This WACV 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-Task Knowledge Distillation for Eye Disease Prediction

Sahil Chelaramani[1], Manish Gupta[1], Vipul Agarwal[1], Prashant Gupta[1], Ranya Habash[2]
[1]Microsoft, [2]Bascom Palmer Eye Institute
{sachelar,gmanish,vagarw, prgup}@microsoft.com, rgh4@med.miami.edu

Abstract

While accurate disease prediction from retinal fundus images is critical, collecting large amounts of high quality labeled training data to build such supervised models is difficult. Deep learning classifiers have led to high accuracy results across a wide variety of medical imaging problems, but they need large amounts of labeled data. Given a fundus image, we aim to evaluate various solutions for learning deep neural classifiers using small labeled data for three tasks related to eye disease prediction: (T_1) predicting one of the five broad categories – diabetic retinopathy, age-related macular degeneration, glaucoma, melanoma and normal, (T_2) predicting one of the 320 fine-grained disease sub-categories, (T_3) generating a textual diagnosis. The problem is challenging because of small data size, need for predictions across multiple tasks, handling image variations, and large number of hyper-parameter choices. Modeling the problem under a multi-task learning (MTL) setup, we investigate the contributions of each of the proposed tasks while dealing with a small amount of labeled data. Further, we suggest a novel MTL-based teacher ensemble method for knowledge distillation. On a dataset of 7212 labeled and 35854 unlabeled images across 3502 patients, our technique obtains ∼83% accuracy, ∼75% top-5 accuracy and ∼48 BLEU for tasks T_1, T_2 and T_3 respectively. Even with 15% training data, our method outperforms baselines by 8.1, 3.2 and 11.2 points for the three tasks respectively.

1. Introduction

Eye diseases significantly impact the quality of life for patients. Four of such diseases are glaucoma, diabetic retinopathy (DR), age-related macular degeneration (AMD), and uveal melanoma. Glaucoma incidence worldwide is 64.3M in 2013 and projected to 76M in 2020 [1]. The disease affects ∼2M people in USA[1]. In 2015, ∼415M people were living with diabetes, of which ∼145M have some of DR[2]. Macular degeneration incidence worldwide is 196M in 2017 and projected to 288M by 2040 [2]. Although it is a relatively rare disease, uveal melanoma is the most common primary intraocular tumor in adults with a mean age-adjusted incidence of 5.1 cases per million per year[3].

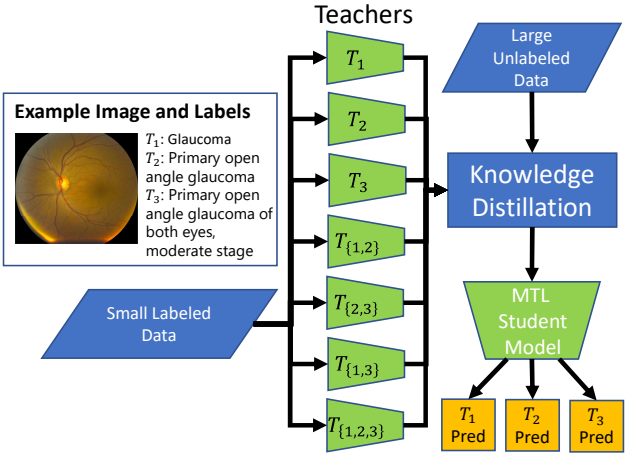

Figure 1: Multi-Task Learning (MTL) combined with Multi-Teacher Knowledge Distillation (KD) for eye disease prediction. Each of the seven teachers is trained on a subset of three tasks. KD is then used to distill "dark knowledge" from all the teachers to the MTL student using both labeled as well as unlabeled instances making our model robust in small labelled data scenarios.

Early diagnosis of these eye diseases can help in effective treatment or at least in avoiding further progression of these diseases. Limited availability of ophthalmologists, lack of awareness and consultation expenses restrict early diagnosis. Hence, automated screening is critical. Recently, there has been a significant focus on applying deep learning techniques for medical imaging. However, deep learning classifiers are known to require large amounts of labeled train-

[1]https://doi.org/10.1016/j.pop.2015.05.008, https://www.ncbi.nlm.nih.gov/pubmed/20711029

[2]http://dx.doi.org/10.1016/S2214-109X(17)30393-5

[3]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5306463/

ing data. Gathering labeled data for retinal imaging at a large scale is difficult because of the extensive human labeling effort, especially by qualified ophthalmologists. Hence, we explore different ways of obtaining high accuracy for disease prediction using small amounts of labeled retinal fundus images. Specifically, we explore three critical tasks namely coarse-grained disease classification, fine-grained disease classification, and detailed disease diagnosis generation. Given small labeled dataset L and large unlabeled dataset U of fundus images, we aim to train a classifier such that given a new fundus image, can predict with high accuracy, one of the following classes: DR, AMD, glaucoma, melanoma, normal, One of the 320 fine-grained sub-disease classes and a detailed textual diagnosis similar to that provided by ophthalmologist.

Disease prediction given fundus images is very challenging because of the following reasons: (1) Gathering large amount of labeled fundus data is difficult. (2) Images have a lot of heterogeneity because of use of different lighting conditions, different devices, and different kinds of fundus images (like single field of view versus montage images), and the field of view captured. (3) Sometimes, minor changes in images are indicative of particular diseases. (4) Presence of artifacts like specular reflection, periphery haze, dust particle marks, fingerprints, blurred images, incorrectly stitched montage makes modeling complicated. (5) Absence of large corpus to learn robust embeddings for rare diagnosis words. (6) Images vary wrt. macula position – center, nasal, inferior, superior. Some images even have disc/macula cut out.

