MED-TEX: Transferring and Explaining Knowledge with Less Data from Pretrained Medical Imaging Models

Thanh Nguyen-Duc, He Zhao, Jianfei Cai and Dinh Phung

***Abstract*— Deep neural network based image classification methods usually require a large amount of training data and lack interpretability, which are critical in the medical imaging domain. In this paper, we develop a novel knowledge distillation and model interpretation framework for medical image classification that jointly solves the above two issues. Specifically, to address the data-hungry issue, we propose to learn a small student model with less data by distilling knowledge only from a cumbersome pretrained teacher model. To interpret the teacher model as well as assisting the learning of the student, an explainer module is introduced to highlight the regions of an input medical image that are important for the predictions of the teacher model. Furthermore, the joint framework is trained by a principled way derived from the information-theoretic perspective. Our framework performance is demonstrated by the comprehensive experiments on the knowledge distillation and model interpretation tasks compared to state-of-the-art methods on a fundus disease dataset.**

***Index Terms*— Knowledge distillation, Model interpretation, Feature selection, Weakly supervised learning.**

I. Introduction

Recently, advanced machine learning methods such as Convolutional Neural Networks (CNNs) [1] have shown remarkable performance in the medical imaging domain such as U-net [2] for image segmentation, ResNet [3] for image classification and medical image reconstruction [4]. In this paper, we consider a practical scenario of medical image classification applications, where a hospital consists of a central headquarter and multiple local branches. Suppose the headquarter has developed a large CNN model for disease classification trained on a large set of data, which is the global model to be distributed to the branches. A branch often needs to develop a customized smaller model on its local data and it usually does not have the access to the large data of the headquarter that is used to train the global model, due to privacy, sensitivity or bandwidth concerns. To assist the development of the local model, a typical way is to learn the knowledge from the global model [5]. Moreover, for medical applications, an explainable model is paramount. Thus, it is highly desirable to have a local branch with two capabilities: explaining the teacher model and transferring the knowledge of the global model to the local model with only the local data.

Explaining a CNN-based image classification model has been studied before. A common way is to use top-down neural saliency such as CAM [6], [7] to locate feature that contributes the most to the classification output. Another way is through feature selection or attention [8]–[11] to generate different weights for different features. Both ways can be used to identify image regions that are important to the prediction of the classifier. However, they are not designed for explaining a pretrained global model. The recent Learning-to-Explain (L2X) method [8] trains an explainer to explain a pretrained teacher model by maximizing the mutual information between selected instance-wise features and the teacher outputs. However, L2X does not address the issue of lack of large training data and has not been reported to have impressive results for image classification. Thus, in this paper, we propose an end-to-end framework to address the above two requirements simultaneously, i.e. we aim to learn a small medical image classification model with less training data but better interpretability. Here we define the interpretability as the ability to identify the areas of an input image that are important to the prediction of the classifier.

Fig. 1(a) gives an overview of the proposed framework, which consists of a teacher $\mathcal{T}$ that is the pretrained globe model, a learnable "student" $\mathcal{S}$ that is the local model extracting the teacher's knowledge, and an explainer $\mathcal{E}$ that explains the teacher. The student is expected to be significantly smaller than the teacher to reduce the computational cost. Specifically, given an input image, the explainer highlights the important pixels for the decision of the teacher and suppresses the unimportant pixels, which addresses the interpretability aim. Then, the explainer facilitates the learning of the student by providing it a simplified input image. In this way, the student does not need to learn from the scratch, but focuses on the important regions that the explainer explains, and at the same time the teacher's knowledge is transferred to the student, which address the aim of training a small model. Interestingly, the above two aims can be jointly achieved by optimizing a joint training objective derived from an information-theoretic perspective by pushing the output of the last layer and intermediate layers of the student close to those of the teacher (see Section III-B).

Manuscript submitted August 1, 2020. We would like to thank you International Cao Thang Eye Hospital for sharing data and Mr. Quang Nguyen from A2DS (a2ds.io) for valuable comments on eye diseases.

The authors are with the Monash University, Melbourne, Australia (e-mail: thanh.nguyen4@monash.edu, ethan.zhao@monash.edu, jianfei.cai@monash.edu and dinh.phung@monash.edu).

It is noteworthy that since the explainer selects important image regions, our work appears to be similar to weakly supervised image segmentation with only image-level annotations [12]–[15]. The fundamental difference is that weakly supervised image segmentation is for the purpose of generating best segmentation with only image-level labels, while our partial goal is to identify the most important image regions w.r.t the teacher's prediction, with the other aim on learning a small student model.

Our contributions can be summarised as follows:

- We propose a new end-to-end MEDical Transferring and EXplaining framework (MED-TEX) from a pretrained gloabl model, which combines knowledge distillation and model interpretation. To our knowledge, our approach is novel on solving two important issues in medical imaging in a joint framework: the lack of training data and the lack of interpretation. Existing methods only focus on either of them. See Section II for more discussion.
- We develop a joint training objective for our framework, derived from an information-theoretic perspective. Specifically, we introduce to maximize the mutual information between not only the output layers but also the intermediate layers of the student and the teacher, which is both theoretically and practically appealing.
- Extensive experimental results demonstrate that our proposed method achieve better performance on the evaluations of both knowledge distillation and model interpretability. Our approach outperforms many others on identifying important image regions, including soft attention [10], [11], hard attention [9], learning to explain [8], and Grad-CAM [7].

The rest of this paper is organized as follows. We review the related work in Section II. Section III introduces our proposed MED-TEX framework in detail. We demonstrate the performance of our framework in Section IV. Finally, we give a conclusion in Section V.

II. Related work

In this section, we review related work, including knowledge transferring (or knowledge distillation), model interpretation by feature selection, and image segmentation with only image-level annotations.

Knowledge distillation (KD): This is a process of transferring knowledge from a complicated pretrained model (teacher) to a smaller lighter-weighted one (student). The student is particularly useful in the cases where computational resources and deployment cost need to be significantly reduced at the inference stage. KD was introduced originally by Hinton et al. [5] to extract knowledge from the distribution of class probabilities predicted by the teacher model. Romero at el. [16] then proposed to distill information from intermediate layers of the teacher to the corresponding layers of the student. More works [17]–[19] introduced the relaxation using the regularization to carefully choose what information to distill from the teacher to the student, e.g., attention maps [18]. Recently, Ahn et al. [20] exploited the information-theoretic perspective as maximizing the mutual information between the teacher and the student in order to transfer knowledge. There are also some attempts to apply KD in the medical imaging domain. For example, Wang et al. [21] used KD to train a student model that speeds up the inference time of a 3D neuron segmentation model. The work of [22] leveraged KD for brain lesion segmentation with soft labels by dilating mask boundaries. Transferring knowledge from multiple sources to promote lung pattern analysis was introduced by Christodoulidis et al. [23]. KD was also explored for improving unpaired multi-modal segmentation in [24]. Compared with these existing KD methods, our framework can not only transfer knowledge from the teacher, but also interpret the teacher's behaviours by introducing the explainer, which is critical to medical imaging applications.

Model interpretation: The rapid growing of machine learning in many applications leads to a strong requirement for model interpretation, especially in the medical imaging domain. One common way to provide visual explanability is using feature localization to locate most relevant features in terms of contributions to model prediction/classification, given an input image. Class Activation Maps (CAM) [7], [25] is the most representative feature localization method, which maps a predicted class to the regions in the feature domain corresponding to locations in the input image. Although CAM can be used to highlight meaningful image locations, its map is very coarse because the spatial size is usually significantly reduced during feature extraction process of neural networks.

