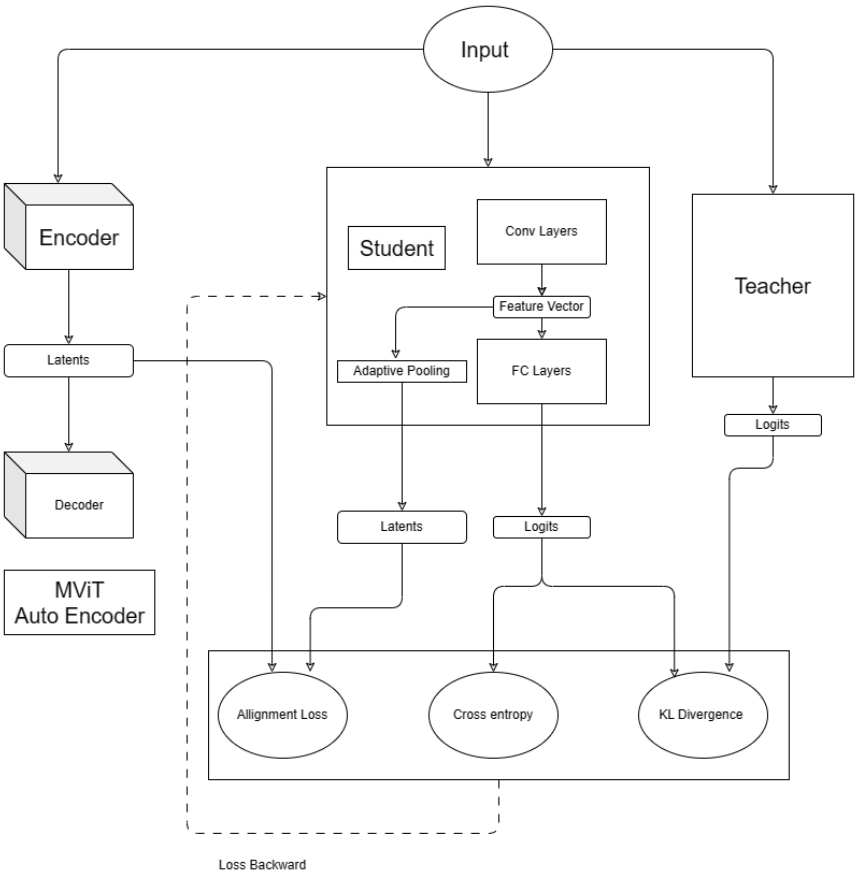


Figure 5.6: Model Architecture

This Distillation objective assigns uniform weights to all the samples, which may not be optimal. Since there might exist which are difficult for the student, then we may want our student to focus more on such samples than focus on easy samples. So we propose our own Sample-Wise Distillation (SWD) Loss which takes into account the difficulty as perceived by the student and assigns a suitable weight for both Cross entropy and the KL Divergence terms.

First, we calculate the factor term as:

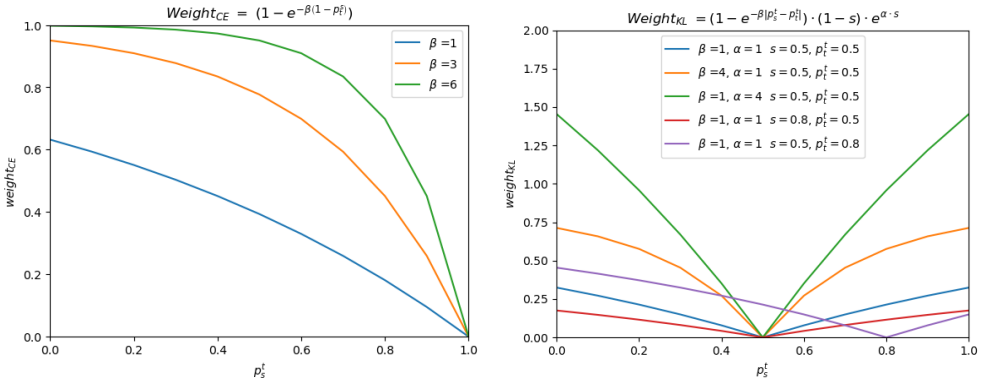
$$factor = \exp\left(-\beta\left(1 - p_t^s\right)\right)$$

where p_t^s is the target class prediction probability of the student, and β is a hyper parameter.

Then, we compute the logit similarity between the student and teacher by first normalizing their output logits and then taking the cosine similarity between them:

$$similarity = \frac{\mathbf{y_s}^\top \mathbf{y_t}}{\|\mathbf{y_s}\|_2 \|\mathbf{y_t}\|_2}$$

where $\mathbf{y_s}$ and $\mathbf{y_t}$ are the L2-normalized logits of the student and teacher



(a) Cross-entropy Weight function (b) KL Divergence Weight function

Figure 5.7: SWD Loss function

models, respectively.