The three proposed tasks are very related; hence we model them using a deep learning based multi-task learning (MTL). Further, in presence of small labeled data, we augment the MTL setup with Knowledge Distillation (KD). In KD, a student model is trained to mimic a teacher model. Fig. 1 provides a conceptual illustration of the proposed approach. Given small amounts of labeled data for the three tasks, we train seven teacher models, one each for a subset of the three tasks. Thus, four of these seven teacher models are multi-task in nature. Further, we learn an MTL student model by distilling the "dark knowledge" from the seven teachers using both small-labeled as well as large-unlabeled data.

Fig. 2 illustrates the detailed architecture of our proposed approach. The MTL teacher model M_1 jointly extracts image description relevant to the three tasks using ImageNet-pretrained ResNet-50 [3], a deep convolutional neural network (CNN), fine-tuned on labeled data L. M_1 is trained end-to-end using gradient descent with a linear combination of cross entropy loss across the three tasks. A complex model with small data can lead to overfitting which can be avoided using regularization. Hence, we perform a two-stage knowledge distillation (KD) [4]. In the first stage, model M_1 is finetuned to obtain model M_2 such that the cross entropy loss on labeled data is small and KL divergence between M_2's output distributions and M_1's output distributions across the three tasks is also minimized. In the second stage, to harness large unlabeled data U, under a semi-supervised setup, we further finetune M_2 to obtain M_3 by minimizing KL divergence between M_2's output distributions and M_3's output distributions across the three tasks.

We extensively experiment with multiple design choices: with and without MTL, with and without KD, with varying temperature for KD, with teacher ensemble and also with varying training data size. We are the only work focusing on modelling multiple eye diseases with the same model, leading to lack of good baselines. Our best method provides an accuracy of 82.52% for task 1, 51.11 (top-1)/75.19 (top-5) for task 2 and 48.1 BLEU for task 3, when 70% of the labeled data is used for training, while demonstrating gains for each task of 8.1, 3.2 and 11.2 points respectively, using only 15% of the labeled data. Our results show the effectiveness of the combined proposed MTL+KD architecture for training retinal disease prediction models from small labeled data.

Overall, we make the following contributions:

- We explore retinal disease classification task using small labeled training data.
- We propose a learning algorithm using MTL and multi-teacher KD on different training data sizes. To the best of our knowledge, this is the first work on using KD in MTL scenario for retinal images.
- Using a dataset of 7212 labeled and 35854 unlabeled fundus images for 3502 patients from 2004 to 2017, we show the effectiveness of the proposed methods. Using only 15% of the labeled data (1082 images), using MTL+KD, we can perform comparably to single task models trained with 70% data.

The paper is organized as follows. We discuss related work on machine learning for eyecare, deep learning for medical imaging, MTL and deep learning with small data in Section 2. In Section 3, we discuss details of the proposed methods. In Section 4, we present dataset details and insights from analysis of results. We conclude with a summary in Section 5.

2. Related Work

Machine Learning for Eyecare. Work on applying predictive analytics for eyecare includes the following: prediction of post-operative surgery outcomes [5, 6, 7], disease prediction (glaucoma [8, 9], AMD [10, 11], DR [12]), segmentation of eye parts and anomalies (blood vessels [13], retina exudates, microaneurysms, drusen, cotton-wool spots), and

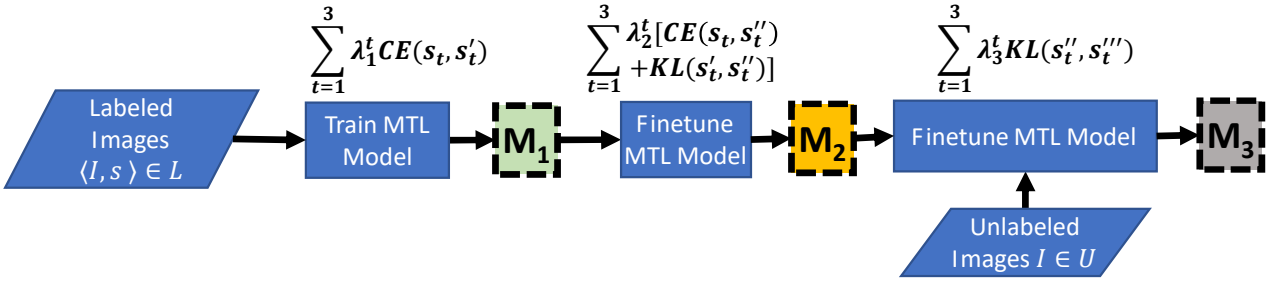

Figure 2: Our model's training phases are depicted in the figure. The loss functions used in each phase are shown above the corresponding stages. CE stands for cross-entropy while KL stands for KL-divergence.

predicting if eye tumor will metastasize [14]. We focus on the critical problem of screening patients concerning four most important eye diseases using fundus images.

Deep Learning for Medical Imaging. Motivated by immense success of deep learning techniques in general vision, speech as well as text problems, there has been a lot of focus on applying deep learning for medical imaging recently [15, 16]. Specifically, deep learning techniques have been applied to medical image data for neuro [17], retinal [12], pulmonary [18], digital pathology [19], breast [20], cardiac [21], abdominal [22], musculoskeletal [23] areas. Specifically, problem areas include image quality analysis [24], image segmentation [25], image/exam classification [26], object/lesion classification [27], and registration (i.e. spatial alignment) of medical images [28], image enhancement, and image reconstruction [29]. Although another work has recently focused on image classification and caption generation from retinal fundus images [30] with *noisy labels*, we build models with *clean but small labeled data*.