Another common way for model interpretation is via feature selection or attention. Feature selection or attention is to automatically generate different weights according to feature content and assign the weights to different feature regions. In general, attention can be divided into "soft" attention and "hard" attention, where the former involves differentiable functions and generates continuous attention weights while the latter involves non-differentiable functions and generates binary weights. Soft attention is often used in feature domain and has been applied in many applications. For example, Xu et al. exploited soft attention [10] for natural image captioning and Yang et al. [26] introduced the guided soft attention for histopathology breast cancer detection. Soft attention can also be applied for weakly supervised image segmentation [12]. For hard attention, which usually involves back-propagation through discrete variables, several tricks can be applied to make the model differentiable suck as REINFORCE [27] and Gumbel-softmax trick [9], [28], [29].

Although attention can highlight semantic regions, it is usually trained for the purpose of maximizing the classification accuracy, not for explaining the pretrained teacher model in our setting. The recent Learning-to-Explain (L2X) approach [8] is the most relevant one. It trains an explainer to explain the pretrained teacher model by maximizing mutual information between selected instance-wise features and the teacher outputs. Its feature selection is based on hard attention with Gumbel-softmax trick. Compared with [8], our method generates soft attention in the pixel domain instead of the feature domain of input images. Moreover, our method also learns a smaller student model with only local data, with information distillation from intermediate layers, which is not

considered in L2X.

Image segmentation with only image-level annotation: Our pixel selection for model explanation essentially generates some segmentation results. This is related to weakly supervised image segmentation with only image-level annotations. Both CAM and attentions have been applied for weakly supervised image segmentation in medical imaging domain. For example, Izadyyazdanabadi et al. [13] applied CAM for diagnostic brain tumor segmentation in confocal laser endomicroscopy glioma images and the work from Feng et al. [14] introduced a coarse image segmentation followed by a fine instance-level segmentation. Rajpurkar et al. [15] also used CAM for chest X-ray segmentation. In the work [30], both "hard" and "soft" attentions are used for robust brain magnetic resonance image segmentation for hydrocephalus patients. In contrast, our method is mainly designed to identify the most important regions of an input image to the teacher's prediction but not to generate best segmentation, although those important regions are highly overlapped with segmentations because of the well trained teacher's behaviors. For example, when the teacher predicts a certain disease, our method is trained to detect which parts of the image that cause the disease based on the prediction of the teacher. Moreover, our goal is to train a smaller student model that can achieve similar classification performance as the teacher with only local data, while simultaneously be able to explain the teacher via pixel selection, which is expected to match segmentation to a certain extent.

III. Transferring and Explaining Knowledge from Medical Pretrained Models (MED-TEX).

In this section, we introduce the details of our framework MED-TEX which includes a fixed pretrained teacher and two trainable modules called explainer and student, as illustrated in Fig. 1(a). Recalling the hosptial example in the introduction, suppose that a CNN-based classifier (teacher, $\mathcal{T}$) is pretrained to classify images which is usually a cumbersome model in order to adapt to large-scale dataset from the headquarter. The student $\mathcal{S}$ is another CNN-based classifier that can be more than a hundred times smaller to significantly reduce computational complexity. It is noteworthy that the dataset used for training the teacher may not be accessible to us due to sensitivity or privacy. With a raw input image, $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ (C, H, W are the channels, height, and width of the image, respectively), the explainer $\mathcal{E}$ inspired by the U-net [2] architecture produces selection scores $\mathbf{\Theta}$, which give high scores for the important pixels for the decision of the teacher and low scores for the unimportant ones. In our framework, $\mathbf{\Theta}$ has the same size to $\mathbf{X}$ and is element-wise multiplied by $\mathbf{X}$ to get a simplified the input image, denoted by $\mathbf{X}'$. This $\mathbf{X}'$ then is input to the student $\mathcal{S}$ to perform predictions. Our goal is training the student to mimic the behaviors of the teacher by pushing teacher's outputs from the last and intermediate layers close to student's outputs while the explainer makes use of $\mathbf{\Theta}$ guide the student by simplifying input $\mathbf{X}$ into $\mathbf{X}'$, as illustrated in Fig. 1. The architecture details of teacher, student and explainer will be elaborated on later in Section IV.

A. Proposed framework

Here we denote the teacher's and student's predicted distributions over the labels as $\boldsymbol{y}^{\mathcal{T}} \in \Delta^L$ and $\boldsymbol{y}^{\mathcal{S}} \in \Delta^L$, respectively, where L is the number of labels and Δ^L denotes the L dimensional simplex. We assume there are M layers of the CNNs of the teacher and the student, where the first to $(M-1)^{\text{th}}$ layers are convolutional layers (or block convolution layers) and the last one is a fully connected layer. These predicted distributions are from the output (M^{th}) layers of the teacher and the student. We further have $\boldsymbol{y}^{\mathcal{T}} = \mathcal{T}(\mathbf{X})$ (i.e., $\mathrm{p}(y_l^{\mathcal{T}} \mid \mathbf{X}) \propto \mathcal{T}(\mathbf{X})_l$), $\mathbf{X}'|\mathbf{X} = \mathcal{E}(\mathbf{X})$, and $\boldsymbol{y}^{\mathcal{S}}|\mathbf{X}' = \mathcal{S}(\mathbf{X}')$ (i.e., $\mathrm{q}(y_l^{\mathcal{S}} \mid \mathbf{X}') \propto \mathcal{S}(\mathbf{X})_l$). With these notations, we can formulate our preliminary goals of explaining and extracting the teacher's knowledge to the student as the following loss derived from mutual information (See the derivation in Eq. (10)).

$$\mathcal{L}^M = \min_{\mathcal{E},\mathcal{S}} -\mathbb{E}_{\mathbf{X}}\Big[\mathbb{E}_{\mathbf{X}'|\mathbf{X}}\Big[\mathbb{E}_{\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}'}[\log q(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')]\Big]\Big], \tag{1}$$

where q corresponds to our student, acting as the variational distribution in the deviation of mutual information in Section III-B.

Essentially, Eq. (1) can be understood as minimizing the cross-entropy loss between the outputs of the teacher and the student and generate $\mathbf{X}'$ by element-wise multiplication between $\mathbf{X}$ and $\mathbf{\Theta}$, aiming to push the predictions of the student close to those of the teacher, with the help from the explainer:

$$\mathcal{L}^M = \min_{\mathcal{E},\mathcal{S}} -\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}'|\mathbf{X}}\left[\sum_l^L \mathrm{p}(y_l^{\mathcal{T}} \mid \mathbf{X}) \log \mathrm{q}(y_l^{\mathcal{S}} \mid \mathbf{X}')\right]\right]. \tag{2}$$

Next, we introduce the detailed construction of the explainer. Specifically, given an input image $\mathbf{X}$, the explainer generates an importance score for each of its pixels. The higher the important score is, the more important the corresponding pixel is to the prediction of the teacher. All the importance scores form the importance map[1], denoted as $\mathbf{\Theta} \in [0,1]^{C\times H\times W}$. In this way, the output of the explainer can be expressed as

$$\mathbf{X}' = \mathbf{\Theta} \odot \mathbf{X}, \tag{3}$$

where $\odot$ is the element-wise multiplication.