Finally we calculate the weight for the KL Divergence as

$$weight_{kl} = (1 - e^{-\beta|p_i^s - p_i^t|}) \cdot (1 - similarity) \cdot e^{(\alpha \cdot similarity)}$$

where p_t^t is the target class prediction probability of the teacher and α is a hyperparameter specific to $weight_{kl}$

Therefore the final Sample-Wise Distillation Loss becomes;

$$\mathcal{L}_{SWD} = (1 - factor) \cdot H_{CE}(y, S(X)) + weight_{kl} \cdot D_{KL}\left(\sigma\left(\frac{z_s(X)}{T}\right) | \sigma\left(\frac{z_t(X)}{T}\right)\right) \tag{5.2}$$

5.5 Model Training

Model training is divided into 3 stages;

5.5.1 Pre-training MViT

In the first stage, we take a pretrained Masked Vision Transformer autoencoder trained on the Imagenet dataset and then finetune the autoencoder-decoder on our dataset. The finetuning is done for 50 epochs at a base learning rate of $5e - 6$ with a batch size of 64. We use MSE loss as the reconstruction loss between the decoder decoded and input images. The mask ratio is 0.75 and is fixed throughout the training. After training, the decoder part of the masked autoencoder (MAE) is discarded, and only the encoder is kept to generate the input data representations.

5.5.2 Fine-tuning of Teacher Models

After the pretraining of the MAE, we finetune the pretrained Resnet50 and other models on the dataset. The other models include Densenet121, Densenet201, Resnet18, ShuffleNetv2_X1_0, ShuffleNetv2_X2_0, VGG16 and WideResNet50_2. We train these models for 50 epochs at a learning rate of 0.001 with an lr decay of 01 at epoch 30. Table 5.1 shows the fine-tuning accuracies.

Table 5.1: Fine tune model Accuracy

Models	Accuracy
Resnet18	76.6
VGG16	87.5
Resnet50	87.5
Densenet121	86.8
Densenet201	86.8
ShuV2_x1_0	65.5
ShuV2_x2_0	84.8
WRN_50_2	56.0
MViT encoder + classifier block	81.5

Chapter 6

Results and Discussion

First, the teacher models were finetuned for 50 epochs at a learning rate of 0.001 with a decay factor of 0.1 at epoch 30. Table 5.1 showcases the best accuracy achieved by the fine-tuned models. After the fine-tuning, we selected the top-performing model Resnet50 as our teacher model and Resnet 18 as our student model and conducted further experiments using this pair. For knowledge Distillation, we train each model for 240 epochs at a base learning rate of 0.01 with a decay factor of 0.1 with decay epochs at 150,180 and 210. For DKD we train with β parameter set to 2, and for our SWD loss we set α and β to 1.0. The results are presented in table 6.1, where RA represents the Representation alignment loss taken wrt. to the input latent representations obtained from the MAE encoder.

Table 6.1: Knowledge distillation using different Methods on ODIR Dataset

Method	**Accuracy**	**F-Score**	**Val Loss**
Finetune Resnet50 (Teacher)	87.5	87.31	0.46
Finetune Resnet18 (Student)	76.6	75.53	0.77
Hinton KD	83.2	82.83	0.73
DKD	83.7	83.6	0.63
RA_{CRD}[1]+ KD	83.9	83.53	0.64
SWD Loss (Ours)	82.6	82.33	0.65
RA_{Cos}+ KD	**84.4**	84.13	0.68
RA_{Cos}[2]+ SWD Loss (Ours)	83.9	83.54	0.63

From the table, we can conclude three things, one. Knowledge distillation shows a massive improvement in student accuracy across the board. This is expected due to the fact that it is easier for the student to learn from a teacher

[1] RA_{CRD} represents the Representation alignment from the MAE encoder using Contrastive Distillation loss instead of Cosine Similarity loss.

[2] RA_{Cos} represents the Representation alignment from the MAE using Cosine Similarity loss.

rather than learning from the complex input data directly. Second. Using a Masked autoencoder as a co-teacher for learning the input data representations benefits the student and makes the student perform even better. This should also be evident from the fact that Maksed autoencoders are very good at latent space representations. So the student model learning from the MAE is able to better learn to represent the input data. And Third. the SWD loss performs worse than the typical Hinton KD, but from the validation loss, we see that the model is much more confident in its predictions compared to other methods.

Chapter 7

Conclusion

Different models were trained on Retinal Eye disease images, extracted from ODIR using the different methods proposed in the study. From table 6.1 we conclude that our usage of MViT encoder as the co-teacher for teaching the input representations showed significant improvements in student performance. And our sample-wise weighting scheme for the distillation objective did not show performance improvement over the normal KD, but when combined with the co-teacher loss, it performed better than DKD, a state-of-the-art method in KD, while having a similar validation loss. This shows that our model is much more confident in its predictions. In the case of medical image analysis, we require the model to be confident in its predictions to a certain degree and over that the model becomes overconfident. But from our results, we conclude that our model using the SWD loss and alignment loss is able to perform on par or beat all the typical methods while being confident in its predictions.