Multi-task Learning. MTL has been used successfully across all applications of machine learning, from natural language processing [31] and speech recognition [32] to computer vision [33] and drug discovery [34]. MTL can be done with soft [35] or hard parameter sharing [36]. Further, recently, there has been some work [37, 38] on performing knowledge distillation in the context of MTL for Transformer-based architectures on the GLUE [39] which is a set of NLP (Natural Language Processing) tasks. In this paper, we perform hard parameter sharing based MTL across three eye diagnosis tasks.

Deep Learning with Small Labeled Data. Previous works have suggested various ways to handle small labeled data. Pre-training using transfer learning [40] from models trained on similar tasks with rich data, is a common approach. Traditionally, self training [41] has been the most popular semi-supervised approach to handle lack of enough labeled data. But self training requires careful threshold tuning for throttling to avoid label noise. Recently, KD [42, 43] has been proposed as an approach for model compression but we observe that it can also be used as a semi-supervised technique [44]. In this paper we explore a combination of multitask learning, distillation, and their combined efficacy when applied to small labelled data.

3. Approach

In this section, we explain the setup of the tasks using deep learning architectures. Next, we describe our proposed MTL architecture. Lastly, we discuss KD with MTL.

3.1. Task Description

For each disease, fundus images show symptoms (Table 1). We attempt to diagnose such symptoms using deep learning methods. Early detection is critical to avoid further loss. We model task T_1 as a multi-class classification of with five classes (DR, AMD, melanoma, glaucoma and normal). Task T_2 is modeled as a 320-class classification problem, where the broad five classes have been divided further based on disease sub-types, disease severity, retinal regions and other important symptoms. Tasks T_1 and T_2 are modeled using ResNet-50 architecture. We initialize model weights using ImageNet pre-training to get a 1024D representation of every fundus image. ResNet output is connected to two dense layers with ReLU and softmax activations respectively. Last layer is of size 5 for T_1 and size 320 for T_2). We use cross entropy loss for both the tasks.

We model task T_3 (diagnosis generation) as an image captioning problem, where given a fundus image, we output the detailed disease diagnosis. The diagnosis generation model consists of a single layered LSTM, which takes the features generated by the CNN encoder (projected to a suitable latent representation using a dense layer) along with the "start-of-sentence (SOS)" token, and trains a language model conditioned on it. The hidden state of each timestep is projected to a vocabulary sized vector, on which softmax activation is applied. This vector is then sampled to generate the corresponding output word from the vocabulary.

Disease	Description	Treatment on early detection	Symptoms in fundus images
DR	damage to the retina due to diabetes mellitus	laser surgery, injection of corticosteroids or anti-VEGF agents	microaneurysms, exudates, cotton wool spots, flame hemorrhages, dot-blot hemorrhages
AMD	deteriorates the macula; distortion and loss of central vision	anti-VEGF medications and supplements	progressive accumulation of drusen in macula; geographic atrophy, increased pigment, and depigmentation
Glaucoma	damage to the optic nerve and cause vision loss	medication, laser treatment, or surgery to slow or stop the progression	analyzing cup to disc ratio, Inferior Superior Nasal Temporal (ISNT) features of cup and disc, and Optic Nerve Head atrophy
Melanoma	cancer of the eye involving the iris, ciliary body, or choroid	radiation therapy; Gamma Knife therapy; Thermotherapy; Surgery (resection or enucleation).	tumors (2.5mm thick) look like pigmented dome shaped mass that extends from the ciliary body or choroid; orange lipofuscin pigmentation or subretinal fluid

Table 1: Disease description, treatments on early detection and visual symptoms in fundus images [9, 11, 12, 45]

.

The generated word is then fed into the next timestep as an input. We use teacher forcing to prevent training biases with ratio to 0.5. We use the cross entropy loss summed across all generated words.

3.2. Multi-Task learning (MTL)

A simple approach is to model the three tasks as three independent multi-class classification problems using neural networks. Such an approach would result in the model parameters growing as a factor of the number of tasks, and does not exploit task dependencies. To address these, we apply MTL by sharing hidden layers between all tasks, while keeping several task-specific output layers. Hence, we use a shared convolutional neural network (CNN) to obtain a latent representation for the input image. Few CNN architectures are popular like: AlexNet [46], Inception v3 [47], VGGNet-19 [48] and Resnet-50 [3]. Since ResNet-50 outperformed other architectures for multiple of our early experiments, we choose to use it to produce our shared encoded representation. Task specific layers for each task are conditioned on this output from ResNet-50. A detailed architecture diagram of the MTL model is provided in the supplementary material.

We use the small labeled data L to train a MTL-based deep ResNet-50 model M_1 (as shown in Fig. 2). Each instance in L consists of an image I along with a label $s = (s_1, s_2, s_3)$. Note that s actually consists of three labels: broad classification label (s_1), fine-grained classification label (s_2), and diagnosis (s_3). Further, each instance $\langle I, s, s' \rangle$ is scored against the model M_1 to obtain a prediction vector $s' = (s'_1, s'_2, s'_3)$. s'_1 is of size 5. s'_2 is of size 320 . s'_3 is a prediction matrix of size $|s_3|$x$|V|$ (where V is the vocabulary) and stores predictions for each time step of the LSTM. Thus, now, for every image $I \in L$, we have hard labels s as well as M_1 predicted soft labels s'. The final loss for the MTL M_1 model is computed as the weighted combination of the individual losses as shown in Eq. 1.

$$L^{M_1}(\langle I, s, s' \rangle) = \sum_{t=1}^{3} \lambda_1^t \text{CE}(s_t, s'_t) \tag{1}$$

where CE represents cross-entropy loss, and $\{\lambda_1^t\}_{t=1}^3$ are tunable hyper-parameters for model M_1.