We construct the explainer $\mathcal{E}$ with a neural network inspired by Unet [2], which can output a high resolution probability map (typically same size as input), denoted as $\mathbf{\Theta}|\mathbf{X} = U_{\mathcal{E}}(\mathbf{X})$. Specifically, the explainer produces $\mathbf{\Theta}$ scores for both channels and spatial locations of $\mathbf{X}$. The channel selection is via a fully connected layer with sigmoid activation function, which takes the output of the explainer's encoder. The channel selection component is especially beneficial for medical images with multiple channels. The spatial selection is the output from the decoder of the explainer, where the last layer is a 1×1 convolution layer with sigmoid activation, as shown in Fig. 1(b).

Note that in the loss of Eq. (2), the student only learns from the predictions (i.e., the final output layer) of the teacher.

[1]For each pixel, we consider its position as well as its channels to be with different importance scores.

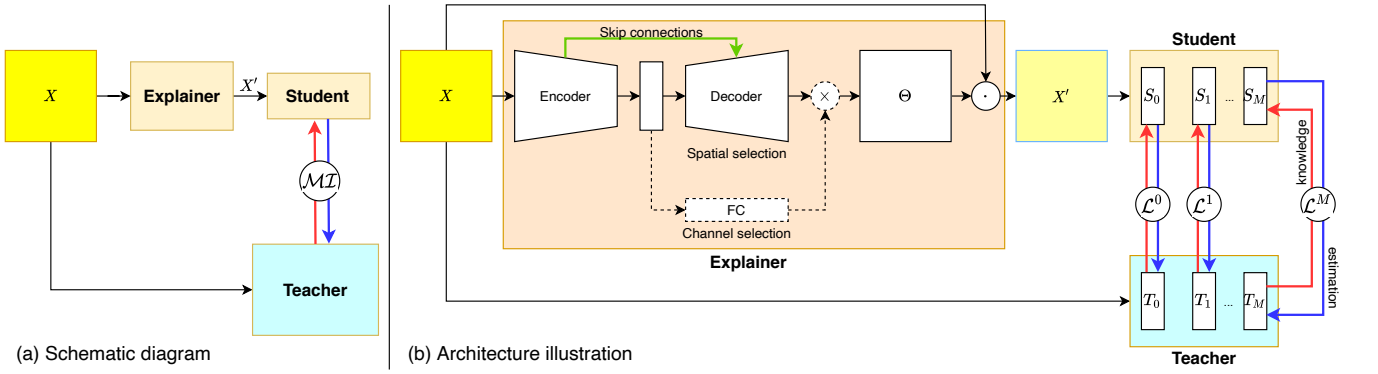


Fig. 1: (a) An overview of our framework, which consists of a fixed pretrained teacher, a learnable explainer and a learnable student. The explainer explains to the student by producing a simplified $\mathbf{X}'$ from input $\mathbf{X}$. The knowledge from teacher is transferred to the student by maximizing the mutual information (MI). (b) The detailed architecture of our framework.

Although our ultimate goal is to let the student generate the same predictions of the teacher, the knowledge in the intermediate layers of the teacher can also be informative to the learning of the student [16]. Inspired by the idea of knowledge distillation in [5], [16], [20], we therefore introduce an additional loss to maximize the mutual information between the outputs of each i^{th} intermediate layer of the teacher ($\mathcal{T}^i(\mathbf{X})$) and the student ($\mathcal{S}^i(\mathbf{X}')$):

$$\mathcal{L}^i = \min_{\mathcal{E},\mathcal{S}} -\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}'|\mathbf{X}}\left[\log r(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}'))\right]\right], \quad (4)$$

where $r(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}'))$ is a variational distribution used for approximating $\mathrm{p}(\mathcal{T}^i(\mathbf{X}) \mid \mathcal{S}^i(\mathbf{X}'))$, which is derived from information-theoretic perspective (see Eq. (12)).

Recall that the output of the i^{th} layer of the teacher is a $C^i \times H^i \times W^i$ feature map (note that the output of the i^{th} layer of the student is of the same spatial dimension but with a smaller number of channels). Following [20], we model $\mathcal{T}^i(\mathbf{X})$ as the following Gaussian distribution conditioned on $\mathcal{S}^i(\mathbf{X}')$:

$$r\left(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}')\right) \sim \prod_{c=1,h=1,w=1}^{C^i,H^i,W^i} \mathcal{N}\left(\mu^i(\mathcal{S}^i(\mathbf{X}'))_{c,h,w}, \sigma_c^{i^2}\right), \quad (5)$$

where μ^i is a subnetwork with 1×1 convolutional layers to match the channel dimensions between $\mathcal{T}^i(\mathbf{X})$ and $\mathcal{S}^i(\mathbf{X}')$, $\mu^i_{c,h,w}$ is a single output unit, and $\sigma_c^{i^2}$ is the learnable parameter specific to each channel at the i^{th} layer. For $\sigma_c^{i^2}$, we exploit the softplus function $\sigma_c^{i^2} = \log(1 + e^{\alpha_c^i}) + \epsilon$ where α_c^i is a learnable parameter and ϵ is used for numerical stability.

With Eq. (5), we can write Eq. (4) as:

$$\mathcal{L}^i = \min_{\mathcal{E},\mathcal{S}} \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}'|\mathbf{X}}\left[\sum_{c=1,h=1,w=1}^{C^i,H^i,W^i} \log \sigma_c^i + \frac{(\mathcal{T}^i(\mathbf{X})_{c,h,w} - \mu^i(\mathcal{S}^i(\mathbf{X}'))_{c,h,w})^2}{2\sigma_c^{i^2}} + \text{const.}\right]\right]. \quad (6)$$

Finally, the overall loss function of our framework can be written as

$$\mathcal{L} = \mathcal{L}^M + \lambda \sum_{i=1}^{M-1} \mathcal{L}^i, \quad (7)$$

where λ is the weight of the losses of the intermediate layers.

B. Derivation from information-theoretic perspective

Previously, we have shown that the objective function of our proposed framework has intuitive interpretations. Here we additionally demonstrate that the objective function can be derived in a theoretical way with mutual information, which is a widely-used measure of the dependence between two random variables and captures how much knowledge of one random variable reduces the uncertainty about the other [31]. In particular, we note: minimizing the training losses in Eq. (2) and Eq. (4) are equal to maximizing the following mutual information: $I(\mathbf{X}'; \boldsymbol{y}^{\mathcal{T}})$ and $I(\mathcal{T}^i(\mathbf{X}); \mathcal{S}^i(\mathbf{X}'))$, respectively.

$$\max_{\mathcal{E},\mathcal{S}} \; I(\mathbf{X}'; \boldsymbol{y}^{\mathcal{T}}) + \lambda \sum_{i=1}^{M-1} I(\mathcal{T}^i(\mathbf{X}); \mathcal{S}^i(\mathbf{X}')). \quad (8)$$

Given the definition of mutual information, the first term of Eq. (8) can be derived as:

$$\begin{aligned} I(\mathbf{X}'; \boldsymbol{y}^{\mathcal{T}}) &= \mathbb{H}(\boldsymbol{y}^{\mathcal{T}}) - \mathbb{H}(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}') \\ &= \mathbb{E}_{\mathbf{X}}\mathbb{E}_{\mathbf{X}'|\mathbf{X}}\mathbb{E}_{\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}'}[\log p(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')] + Const. \end{aligned} \quad (9)$$

In general, it is impossible to compute expectations under the conditional distribution of $p(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')$. Hence, we define a variational distribution $q(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')$ that approximates $p(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')$:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}'}[\log p(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')] \\ &\quad = \mathbb{E}_{\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}'}[\log q(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')] \\ &\qquad + \mathbb{D}_{KL}[q(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')||p(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')] \\ &\quad \geq \mathbb{E}_{\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}'}[\log q(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')], \end{aligned} \quad (10)$$

where $\mathbb{D}_{KL}$ is the KullbackLeibler divergence and equality holds if and only if $q(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')$ and $p(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')$ are equal in distribution. Note that it is not hard to show that our student corresponds to the variational distribution q.

