Multi-Task Knowledge Distillation for Eye Disease Prediction – Supplementary

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

1. MTL Architecture

Fig. 1 shows the basic architecture of our proposed MTL disease prediction system. As shown in Fig.1, we share the ResNet-50 encoder weights across all the tasks. Task specific layers for each task are conditioned on the shared ResNet-50 encoder. For task 1 and 2, we use a fully-connected layer with ReLU, and then a softmax output layer. For task 3, we feed output of the ResNet-50 encoder to an LSTM.

2. Results for M_2

Table 1 summarizes results across multiple task combinations with varying train data sizes using model M_2. As discussed in the main part of the paper, results for model M_2 are significantly better than those for model M_1 and worse compared to those for model M_3. Thus, model M_2 is in some sense an intermediate model for our main proposed model M_3.

3. Error Analysis for our Best Model (M_3)

Of the 1082 test samples, (1) Only 42 cases have coarse label correctly predicted but low BLEU (< 0.2) $\implies$ a low probability of getting wrong caption if disease label is correct. (2) 20 cases have coarse label wrong but high BLEU (> 0.5) $\implies$ it is unlikely to get the diagnosis without predicting coarse label correctly (3) 100 cases have fine grained label correctly predicted but the coarse label prediction is wrong $\implies$ With a high probability, if T_1 goes wrong, then it is likely that T_2 may still be correct. (4) 13 have coarse label correctly predicted but fine-grained label is wrong $\implies$ there is high chance of predicting both correctly together.

4. Grad-CAM Visualizations

We used Gradient-weighted Class Activation Mapping (Grad-CAM) [1] 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. 2 shows class activation mapping visualizations for six randomly selected images, across two diseases: DR and AMD. 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 our best method M_3 (MTL+KD) for two different dataset sizes (15% and 70%). We also show predicted outputs (Green$\sim$Correct and Red$\sim$Incorrect). We observe that the Grad-CAM activations highly correlate with expert annotations across all the images for a dataset size of 70%. However, for small (15%) dataset size, some cases display errors. We show similar results in Fig. 3 for glaucoma and melanoma.

References

[1] 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.