3.3. Knowledge Distillation (KD) with MTL

KD (or student-teacher networks [42]) is a model compression method in which a small model (student) is trained to mimic a pre-trained, larger model (teacher), or an ensemble of models. In our paper, the student model has the same capacity as the teacher model. We adapt the KD approach for improved regularization and semi-supervised learning.

KD with Labeled Data. KD uses the predicted distributions from the teacher and the student to define a KL divergence loss for training the student. These predicted distributions are obtained after the softmax activation which takes a hyper-parameter called temperature τ (usually $\tau > 1$). Our student model M_2 has the same architecture as teacher M_1. Loss for M_2 is a linear combination of cross entropy loss with respect to hard labels (between s and s') and KL-divergence loss with respect to soft labels (between s' and s'') where $s'' = (s''_1, s''_2, s''_3)$ is output prediction from M_2 as shown in Eq. 2.

$$L^{M_2}(\langle I, s, s', s'' \rangle) = \sum_{t=1}^{3} \lambda_2^t \left[\text{CE}(s_t, s'_t) + \text{KL}(s'_t, s''_t)\right] \tag{2}$$

	DR	AMD	Glaucoma	Melanoma
1-gram	retinopathy, diabetic, edema, macular, proliferative	macular, degeneration, age-related, nonexudative, retina	glaucoma, open-angle, stage, primary, severe	melanoma, malignant, choroidal, uvea, ciliary
2-gram	macular edema, diabetic retinopathy, diabetes mellitus, proliferative diabetic, type 2	macular degeneration, age-related macular, of retina, senile macular, degeneration of	glaucoma suspect, open-angle glaucoma, primary open-angle, severe stage, glaucoma severe	malignant melanoma, choroidal malignant, melanoma of, of uvea, ciliary body
3-gram	proliferative diabetic retinopathy, type 2 diabetes, 2 diabetes mellitus	age-related macular degeneration, senile macular degeneration, degeneration of retina, nonexudative senile macular	primary open-angle glaucoma, glaucoma severe stage, glaucoma moderate stage	choroidal malignant melanoma, melanoma of choroid, melanoma of uvea, melanoma of ciliary

Table 2: Most popular unigrams, bigrams and trigrams per disease, from the diagnosis provided by ophthalmologists.

where CE and KL represent cross-entropy and KL divergence loss respectively, and $\{\lambda_2^t\}_{t=1}^3$ are tunable hyper-parameters for model M_2.

Learning from Unlabeled Data. Further, we adapt the KD approach for improved semi-supervised learning as follows. Images I from unlabeled set U are scored against the model M_2 to obtain a soft prediction s''. These predictions s'' are then used to further fine-tune model M_2 to obtain model M_3 using the KL-divergence loss between M_3's predicted class/ diagnosis distributions (s''') and s'' as shown in Eq. 3.

$$L^{M_3}(\langle I, s'', s''' \rangle) = \sum_{t=1}^{3} \lambda_3^t \text{KL}(s_t'', s_t''') \tag{3}$$

where KL represents KL divergence loss, and $\{\lambda_3^t\}_{t=1}^3$ are tunable hyper-parameters for model M_3. Given that models M_1 and M_2 have been trained on a relatively small set L, we hope that the addition of soft labels from U, will ensure that model M_3 generalizes better than the prior models.

Learning from Multiple Teachers. In the MTL setup, different tasks may need variable amounts of data as well as number of steps to generalize adequately. Combinations of tasks may improve or hinder the learning process. We train models for every sub-combination of the three tasks, namely $\{T_1,\ T_2,\ T_3,\ (T_1, T_2),\ (T_1, T_3),\ (T_2, T_3),\ (T_1, T_2, T_3)\}$. We use these models as the teacher ensemble to guide learning. Specifically, during training, a student model with task T_i will distill knowledge (using KL-divergence) from teacher models which have learned task T_i. Each teacher captures a different aspect of dynamics between tasks and provides the same as supervision to the student model. As discussed above, teachers are also trained on the labeled set L and fine-tuned on U.

4. Experiments

4.1. Dataset and Experimental Settings

Our dataset consists of 7212 labeled and 35854 unlabeled fundus images corresponding to 3502 patients who visited the hidden-for-review institute from 2004 to 2017. Fundus images have been captured using Topcon 50X with Ophthalmic Imaging Systems (OIS) capture station. The images vary in resolution and field of view. We resized all the images to a standard size of 224x224x3. We also normalized the pixel values to be between -1 and 1. The labeled images are assigned to either normal category or to one of the four broad disease categories: DR, AMD, glaucoma or melanoma. Labeled data is almost balanced. Our cleaned dataset contains 320 fine-grained sub categories of diseases. Each image is also labeled with a diagnosis description provided by ophthalmologists, with an average length of 7.69 words and mode of 4 words. We filtered out diagnosis that occur less than 10 times. These diagnosis contain a vocabulary of 237 words. Disease specific diagnosis vocabulary sizes are as follows: DR 130, AMD 59, Glaucoma 71, and Melanoma 15. Fig. 3 shows the diagnosis caption length distribution in words. Of these 237 words, we retained the most frequent 193. Also, we set maximum caption length to 14 and truncate longer captions. Table 2 shows top most popular unigrams, bigrams and trigrams per disease.