For the second term of Eq. (8), we have:

$$\begin{aligned} &I(\mathcal{T}^i(\mathbf{X}); \mathcal{S}^i(\mathbf{X}')) = \mathbb{H}(\mathcal{T}^i(\mathbf{X})) - \mathbb{H}(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}')) \\ &= \mathbb{E}_{\mathbf{X}}\mathbb{E}_{\mathbf{X}'|\mathbf{X}}\mathbb{E}_{\mathcal{T}^i|\mathbf{X},\mathcal{S}^i|\mathbf{X}'}[\log p(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}'))] + Const \end{aligned} \quad (11)$$

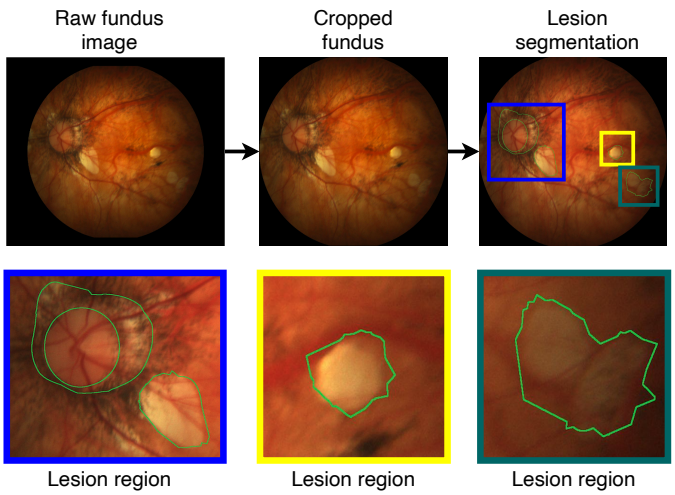


Fig. 2: An example in fundus dataset. Fine-grained lesion regions are inside the contour of the three images in the 2nd row, which are the zoom-in versions of the three lesion regions identified in the 3rd column of the 1st row.

TABLE I: Teacher and student model architecture.

Teacher	Student
3×3 conv, 32, pad=1; ReLU [b1]	3×3 conv, 2, pad=1; ReLU [b1]
2×2 max pooling	2×2 max pooling
3×3 conv, 64, pad=1; ReLU [b2]	3×3 conv, 4, pad=1; ReLU [b2]
2×2 max pooling	2×2 max pooling
3×3 conv, 128, pad=1; ReLU [b3]	3×3 conv, 8, pad=1; ReLU [b3]
2×2 max pooling	2×2 max pooling
3×3 conv, 256, pad=1; ReLU [b4]	3×3 conv, 16, pad=1; ReLU [b4]
fully connected layer	fully connected layer
softmax	softmax

Given Eq. (11), we can derive the following formula, similar to Eq. (10):

$$\begin{aligned}&\mathbb{E}_{\mathcal{T}^i|\mathbf{X},\mathcal{S}^i|\mathbf{X}'}[\log p(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}'))]\\ &\geq\mathbb{E}_{\mathcal{T}^i|\mathbf{X},\mathcal{S}^i|\mathbf{X}'}[\log r(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}'))],\end{aligned} \tag{12}$$

where r is the variational distribution to approximate the conditional distribution.

By using the two variational distributions q and r, the problem (8) can be relaxed to Eq. (13), i.e. maximizing the variational lower bounds.

$$\max_{\mathcal{E},\mathcal{S}} \mathbb{E}[\log q(\boldsymbol{y}^{\mathcal{T}}|\mathbf{X}')] + \lambda \sum_{i=1}^{M-1} \mathbb{E}[\log r(\mathcal{T}^i(\mathbf{X})|\mathcal{S}^i(\mathbf{X}'))]. \tag{13}$$

IV. Experiments

In this section, we present the experiments conducted on real-world datasets to exam the performance of the proposed MED-TEX against the state-of-the-art methods.

A. Architectures and settings of MED-TEX

For the teacher and student, we adopt a deep architecture with 4 block CNN layers, shown in Table I. It is important to note that with less number of filters, the size of the student model is much (226 times) smaller than the teacher, i.e., 1.7k parameters of the student versus 390.5k parameters of the teacher. We pretrained the teacher on the training set, which achieves 96.33% accuracy, 0.964 precision, 0.963 recall and 0.96 F1 score on the testing data.

TABLE II: Explainer model architecture.

Encoder
2× (3×3 conv, 32, pad=1; Batch Norm; ReLU) [e0]
2×2 max pooling; 2× (3×3 conv, 64, pad=1; batch norm; ReLU) [e1]
2×2 max pooling; 2× (3×3 conv, 128, pad=1; batch norm; ReLU) [e2]
2×2 max pooling; 2× (3×3 conv, 256, pad=1; batch norm; ReLU) [e3]
2×2 max pooling; 2× (3×3 conv, 512, pad=1; batch norm; ReLU) [e4]
2×2 max pooling; 2× (3×3 conv, 512, pad=1; batch norm; ReLU) [e5]
Decoder
2×2 nearest upsample; 2× (3×3 conv, 512, pad=1; Batch Norm; ReLU) [d4]
concatenate [d4, e4]; 2×2 nearest upsample; 2× (3×3 conv, 256, pad=1; Batch Norm; ReLU) [d3]
concatenate [d3, e3]; 2×2 nearest upsample; 2× (3×3 conv, 128, pad=1; Batch Norm; ReLU) [d2]
concatenate [d2, e2]; 2×2 nearest upsample; 2× (3×3 conv, 128, pad=1; Batch Norm; ReLU) [d1]
concatenate [d1, e1]; 2×2 nearest upsample; 2× (3×3 conv, 64, pad=1; Batch Norm; ReLU) [d0]
1×1 conv, 3, pad=1; sigmoid [output]