Chapter 8

Work Plan

The work plan for the study involves

- **Dataset:** The previous studies focussed on building deep learning models specific to a single disease or for a few diseases. Our work focuses on building a generalized deep-learning model for eye diseases. So this requires a large dataset containing a wide variety of eye diseases. So for our study, we are using the publicly available ODIR dataset which contains Retinal Eye images classified into 8 different categories. After the collection of data, the next tasks are
  - Data Preprocessing
  - Data Augmentation
  - Data Resampling
- **Finetuning:** The models which are to be finetuned are selected, and the imagenet pretrained models are finetuned for 50 epochs with a low learning rate. Then we select a suitable teacher and student model for further experiments.
- **Masked Vision Transformer:** We propose to use an MViT encoder as a distillation co-teacher, so we require a well-trained MViT MAE; for this, we take the imagenet pretrained MViT MAE, then finetune the MAE on our dataset for 50 epochs.
- **Knowledge Distillation:** It is a process of transferring knowledge from a bigger teacher model to a smaller student model. It is already being used to improve the model performance in classification tasks in other domains. Knowledge Distillation for Eye Disease classification is not fully explored, and in our work, we are using two distillation methods. In one method

we are using our proposed Sample-Wise Distillation Loss, which is a per-sample weighted loss function. On the other hand, we are using the MViT co-teacher to train the student to learn input data representations more efficiently and effectively. Though these two methods are complementary to each other, we compare the performance of our distillation methods using these two methods separately and in combination with each other.

Bibliography

[1] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[2] Odir-2019 - grand challenge.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[4] Rupert RA Bourne, Seth R Flaxman, Tasanee Braithwaite, Maria V Cicinelli, Aditi Das, Jost B Jonas, Jill Keeffe, John H Kempen, Janet Leasher, Hans Limburg, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health*, 5(9):e888–e897, 2017.

[5] Vision impairment and blindness, Oct 2022.

[6] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[7] Yifan Peng, Shazia Dharssi, Qingyu Chen, Tiarnan D Keenan, Elvira Agrón, Wai T Wong, Emily Y Chew, and Zhiyong Lu. Deepseenet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, 126(4):565–575, 2019.

[8] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep

learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[9] Ramachandran Rajalakshmi, Radhakrishnan Subashini, Ranjit Mohan Anjana, and Viswanathan Mohan. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye*, 32(6):1138–1144, 2018.

[10] Zhixi Li, Stuart Keel, Chi Liu, Yifan He, Wei Meng, Jane Scheetz, Pei Ying Lee, Jonathan Shaw, Daniel Ting, Tien Yin Wong, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes care*, 41(12):2509–2516, 2018.

[11] Philippe M Burlina, Neil Joshi, Michael Pekala, Katia D Pacheco, David E Freund, and Neil M Bressler. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA ophthalmology*, 135(11):1170–1176, 2017.

[12] Sidong Liu, Stuart L Graham, Angela Schulz, Michael Kalloniatis, Barbara Zangerl, Weidong Cai, Yang Gao, Brian Chua, Hemamalini Arvind, John Grigg, et al. A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs. *Ophthalmology Glaucoma*, 1(1):15–22, 2018.

[13] Philippe Burlina, Katia D Pacheco, Neil Joshi, David E Freund, and Neil M Bressler. Comparing humans and deep learning performance for grading amd: a study in using universal deep features and transfer learning for automated amd analysis. *Computers in biology and medicine*, 82:80–86, 2017.

[14] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, 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. *Jama*, 318(22):2211–2223, 2017.

[15] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[16] Abhishek Deshpande and Jatin Pardhi. Automated detection of diabetic retinopathy using vgg-16 architecture. *Irjet*, 8(03), 2021.

[17] Neha Sengar, Rakesh Chandra Joshi, Malay Kishore Dutta, and Radim Burget. Eyedeep-net: a multi-class diagnosis of retinal diseases using deep neural network. *Neural Computing and Applications*, pages 1–21, 2023.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[19] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[20] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[21] Sahil Chelaramani, Manish Gupta, Vipul Agarwal, Prashant Gupta, and Ranya Habash. Multi-task knowledge distillation for eye disease prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3983–3993, 2021.

[22] Junjun He, Cheng Li, Jin Ye, Yu Qiao, and Lixu Gu. Self-speculation of clinical features based on knowledge distillation for accurate ocular disease classification. *Biomedical Signal Processing and Control*, 67:102491, 2021.