Experimental Settings. All architecture and hyper-parameters choices were based on cross-validation. We choose to use Resnet-50 (initialized with pre-trained ImageNet weight) for the architecture of our shared image encoder. The architecture is trained with stochastic gradient descent (SGD) with Nesterov momentum. We set the learning rate for SGD to 0.01, with momentum value of 0.9. Our experiments showed that Adam optimizer also performed similar to SGD. We use early stopping, which on average

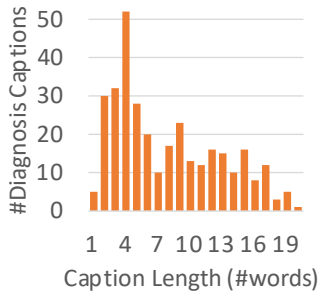

Figure 3: Length distribution of captions (in words).

completes training in 30 epochs. All models are trained on an Intel Xeon CPU machine with 32 cores and 2 Titan X GPUs (each with 12GB RAM) with a batch size of 128. For MTL, we set $\{\lambda^t_{\{1,2,3\}}\}^3_{t=1}$ to 1 so as to give equal importance to each task. We additionally use an $L2$ weight regularization, whose coefficient we set to 10^{-6}. For the LSTM decoder, we use embeddings of size 256, hidden layer with 512 units and teacher forcing ratio of 0.5. We learn embeddings dynamically and initialize with random embeddings. We use a fixed 15% of the labeled data for validation and test each. We learn various models by varying the training set percentages as p=[15, 30, 45, 60, 70]. All experiments are conducted in PyTorch. We make the code publicly available here[4].

4.2. Results

We attempt to answer the following questions: (1) How does training data size effect the MTL? (2) Does KD help individual tasks? (3) How do we find the optimal KD temperature for each task? (4) Behavior of task combinations with KD, and (5) Does using teacher ensemble for KD help? To answer these, we need to experiment with these design choices (with cardinality in brackets): (1) use KD or not (two), (2) task combinations (seven), (3) training data size (five), (4) temperature for KD (four), and (5) use of teacher ensemble or not for KD (two). Overall, there are a total of 2x7x5x4x2=560 combinations of hyper-parameters. After pruning hyper-parameters (as described in the subsections below), we end up performing 121 experiments to answer these questions. We are the only work focusing on modelling multiple eye diseases with the same model, leading to lack of good baselines.

Effect of varying training data size on MTL (M_1 results). We investigate whether MTL helps with varying amount of labeled data without any KD. Table 3 highlights the performance on various tasks after training on a combination of tasks with varying dataset size p. Performance is measured using multi-class classification accuracy for tasks T_1 and T_2, and BLEU [49] score for T_3. Note that if we do not include a task in the MTL combination for training, we cannot use it for evaluation; hence some table cells appear empty. Clearly, MTL leads to improved accuracies compared to independently learning for either of the three tasks, across most (dataset size, task) combinations. Comparing with our best MTL model (last row from Table 3), T_3 benefits most with an average of 9.6% gain; while T_1 and T_2 have average gains of 1.4 and 4.8%. Also, benefits from MTL are higher for larger dataset sizes.

MTL + KD with Labeled Data (M_2 results). Comparing the last row of Table 3 with row 4 of Table 5, we observe that M_2 improves over the M_1 accuracy by 2.2%, 2.2% and 5.2% across the three tasks resp. For detailed performance of model M_2 across various task combinations, please refer to the supplementary.

KD + Temperature Optimization for each task. For each (task, data size) pair, we find the best temperature setting. To manage the burden of extensive computation, we follow a greedy exploration of τ, i.e., we select τ based on the best performance on individual tasks rather than a combination. Intuitively, if we had a perfect teacher, we could use an output distribution closer to a one-hot vector; if the teacher is imperfect, we would prefer to use a softer output distribution. This effect can be emulated by experimenting with different temperature values τ as [0.5, 1, 5, 10]. Small τ implies sharper distributions, while larger τ implies softer distributions. Balancing parameter α for classifier with hard+soft labels was set to 0.95 (for soft loss). Table 4 shows that across various data fractions and across tasks, we observe better results than in Table 3. Thus, KD seems to be helpful across dataset sizes for retinal disease prediction. The best τ across tasks is 10 for dataset size p=15% while it is 5 for p=70% for most tasks. This leads to an intuitive observation: when p is small, the task is harder to learn and hence higher τ is desired. We show the analysis for M_3 only but we followed a similar approach for M_2 as well.

MTL + KD with Unlabeled Data + Teacher Ensemble (M_3 results). Table 5 shows results when we combine MTL with KD. (T_1, T_2, T_3) is the best among all task combinations. Compared to only MTL (Table 3) with all three tasks together, MTL+KD shows massive gains especially with smaller data fractions for each task. Finally, we enable teacher ensemble for KD, rather than learning from a single teacher. Learning from teacher ensemble in an MTL setup implies that a model with tasks (T_1, T_2, T_3) can distill knowledge from teachers which were trained on task combinations (T_1), (T_2), (T_3), (T_1, T_2), (T_1, T_3), (T_2, T_3), (T_1, T_2, T_3). Last line in Table 5 shows that MTL+KD with teacher ensemble is the best method, significantly better than the initial baseline. Comparing rows 4 and 6, M_3 improves task accuracies by 1.7%, 3.3%, 14.1%

