MDNet: A Semantically and Visually Interpretable
Medical Image Diagnosis Network

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, Lin Yang
University of Florida
`zizhao@cise.ufl.edu`

Abstract

The inability to interpret the model prediction in semantically and visually meaningful ways is a well-known shortcoming of most existing computer-aided diagnosis methods. In this paper, we propose MDNet to establish a direct multimodal mapping between medical images and diagnostic reports that can read images, generate diagnostic reports, retrieve images by symptom descriptions, and visualize attention, to provide justifications of the network diagnosis process. MDNet includes an image model and a language model. The image model is proposed to enhance multi-scale feature ensembles and utilization efficiency. The language model, integrated with our improved attention mechanism, aims to read and explore discriminative image feature descriptions from reports to learn a direct mapping from sentence words to image pixels. The overall network is trained end-to-end by using our developed optimization strategy. Based on a pathology bladder cancer images and its diagnostic reports (BCIDR) dataset, we conduct sufficient experiments to demonstrate that MDNet outperforms comparative baselines. The proposed image model obtains state-of-the-art performance on two CIFAR datasets as well.

1. Introduction

In recent years, the rapid development of deep learning technologies has shown remarkable impact on the biomedical image domain. Conventional image analysis tasks, such as segmentation and detection [2], support quick knowledge discovery from medical metadata to help specialists' manual diagnosis and decision-making. Automatic decision-making tasks (e.g. diagnosis) are usually treated as standard classification problems. However, generic classification models are not an optimal solution for intelligent computer-aided diagnosis, because such models conceal the rationale for their conclusions, and therefore lack the interpretable justifications to support their decision-making process. It is rather difficult to investigate how well the model captures and understands the critical biomarker information. A
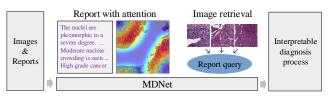

Figure 1: Overview of our medical image diagnosis network (MDNet) for interpretable diagnosis process.

model that is able to visually and semantically interpret the underlying reasons that support its diagnosis results is significant and critical (Figure 1).

In clinical practice, medical specialists usually write diagnosis reports to record microscopic findings from images to diagnose and select treatment options. Teaching machine learning models to automatically imitate this process is a way to provide interpretability to machine learning models. Recently, image to language generation [14, 22, 4, 33] and attention [36] methods attract some research interests.

In this paper, we present a unified network, namely MDNet, that can read images, generate diagnostic reports, retrieve images by symptom descriptions, and visualize network attention, to provide justifications of the network diagnosis process. For evaluation, we have applied MDNet on a pathology bladder cancer image dataset with diagnostic reports (Section 5.2 introduces dataset details). In bladder pathology images, changes in the size and density of urothelial cell nuclei or thickening of the urothelial neoplasm of bladder tissue indicate carcinoma. Accurately describing these features facilitates the accurate diagnosis and is critical for the identification of early-stage bladder cancer. The accurate discrimination of those subtle appearance changes is challenging even for observers with extensive experience. To train MDNet, we address the problem of directly mining discriminative image feature information from reports and learn a direct multimodal mapping from report sentence words to image pixels. This problem is significant because discriminative image features to support diagnostic conclusion inference is "latent" in reports rather than offered by specific image/object labels. Effectively utilizing

these semantic information in reports is necessary for effective image-language modeling.

For image modeling based on convolutional neural networks (CNNs), we address the capability of the network to capture size-variant image features (such as mitosis depicted in pixels or cell polarity depicted in regions) for image representations. We analyze the weakness of the residual network (ResNet) [6, 7] from the ensemble learning aspect and propose *ensemble-connection* to encourage multi-scale representation integration, which results in more efficient feature utilization according to our experiment results. For language modeling, we adopt Long Short-Term Memory (LSTM) networks [33], but focus on investigating the usage of LSTM to mine discriminative information from reports and compute effective gradients to guide the image model training. We develop an optimization approach to train the overall network end-to-end starting from scratch. We integrate the attention mechanism [36] in our language model and propose to enhance its visual feature alignment with sentence words to obtain sharper attention maps.

To our knowledge, this is the first study to develop an interpretable attention-based model that can explicitly simulate the medical (pathology) image diagnosis process. We perform sufficient experimental analysis with complementary evaluation metrics to demonstrate that MDNet can generate promising and reliable results, also outperforms well-known image captioning baselines [14] on the BCIDR dataset. In addition, we validate the state-of-the-art performance of the proposed image model belonging to MDNet on two public CIFAR datasets [18].