TABLE III: Abbreviation of the compared methods

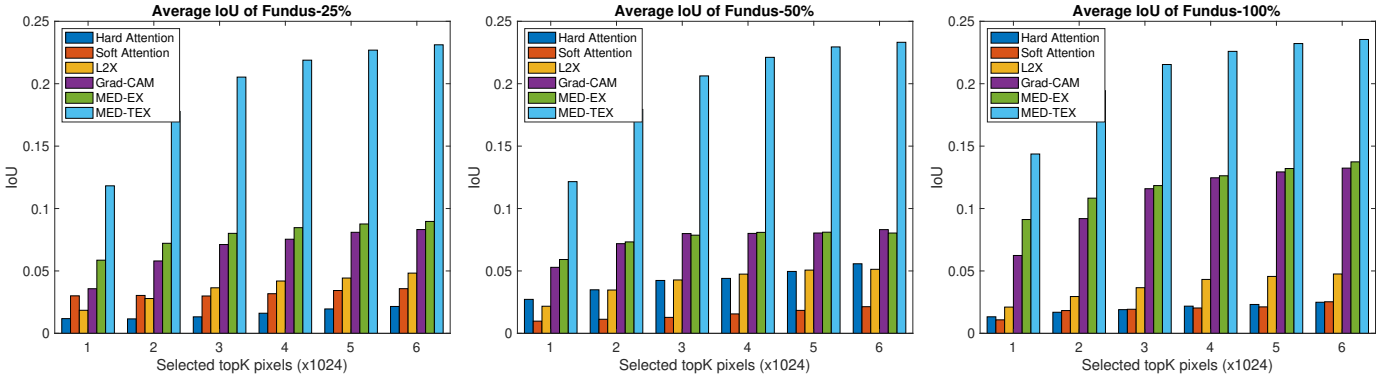
Method	Abbreviation
Resnet18 + hard attention using Gumbel-softmax [9]	Hard attention
Resnet18 + soft attention using [10], [11]	Soft attention
Resnet18 + patch-based image selection using Gumbel-softmax [8]	L2X
Resnet18 + grad class activation map [6], [7]	Grad-CAM
Our student without explainer	Student (only)
Our Transfer and EXplain framework using only explain $\mathcal{L}^M$ (Eq. 1) loss	MED-EX
Our full Transfer and EXplain framework	MED-TEX

For the explainer, we adopt the Unet architecture [2], which takes an image $\mathbf{X}$ as input and outputs the selection score $\boldsymbol{\Theta}$, shown in Table II. Note that of $\boldsymbol{\Theta} \in \mathbb{R}^{C\times H\times W}$ can be decomposed into a $1 \times H \times W$ tensor that models the spatial selection and a $C \times 1$ tensor that models the channel selection. The spatial selection tensor is the output from the decoder and the channel selection tensor is generated by passing the output of the encoder (i.e., [e5] in Table II) through a fully connected neural network with sigmoid activation function. Finally, $\boldsymbol{\Theta}$ is obtained by matrix multiplication between the spatial and channel selection tensors.

There are four $\mathcal{L}^i$ losses for the intermediate layers of the teacher and student, i.e., convolutional blocks b1, b2, b3 and b4 in Table I. Each $\mathcal{L}^i$ consists of a subnetwork μ^i and learnable scalar α^i_c. The subnetwork of μ^i consists of a 1×1 convolutional layer (with 16, 32, 64 and 128 numbers of filters respectively), ReLU activation, and a 1×1 convolutional layer (with 32, 64, 128 and 256 numbers of filters respectively). We empirically set $\lambda = 0.001$.

TABLE IV: Post-hoc evaluation on the fundus dataset.

	Fundus-25%				Fundus-50%				Fundus-100%			
Method	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Hard attention	0.895	0.984	0.849	0.912	0.918	0.974	0.896	0.934	0.928	0.955	0.933	0.944
Soft attention	0.895	0.982	0.852	0.912	0.915	0.937	0.93	0.933	0.948	0.975	0.943	0.959
L2X	0.863	0.839	0.974	0.901	0.931	0.964	0.927	0.945	0.94	0.94	0.981	0.961
Grad-CAM	0.891	0.959	0.867	0.911	0.921	0.964	0.911	0.937	0.921	0.918	0.963	0.940
Student (only)	0.863	0.903	0.813	0.856	0.90	0.916	0.879	0.897	0.927	0.960	0.890	0.923
MED-EX	0.908	0.958	0.894	0.925	0.938	0.978	0.924	0.950	0.951	0.994	0.93	0.961
MED-TEX	**0.915**	0.961	0.904	**0.933**	**0.955**	0.984	0.946	**0.964**	**0.975**	0.989	0.972	**0.98**



(a) Fundus-25% (b) Fundus-50% (c) Fundus-100%

Fig. 3: Average IoU evaluation among various methods at different topKs.

TABLE V: Average IoU evaluation when top$K$ is equal to the number of ground-truth lesion pixels for every individual image.

	Fundus-100%	Fundus-50%	Fundus-25%
MED-EX	0.091	0.06	0.058
MED-TEX	0.405	0.313	0.304

B. Datasets

We conducted our experiment on a fundus dataset collected from two sources: Baidu iChallenge[2] and Cao Thang International Eye Hospital (CTEH)[3]. The dataset consists of two kinds of fundus images: the ones with pathological myopia[4] (abnormal images) and the normal ones without the disease. Fig. 2 shows the data processing procedure. Given a raw image with pathological myopia, it was firstly preprocessed by cropping off the background. Next, the lesion regions of the image were identified and segmented by medical experts. For normal images, they were collected from the CTEH electronic health record. Finally, we have 1873 images in total, which consists of 1073 normal and 800 abnormal images. For the abnormal images, there are 200 of them with fine-grained lesion segmentations.

Originally, these images are in various sizes, so we rescaled all of them to $3\times256\times256$ (3 is the number of channels) and normalized their values in the range between 0 and 1. We split the dataset into the training (773 normal and 500 abnormal images) and testing (300 normal and 300 abnormal images) sets. It is noteworthy that all the 200 images with lesion segementaions are in the testing set. To mimic the case where we have less data to learn from and explain the teacher, we further reduce the number of training images for MED-TEX, i.e., 25%, 50% and 100% training images are used, denoted as Fundus-25%, Fundus-50% and Fundus-100%, respectively.

In addition to the fundas dataset, we also conduct our experiments on the Tiny ImageNet dataset[5], whose settings are shown in the Appendix.

C. Compared methods

There is no existing method with the same problem setting as ours. So we compare our MED-TEX with the representative model interpretation methods that can be adapted to our scenario, including "hard" attention using Gumbel-softmax trick [9], "soft" attention [10], [11], Grad-CAM [6], [7] and L2X [8]. In particular, for all the compared methods, we leveraged ResNet18 [3] without fully connected layers for feature extraction, which are further input into those methods to generate model interpretations. The "hard" attention using Gumbel-softmax trick [9] discretely samples the feature domain extracted by ResNet18, followed by a fully connected layer with softmax to produce predictions. In the same context, "soft" attention [10], [11] and Grad-CAM [6], [7] also perform feature selection on the feature domain. All these three models are trained by the cross-entropy loss with the labels being generated by the teacher model. We adapt L2X [8], which has not been carefully studied for image classification, by using ResNet18 for the explainer and a similar student architecture

[2]https://ichallenge.baidu.com
[3]http://cteyehospital.com
[4]It is an eye disease that causes distant objects to be blurry.
[5]https://www.kaggle.com/c/tiny-imagenet