[4] https://github.com/SahilC/mtl-kd-disease-recognition

#	Test→	T_1 (Accuracy)					T_2 (Accuracy)					T_3 (BLEU)				
	MTL Train↓ p →	15	30	45	60	70	15	30	45	60	70	15	30	45	60	70
1	T_1	0.684	0.755	0.755	0.774	0.744										
2	T_2						0.382	**0.417**	0.434	0.392	0.415					
3	T_3											0.234	0.297	0.321	0.349	0.354
4	T_1,T_2	**0.729**	0.760	0.766	**0.778**	0.775	0.336	0.379	0.409	0.392	0.429					
5	T_1,T_3	0.701	0.731	**0.772**	0.770	**0.800**						0.225	0.306	0.336	0.350	0.374
6	T_2,T_3						0.361	0.394	0.394	0.452	0.439	0.237	0.306	**0.338**	0.370	0.369
7	T_1,T_2,T_3	0.693	**0.765**	0.754	0.769	0.781	**0.386**	0.407	**0.435**	**0.461**	**0.447**	**0.258**	**0.314**	0.337	**0.411**	**0.387**

Table 3: Test Accuracy for MTL on different combinations of tasks with varying dataset size p. No KD. Empty cells represent cases where evaluation does not make sense. These are the results for model M_1.

#	Test→	T_1 (Accuracy)					T_2 (Accuracy)					T_3 (BLEU)				
	$\tau \downarrow p \rightarrow$	15	30	45	60	70	15	30	45	60	70	15	30	45	60	70
1	0.5	0.687	0.750	0.749	0.784	0.779	0.390	**0.416**	0.439	0.418	0.413	0.249	0.311	0.336	0.348	0.366
2	1	0.752	0.769	**0.766**	0.778	0.789	0.387	0.410	0.434	0.464	**0.479**	**0.319**	**0.327**	0.342	0.429	0.449
3	5	**0.769**	**0.787**	0.763	**0.827**	**0.818**	0.392	0.398	**0.444**	0.458	0.477	0.285	0.289	**0.376**	**0.455**	**0.454**
4	10	**0.769**	0.767	0.755	0.825	0.816	**0.407**	0.408	0.437	**0.473**	0.467	0.288	0.295	0.364	0.394	0.443

Table 4: Test Accuracy for KD on individual tasks (no MTL) with varying dataset size p and varying temperature τ. These are the results for model M_3.

#	Test→	T_1 (Accuracy)					T_2 (Accuracy)					T_3 (BLEU)				
	MTL Train↓ p →	15	30	45	60	70	15	30	45	60	70	15	30	45	60	70
1	T_1,T_2(M_3)	0.740	0.752	0.783	0.800	0.774	0.384	0.399	0.422	0.411	0.474					
2	T_1,T_3(M_3)	0.732	0.733	0.750	0.777	0.783						0.330	0.338	0.394	0.413	0.454
3	T_2,T_3(M_3)						0.365	0.408	0.459	0.491	0.462	0.333	0.315	0.390	0.438	0.428
4	T_1,T_2,T_3(M_2)	0.746	0.771	0.760	0.782	0.782	0.391	0.407	0.438	0.476	0.473	0.261	0.325	0.357	0.445	0.415
5	T_1,T_2,T_3+Ensemble(M_2)	0.760	0.779	0.759	**0.803**	0.782	0.397	0.411	0.438	0.480	0.481	0.258	0.333	0.353	0.447	0.421
6	T_1,T_2,T_3(M_3)	0.751	0.764	**0.790**	0.795	0.808	0.384	0.436	0.454	0.482	0.503	0.337	0.363	0.400	0.460	0.475
7	T_1,T_2,T_3+Ensemble(M_3)	**0.765**	**0.782**	0.789	**0.803**	**0.825**	**0.414**	**0.441**	**0.487**	**0.507**	**0.511**	**0.346**	**0.371**	**0.432**	**0.473**	**0.481**

Table 5: Test Accuracy for KD+MTL on different combinations of tasks with varying dataset size p. For each cell, τ is the best temperature chosen for the (task combination, dataset size) pair from Table 4. Rows with "+Ensemble" correspond to using teacher ensemble for distillation. The results of using the combination of all three tasks in M_2 are shown for comparison. The rest of the results are for model M_3.

respectively over M_2. Similarly, comparing rows 5 and 7, M_3 improves task accuracies by 2.1%, 6.9%, 17.6% respectively over M_2. This proves that distillation with unlabeled data is extremely useful.

Detailed Analysis of our Best Model. Table 6 shows confusion matrix for the best result for task T_1. Interestingly the precision and recall across all classes is between 0.71 and 0.97. Notably, recall for melanoma is as high as 0.97. Fig. 4 shows top-K accuracy values for the fine-grained disease classification task across diseases for our best T_2 classifier. Top-5 accuracy values for task T_2 are ∼96% for melanoma but on average top-5 accuracy across all diseases is ∼75%. Further, for task T_3, BLEU scores across the diseases are

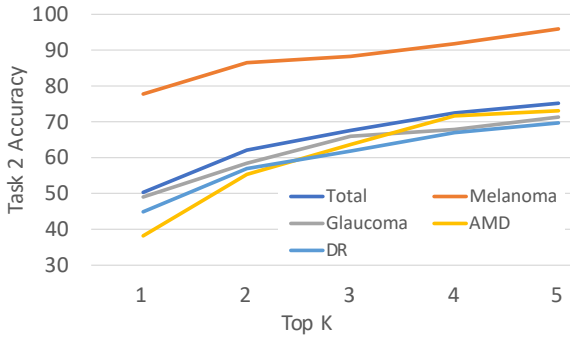

Figure 4: Top-K accuracy achieved by our best model values for fine-grained disease prediction (T_2) across diseases.