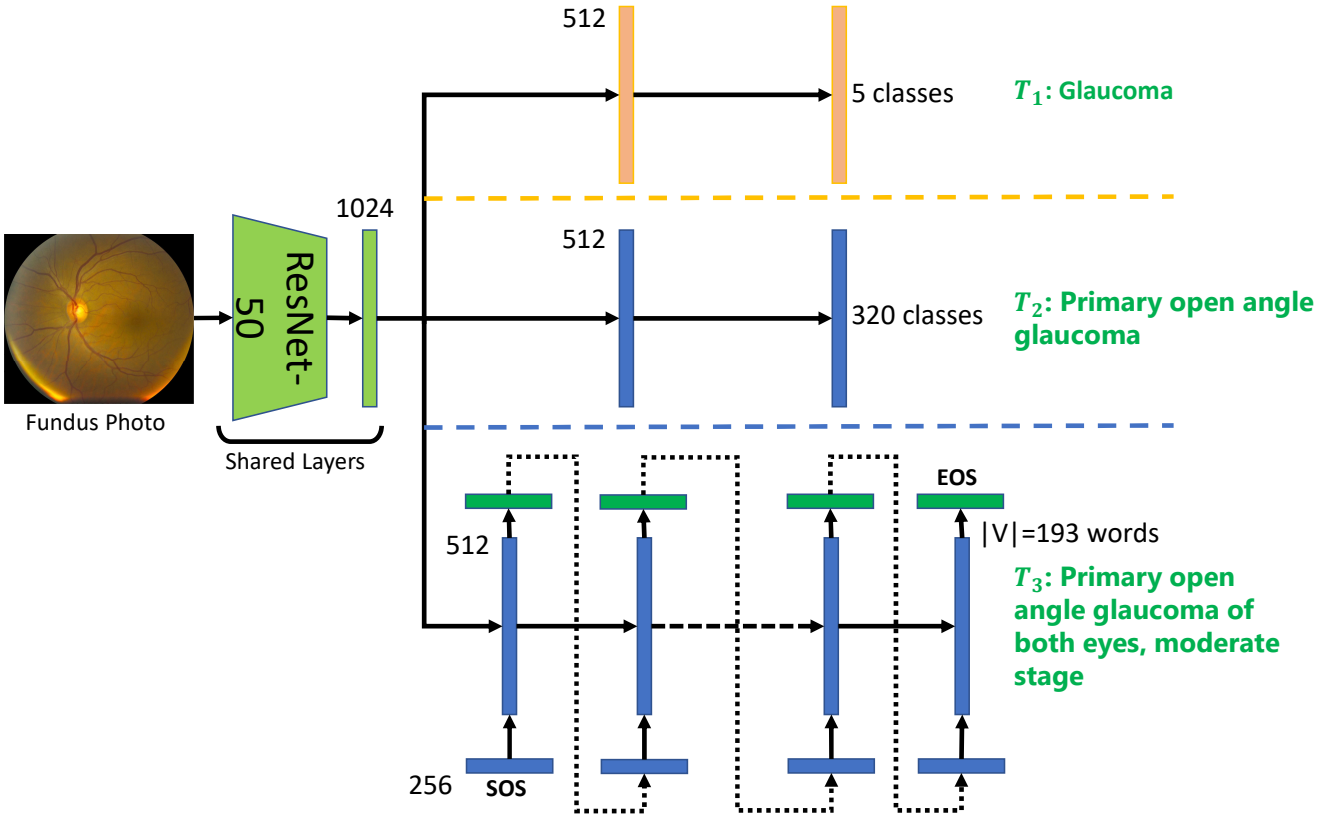

Figure 1: Architecture for the MTL model used for disease prediction. Depicted are the shared layers of a CNN from which features are extracted and fed into the corresponding tasks

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
T_1,T_2	0.73	0.761	0.780	0.803	0.77	0.35	0.38	0.428	0.407	0.432					
T_1,T_3	0.72	0.731	0.762	0.77	0.781						0.252	0.317	0.34	0.379	0.429
T_2,T_3						0.379	0.398	0.416	0.461	0.443	0.276	0.309	0.350	0.386	0.396
T_1,T_2,T_3	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
T_1,T_2,T_3+Ensemble	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

Table 1: Test Accuracy for KD+MTL on different combinations of tasks using ResNet-50 with varying dataset size p. For each cell, τ is the best temperature chosen for the (task combination, dataset size). Last row corresponds to using teacher ensemble for distillation. The rest of the results for model M_2.