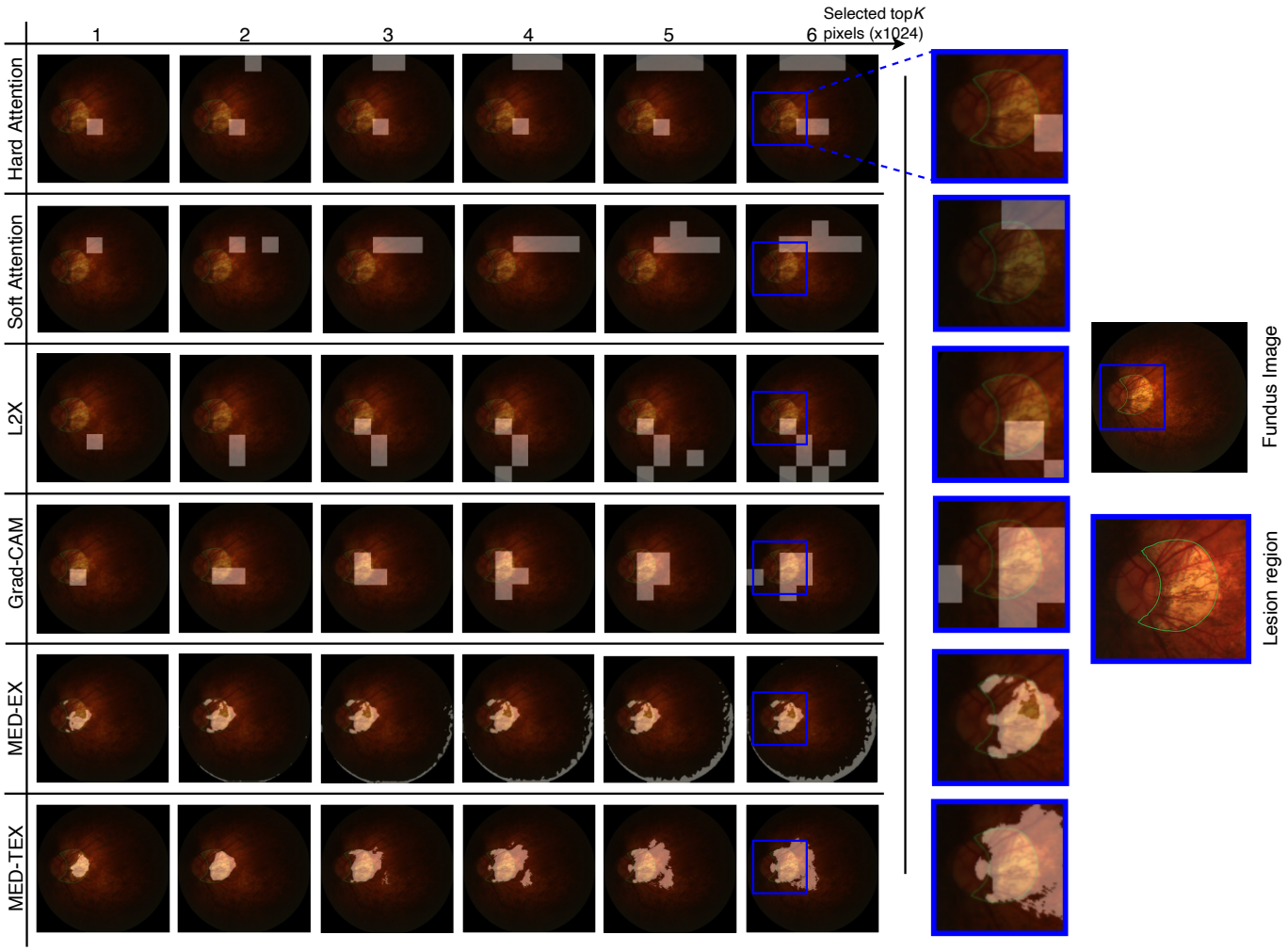


Fig. 4: Visualization results of top$K$ highlighted image regions of different methods trained on Fundus-100%, compared with the ground-truth lesion segmentation (specified by the green contour). While hard attention [9], soft attention [10], [11], Grad-CAM [7], and L2X [8] output patch-based region selection maps, our MED-EX and MED-TEX give pixel-level selection scores which is more accurate and fine-grained than others.

as ours. Note that L2X is trained by minimizing L^M only, without the intermediate losses.

To demonstrate the effectiveness of knowledge transformation of the intermediate layers, we compare MED-TEX with its variant without information transfer losses (Eq. 4), denoted as MED-EX. The loss for training MED-EX is the same to L2X, but the ways of constructing the explainer are totally different in the two models. To illustrate the importance of explainer, we also consider another variant, i.e., the student without the explainer, which was trained on the raw input image X. We summarize all these comparison methods and their abbreviations in Table III. All models are trained by using Adam with 0.001 learning rate and a batch size of 64 on an NVIDIA RTX Titan GPU with 24GB memory.

D. Evaluation metrics

Post-hoc metric: To evaluate our MED-TEX, we use post-hoc metric [8] which compares the predictive distributions of the student given $\mathbf{X}'$ and the teacher given $\mathbf{X}$. In other words, we compute accuracy, precision, recall and F1 score of the outputs from different methods against the output of the pretrained teacher on the testing dataset.

Intersection over Union (IoU): We compare Intersection over Union (IoU) between the highlighted image regions and the ground-truth lesion segmentation of abnormal images. Note that hard attention, soft attention, Grad-CAM, and L2X output patch-based region selection maps (the ResNet18 feature extraction outputs a feature map of 8×8 spatial size each of which is corresponding to a 32×32 region in image domain), while our MED-EX and MED-TEX give pixel-level selection scores. For a better comparison, we rank feature scores and select the number of pixels corresponding to the top K highest scores (e.g., top$K$ $\in \{k \times 32 \times 32 \mid k = 1, 2, 3, 4, 5, 6\}$):

$$IoU_{topK} = 2\frac{\mathbf{\Theta}_{topK} \bigcap \mathbf{X}_{lesion}}{\mathbf{\Theta}_{topK} \bigcup \mathbf{X}_{lesion}}, \tag{14}$$

where $\mathbf{\Theta}_{topK}$ indicates the selected pixels corresponding to the top$K$ feature scores and $\mathbf{X}_{lesion}$ denotes ground-truth lesion segmentation pixels.