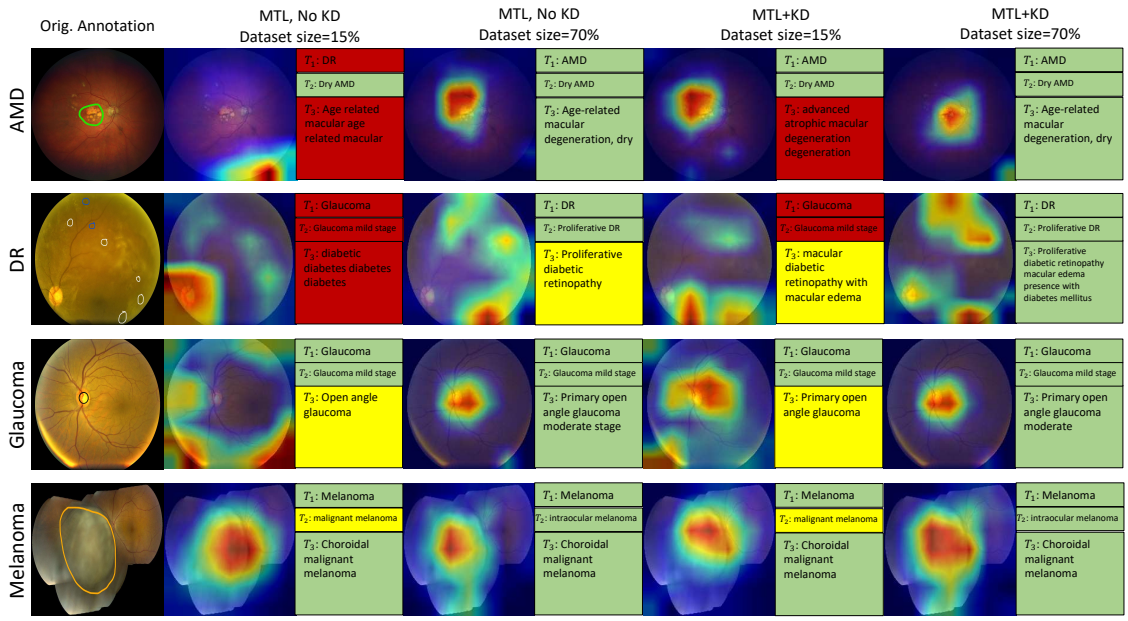

Figure 5: Grad-CAM visualization for predictions using the model M_1 versus the proposed MTL+KD model M_3 across different diseases and different training dataset sizes along with their corresponding model outputs for T_1, T_2, T_3. (Green∼Correct, Yellow∼Partially correct, Red∼Incorrect)

as follows: Melanoma (73.84), Glaucoma (50.62), AMD (73.84) and DR (39.86). We achieve high BLEU scores for Melanoma, AMD, moderate BLEU for Glaucoma and relatively low BLEU for DR. Note that DR has the largest vocabulary size of 130 words, making the captions associated with the disease vary more than for other diseases, which had smaller vocabulary sizes. BLEU scores remain similar across diseases even for longer captions. Detailed error analysis can be found in the supplementary.

		Predicted				
		Melanoma	Glaucoma	AMD	DR	Normal
Actual	Melanoma	**195**	0	2	4	0
	Glaucoma	0	**177**	9	3	0
	AMD	3	5	**178**	5	23
	DR	23	8	16	**181**	23
	Normal	17	21	15	12	**162**

Table 6: Confusion matrix for broad disease prediction (T_1). As shown, we have a high correlation between actual and predicted values, indicating our model is effective.

Grad-CAM Visualization. We used Gradient-weighted Class Activation Mapping (Grad-CAM) [50] to visualize the regions of fundus image that are "important" for disease predictions. It captures how intensely the input image activates different channels by computing how important each channel is with regard to the class. Fig. 5 shows class activation mapping visualizations for four randomly selected images, across the four diseases. The first column shows images with anomaly annotations by an ophthalmologist. The remaining columns show class activation mappings obtained using Grad-CAM for predictions by model M_1 (MTL but no KD) versus the best method M_3 (MTL+KD) for two different dataset sizes (15% and 70%). We show predicted outputs for all the three tasks: T_1, T_2 and T_3 (Green∼Correct, Yellow∼Partially correct, Red∼Incorrect). We observe that the Grad-CAM activations highly correlate with expert annotations across all the four images for a dataset size of 70%. However, for small (15%) dataset size, activations generated using KD+MTL have much higher correlation with expert annotations compared to those generated using the basic classifier, and hence lead to more accurate predictions.

5. Conclusion

We proposed the use of MTL and KD methods to improve fine-grained recognition of eye diseases using small labeled dataset of fundus images. Both KD and MTL result in massive boosts in performance across metrics. Tuning softmax temperature is important. Teacher ensemble method is more effective than just one teacher for KD. In the future, we plan to experiment with additional auxiliary tasks and images corresponding to more diseases.

References

[1] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *ophth.*, vol. 121, no. 11, pp. 2081–2090, 2014.

[2] J. B. Jonas, C. M. G. Cheung, and S. Panda-Jonas, "Updates on the epidemiology of age-related macular degeneration," *The Asia-Pacific Journal of ophth.*, vol. 6, no. 6, pp. 493–497, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.

[4] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, 2016.

[5] M. Gupta, P. Gupta, P. K. Vaddavalli, and A. Fatima, "Predicting post-operative visual acuity for lasik surgeries," in *PAKDD*, pp. 489–501, 2016.