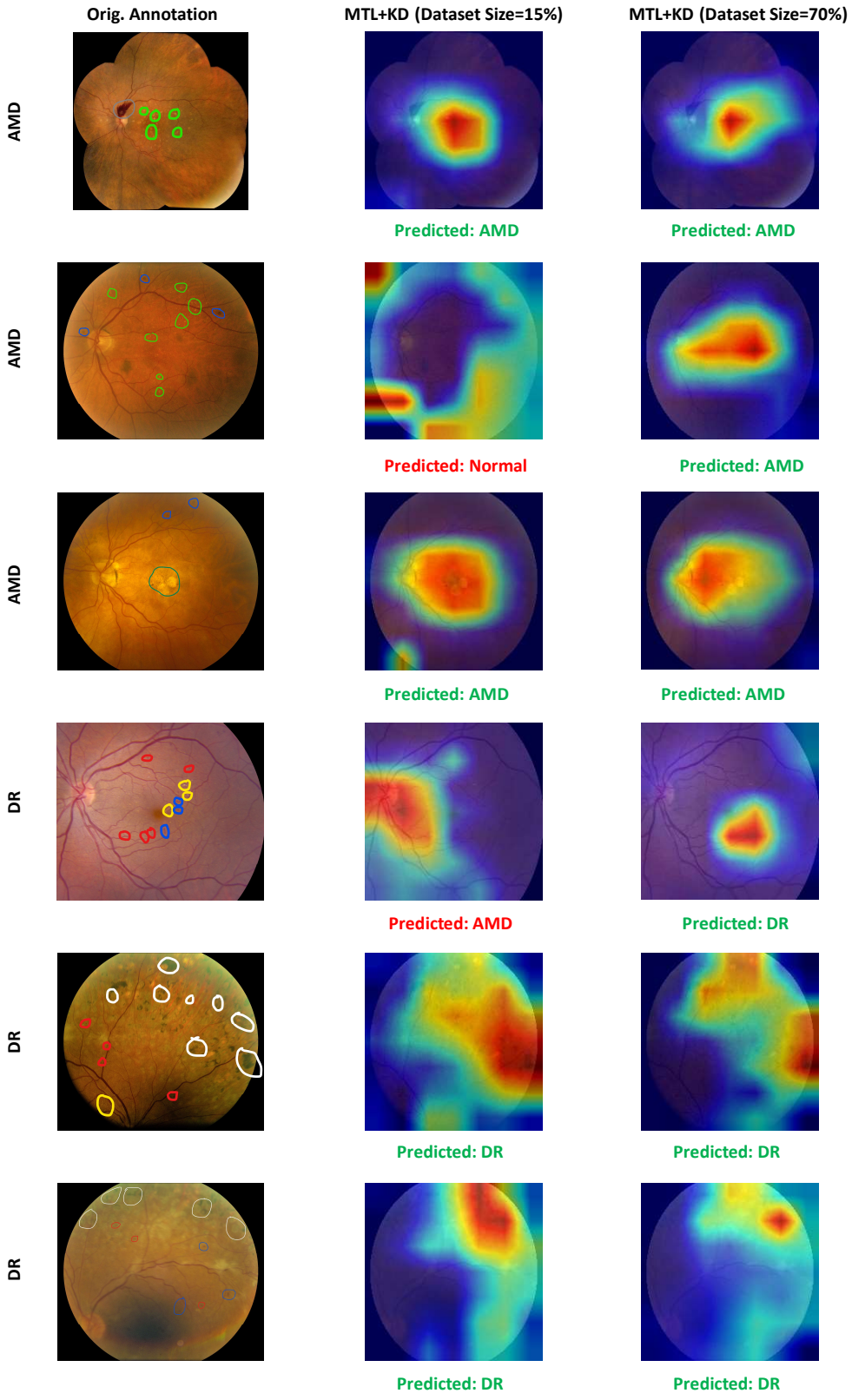

Figure 2: Grad-CAM visualization for predictions using the proposed MTL+KD model M_3 across AMD and DR and training dataset sizes set as 15% or 70% along with their corresponding model outputs. (Green∼Correct, Red∼Incorrect)

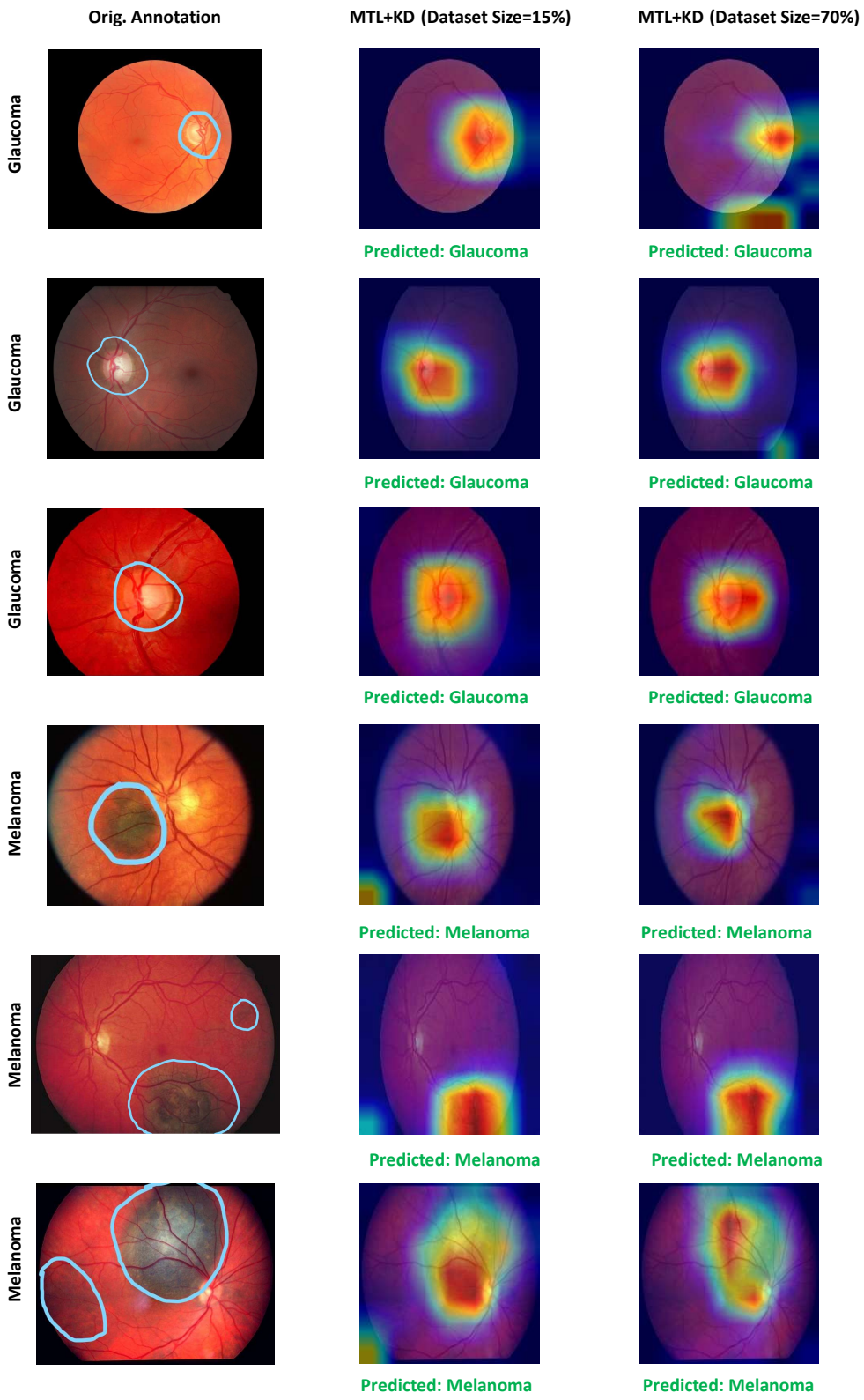

Figure 3: Grad-CAM visualization for predictions using the proposed MTL+KD model M_3 across Glaucoma and Melanoma and training dataset sizes set as 15% or 70% along with their corresponding model outputs. (Green∼Correct, Red∼Incorrect)