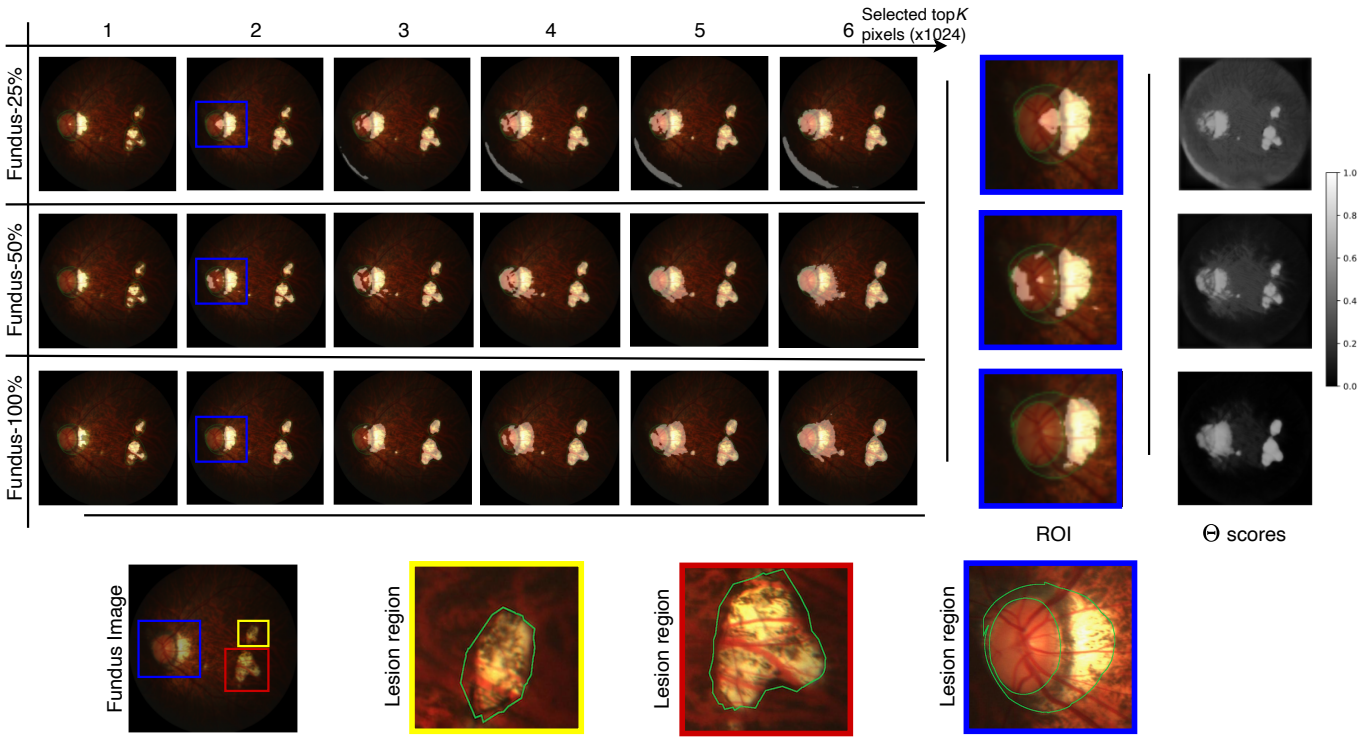


Fig. 5: Visualization results of top$K$ highlighted image regions of different methods with different number of training data, compared with the ground-truth lesion segmentations (specified by the green contours). Feature selection scores Θ are plot in heatmaps on the right.

E. Results

In order to compare our MED-TEX to other methods, we first use post-hoc metric [8]. Our method consistently outperforms hard attention using Gumbel-softmax trick [9], soft attention [10], [11], Grad-CAM [6], [7], and learning to explain [8] in term of both accuracy and F1 score, as shown in Table IV. Especially, for our proposed method, MED-TEX, it achieves reasonably good results on approximating the teacher with only 25% data. In addition, when the full dataset is used, MED-TEX reaches 0.98 F1 score, meaning that the student can perform nearly as well as the teacher on the image classification task.

If we compare MED-TEX with its variant, Student (only), it can be observed that Student (only) trained directly from raw input images cannot perform well. This suggests that the explainer with feature selection at pixel-level plays a central role to guide the student to achieve better performance.

Fig. 3 shows the IoU results in bar charts. We can see that our MED-TEX achieves significantly higher IoU than others on fundus dataset. Specifically, MED-TEX performs approximately 2$\times$ better than MED-EX and Grad-CAM and more than 4$\times$ better than others. We also observe that even our approach is trained with small amount data of fundus dataset, the IoU is still remaining relatively high (e.g., when top$K$ is equal to 1024, MED-TEX achieves 0.118 IoU in fundus-25%). In addition, we further evaluate the performance of MED-EX and MED-TEX when top$K$ is equal to the number of ground-truth lesion pixels for each individual image. Table V reports the average IoU results, where MEX-TEX achieves 0.4, 0.31 and 0.3 for Fundus-100%, Fundus-50% and Fundus-25%, respectively.

Fig. 4 shows the visualization results of top$K$ highlighted image regions of different methods. Hard attention [9], soft attention [10], [11], Grad-CAM [6], [7], and L2X [8] can only give patch-based region selection maps, while our MED-EX and MED-TEX produces pixel-level selection scores. It can be seen from Fig. 4 that our method on an abnormal fundus image highlights the regions that well match the ground-truth lesion segmentations. In general, MED-EX and MED-TEX produce more accurate and fine-grained lesion segmentations on the pixel level than those segmentations on the feature level in the other methods. Moreover, MED-TEX clearly outperforms MED-EX due to the use of the intermediate knowledge distillation losses.

We also qualitatively evaluate our method with different proportions of the training data in Fig. 5, where the example fundus image has multiple lesion regions (red, blue and yellow regions). We can see that MED-TEX is able to precisely point out these regions, even trained with less data. With more training data used, our method gradually improves the quality and accuracy of identifying the lesion regions. Finally, we also evaluate MED-TEX using Tiny ImageNet dataset with the same settings, whose results are shown in the appendix. It can also observed the same trend that our method consistently outperforms the other compared approaches.

V. Conclusion

In this paper, we have introduced our novel framework MED-TEX, which is a joint knowledge distillation and model

TABLE VI: Post-hoc evaluation on the Tiny ImageNet dataset

	25%		50%		100%	
Method	Acc	F1	Acc	F1	Acc	F1
Hard attention	0.92	0.923	0.907	0.901	0.966	0.968
Soft attention	0.913	0.91	0.935	0.94	0.953	0.955
L2X	0.833	0.828	0.86	0.859	0.87	0.87
Grad-CAM	0.90	0.906	0.92	0.918	0.953	0.953
Student (only)	0.686	0.711	0.80	0.779	0.826	0.839
MED-EX	0.913	0.912	0.933	0.932	0.953	0.955
MED-TEX	**0.927**	**0.928**	**0.94**	**0.943**	**0.967**	**0.969**

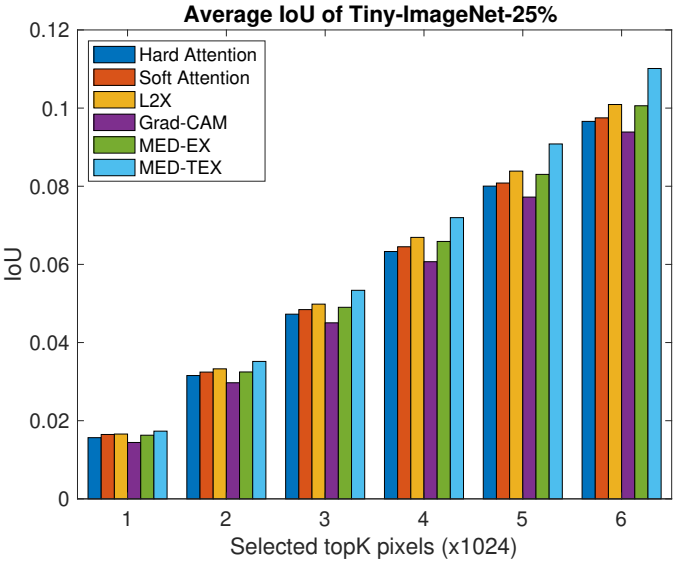


Fig. 6: Average IoU evaluation among various methods.