[6] H. L. Rao, U. K. Addepalli, R. K. Yadav, N. S. Choudhari, S. Senthil, A. K. Mandal, and C. S. Garudadri, "Accuracy Of Ordinary Least Squares And Empirical Bayes Estimates Of Short Term Visual Field Progression Rates To Predict Long Term Outcomes In Glaucoma," *Investigative ophth. & Visual Science*, vol. 53, no. 14, pp. 182–182, 2012.

[7] L. Torquetti, G. Ferrara, and P. Ferrara, "Predictors of Clinical Outcomes after Intrastromal Corneal Ring Segments Implantation," *Int J Keratoconus Ectatic Corneal Dis*, vol. 1, pp. 26–30, 2012.

[8] C. Bowd, R. N. Weinreb, M. Balasubramanian, I. Lee, G. Jang, S. Yousefi, L. M. Zangwill, F. A. Medeiros, C. A. Girkin, J. M. Liebmann, *et al.*, "Glaucomatous Patterns in Frequency Doubling Technology (FDT) Perimetry Data identified by Unsupervised Machine Learning Classifiers," *PloS one*, vol. 9, no. 1, p. e85941, 2014.

[9] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao, "Disc-aware ensemble network for glaucoma screening from fundus image," *TMI*, vol. 37, no. 11, pp. 2493–2501, 2018.

[10] P. Fraccaro, M. Nicolo, M. Bonetto, M. Giacomini, P. Weller, C. E. Traverso, M. Prosperi, and D. OSullivan, "Combining Macula Clinical Signs and Patient Characteristics for Age-related Macular Degeneration Diagnosis: A Machine Learning Approach," *BMC ophth.*, vol. 15, no. 1, p. 1, 2015.

[11] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration oct images," *ophth. Retina*, vol. 1, no. 4, pp. 322–327, 2017.

[12] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[13] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An Ensemble Classification-based Approach applied to Retinal Blood Vessel Segmentation," *Bio Engg*, vol. 59, no. 9, pp. 2538–2548, 2012.

[14] J. Harbour, "Molecular Prediction of Time to Metastasis from Ocular Melanoma Fine Needle Aspirates," *Clinical Cancer Research*, vol. 12, no. 19 Supplement, pp. A77–A77, 2006.

[15] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *TMI*, vol. 35, no. 5, pp. 1153–1159, 2016.

[16] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[17] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *MICCAI*, pp. 212–220, 2016.

[18] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using cnns," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[19] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J. Path. informatics*, vol. 7, 2016.

[20] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images," *TMI*, vol. 35, no. 1, pp. 119–130, 2015.

[21] M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri," *Medical Image Analysis*, vol. 30, pp. 108–119, 2016.

[22] P. M. Cheng and H. S. Malhi, "Transfer learning with cnns for classification of abdominal ultrasound images," *J. digital imaging*, vol. 30, no. 2, pp. 234–243, 2017.

[23] F. Liu, Z. Zhou, H. Jang, A. Samsonov, G. Zhao, and R. Kijowski, "Deep cnn and 3d deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging," *Magnetic resonance in medicine*, vol. 79, no. 4, pp. 2379–2391, 2018.

[24] M. Lalonde, L. Gagnon, M.-C. Boucher, *et al.*, "Automatic visual quality assessment in optical fundus images," in *Vision Interface*, vol. 32, pp. 259–264, 2001.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241, 2015.

[26] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *ICASSP*, pp. 1626–1630, 2014.

[27] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[28] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a cnn," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 204–212, 2017.

[29] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic resonance in medicine*, vol. 79, no. 6, pp. 3055–3071, 2018.

[30] S. Chelaramani, M. Gupta, V. Agarwal, P. Gupta, and R. Habash, "Multi-task learning for eye disease prediction," in *ACPR*, 2019.

[31] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *ICML*, pp. 160–167, 2008.

[32] L. Deng, G. Hinton, and B. Kingsbury, "New types of dnn learning for speech recognition and related applications: An overview," in *ICASSP*, pp. 8599–8603, 2013.

[33] R. Girshick, "Fast r-cnn," in *ICCV*, pp. 1440–1448, 2015.

[34] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively multitask networks for drug discovery," *arXiv*, 2015.

[35] L. Duong, T. Cohn, S. Bird, and P. Cook, "Low resource dependency parsing: Cross-lingual parameter sharing in a nn parser," in *IJCNLP*, pp. 845–850, 2015.

[36] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *ICML*, pp. 41–48, 1993.

[37] L. Liu, H. Wang, J. Lin, R. Socher, and C. Xiong, "Attentive student meets multi-task teacher: Improved knowledge distillation for pretrained models," *arXiv preprint arXiv:1911.03588*, 2019.

[38] X. Liu, P. He, W. Chen, and J. Gao, "Improving multi-task deep neural networks via knowledge distillation for natural language understanding," *arXiv preprint arXiv:1904.09482*, 2019.

[39] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[40] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *NIPS*, pp. 3320–3328, 2014.

[41] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Trans. on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[42] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[43] S.-I. Mirzadeh, M. Farajtabar, A. Li, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher," *arXiv preprint arXiv:1902.03393*, 2019.

[44] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

[45] A. D. Singh, M. E. Turell, and A. K. Topham, "Uveal melanoma: trends in incidence, treatment, and survival," *Ophthalmology*, vol. 118, no. 9, pp. 1881–1885, 2011.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, pp. 1097–1105, 2012.

[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, pp. 2818–2826, 2016.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, 2014.

[49] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, pp. 311–318, 2002.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, pp. 618–626, 2017.