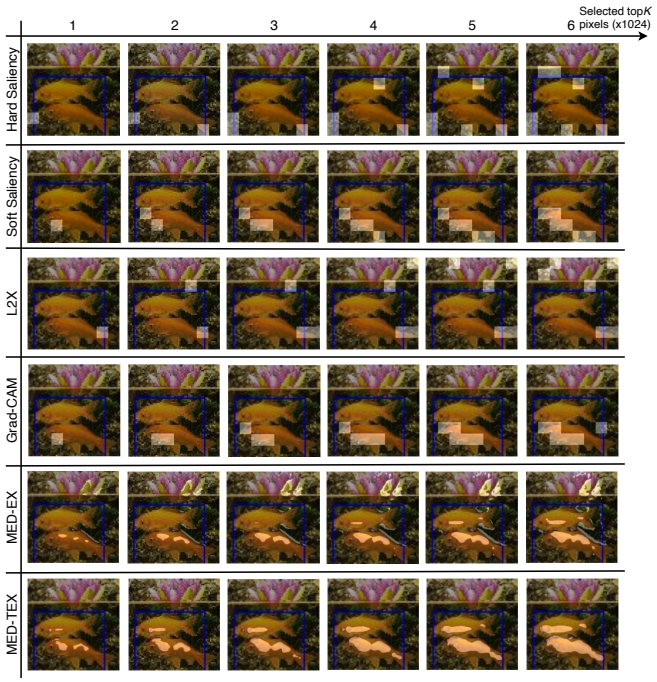


Fig. 7: Visualization results on the example golden fish image.

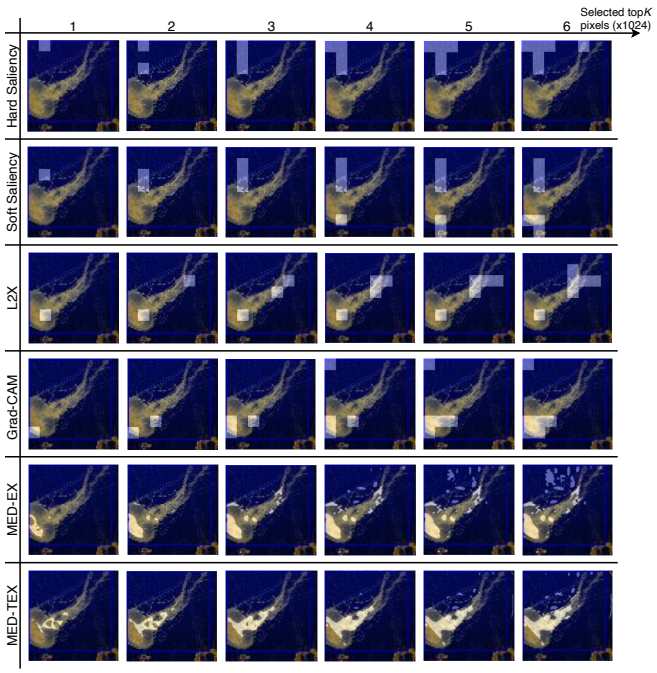


Fig. 8: Visualization results on the example jelly fish image.

interpretation framework that learns the significantly smaller student (compared to the teacher) and explainer models by leveraging the knowledge only from the pretrained teacher model. With the proposed framework, we can tackle two important issues in medical imaging: the lack of training data and the lack of interpretation. Specifically, the student is trained with less data to learn from the knowledge of pretrained teacher with the assistance of the explainer designed to highlight the important image areas to the teacher's predictions. The output of the explainer can also be used as low-level strong annotations trained by high-level weak ones (teacher's knowledge). In addition, to train the framework, we have proposed to maximize the mutual information between the intermediate and output layers of the student and teacher, which forms a novel training objective of our framework. In our experiment, we show that MED-TEX outperforms several widely-used knowledge distillation and model interpretation techniques, including: soft attention [10], [11], hard attention [9], L2X [8], Grad-CAM [6], [7] on the fundus dataset in terms of both quantitative and qualitative evaluations. In the future work, we would like to apply our framework to COVID-19 analysis, aiming to identify the most important regions of lung X-ray images that are related to COVID-19.

APPENDIX

In this section, we demonstrate that our framework works not only on medical imaging datasets but also on natural imaging datasets. We select 500 golden fish images (425 for training and 75 for testing) and 500 jellyfish images (also 425 for training and 75 for testing). All the images are with labeled bounding boxes identifying the regions of the fishes. Examples of the images are shown in Fig. 7 and 8. The teacher model is then trained by 850 images and tested with 250 images that reaches 95.4% accuracy (0.947 precision, 0.956 recall and 0.954 F1 score). In the post-hoc evaluation, MED-TEX gives higher F1 score and accuracy than the other methods. There is also a significant improvement of IoU with less data as shown

in Fig. 6. We qualitatively evaluate MED-TEX by visualizing the demo images (golden fish and jelly fish) and highlighted the regions with different top$K$s trained by Tiny ImageNet 100%, as shown in Fig. 7 and 8, respectively. It can be seen than our approaches are able to locate the fishes more precisely than other methods.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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

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

[4] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing mri reconstruction using a generative adversarial network with a cyclic loss," *Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1488–1497, 2018.

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

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of CVPR*. IEEE, 2016, pp. 2921–2929.

[7] 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 *Proceedings of ICCV*. IEEE, 2017, pp. 618–626.

[8] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," *arXiv preprint arXiv:1802.07814*, 2018.

[9] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of ICML*, 2015, pp. 2048–2057.

[12] T. Zhang, G. Lin, J. Cai, T. Shen, C. Shen, and A. C. Kot, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *Transactions on Multimedia*, vol. 21, no. 11, pp. 2930–2941, 2019.

[13] M. Izadyyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L. B. Moreira, J. Eschbacher, P. Nakaji, M. C. Preul, and Y. Yang, "Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images," in *Proceedings of MICCAI*. Springer, 2018, pp. 300–308.

[14] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," in *Proceedings of MICCAI*. Springer, 2017, pp. 568–576.

[15] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[16] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[17] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[18] V. Belagiannis, A. Farshad, and F. Galasso, "Adversarial network compression," in *Proceedings of ECCV*, 2018, pp. 0–0.

[19] S. Chen, C. Zhang, and M. Dong, "Coupled end-to-end transfer learning with generalized fisher information," in *Proceedings of CVPR*. IEEE, 2018, pp. 4329–4338.

[20] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of CVPR*. IEEE, 2019, pp. 9163–9171.

[21] H. Wang, D. Zhang, Y. Song, S. Liu, Y. Wang, D. Feng, H. Peng, and W. Cai, "Segmenting neuronal structure in 3D optical microscope images via knowledge distillation with teacher-student network," in *Proceedings of ISBI*. IEEE, 2019, pp. 228–231.

[22] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved ms lesion segmentation," in *Proceedings of ISBI*. IEEE, 2019, pp. 1563–1566.

[23] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource transfer learning with convolutional neural networks for lung pattern analysis," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 76–84, 2016.

[24] Q. Dou, Q. Liu, P. Heng, and B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation." *Transactions on Medical Imaging*, vol. 39, no. 7, 2020.

[25] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proceedings of WACV*. IEEE, 2018, pp. 839–847.

[26] H. Yang, J.-Y. Kim, H. Kim, and S. P. Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1306–1315, 2019.

[27] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[28] C. J. Maddison, D. Tarlow, and T. Minka, "A* sampling," in *Proceedings of NIPS*, 2014, pp. 3086–3094.

[29] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv preprint arXiv:1611.00712*, 2016.

[30] X. Ren, J. Huo, K. Xuan, D. Wei, L. Zhang, and Q. Wang, "Robust brain magnetic resonance image segmentation for hydrocephalus patients: Hard and soft attention," in *Proceedings of ISBI*. IEEE, 2020, pp. 385–389.

[31] T. M. Cover and J. A. Thomas, "Elements of information theory," 2012.