2. Related Work

Image and language modeling: Joint image and language modeling enables the generation of semantic descriptions, which provides more intelligible predictions. Image captioning is one typical of application [16]. Recent methods use recurrent neural networks (RNNs) to model natural language conditioned on image information modeled by CNNs [14, 33, 13, 38]. They typically employ pre-trained powerful CNN models, such as GoogLeNet [28], to provide image features. Semantic image features play a key role in accurate captioning [22, 4]. Many methods focus on learning better alignment from natural language words to provided visual features, such as attention mechanisms [36, 38, 37], multimodal RNN [22, 14, 4] and so on [24, 37]. However, in the medical image domain, pre-trained universal CNN models are not available. A complete end-to-end trainable model for joint image-sentence modeling is an attractive open question, and it can facilitate multimodal knowledge sharing between the image and language models.

Image-sentence alignment also encourages visual explanations for network inner workings [15]. Hence, attention mechanisms become particularly necessary [36]. We wit-

ness growing interests of its exploration to achieve the network interpretability [41, 27]. The full power of this field has vast potentials to renovate computer-aided medical diagnosis, but a dearth of related work exists. To date, [25] and [17] deal with the problem of generating disease keywords for radiology images.

Skip-connection: Based on the residual network (ResNet) [6], the new pre-act-ResNet [7] introduces identity mapping skip-connection [7] to address the network training difficulty. Identity mapping gradually becomes an acknowledged strategy to overcome the barrier of training very deep networks [7, 11, 39, 10]. Besides, skip-connection encourages the integration of multi-scale representations for more efficient feature utilization [21, 1, 35].

3. Image model

3.1. Residual networks

The identity mapping in the newest ResNet [7] is a simple yet effective skip-connection to allow the unimpeded information flow inside the network [29]. Each skip-connected computation unit is called a residual block. In a ResNet with L residual blocks, the forward output y_L from the l-th residual block and the gradient of the loss $\mathcal{L}$ w.r.t its input y_l is defined as

$$y_L = y_l + \sum_{m=l}^{L-1} \mathcal{F}_m(y_m), \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial y_l} = \frac{\partial \mathcal{L}}{\partial y_L}(1 + \frac{\partial}{\partial y_l} \sum_{m=l}^{L-1} \mathcal{F}_m(y_m)), \tag{2}$$

where $\mathcal{F}_m$ is composed by consecutive batch normalization [12], rectified linear units (ReLU), and convolution. Thanks to the addition scheme, the gradient (i.e. $\frac{\partial \mathcal{L}}{\partial y_L}$) in backward can flow directly to preceding layers without passing through any convolutional layer. Since the weights of convolutional layers can scale gradients, this property alleviates the gradient vanishing effect when the depth of the network increases [23, 7].

3.2. Decouple ensemble network outputs

One skip-connection in a residual block offers two information flow paths, so the total path increases exponentially as network goes deeper [11]. Recent work [32] shows that ResNet with n residual blocks can be interpreted as the ensemble of 2^n relatively shallow networks. It can be viewed that the exponential ensembles boost the network performance [32]. Consequently, this viewpoint reveals a weakness of ResNet by our probes into its classification module.

In ResNet and other related networks [7, 11, 19, 30], the classification module connecting convolutional layers includes a global average pooling layer and a fully connected layer. The two layers are mathematically defined as

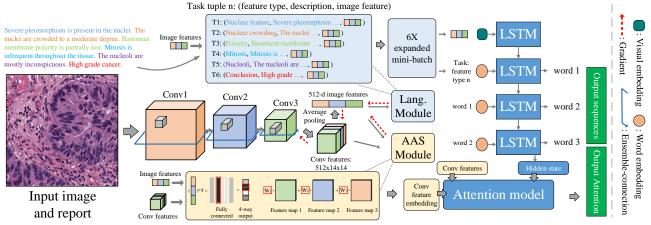$$p^c = \sum_k w_k^c \cdot \sum_{i,j} y_L^{(k)}(i,j), \tag{3}$$

Figure 2: Overall illustration of MDNet. We use a bladder image with its diagnostic report as an example. The image model generates an image feature to pass to LSTM in the form of a task tuple and a Conv feature embedding (for the attention model) computed by the AAS module (defined in the method). LSTM executes prediction tasks according to the specified image feature type (best viewed in color).

where p^c is the probability output of class c. (i,j) denotes spatial coordinates. $\boldsymbol{w}^c = [w_1^c, ..., w_k^c, ...]^T$ is the c-th column of the weight matrix of the fully connected layer applied onto p^c. $y_L^{(k)}$ is the k-th feature map of the last residual block. By plugging Eq. (1) into Eq. (3)[1], we can see that p^c is the weighted average of the summed ensemble output:

$$p^c = \sum_{i,j} \boldsymbol{w}^c y_L = \sum_{i,j} \boldsymbol{w}^c \Big(y_1 + \sum_{m=1}^{L-1} \mathcal{F}_m\Big). \quad (4)$$

In this paper, we argue that using a single weighting function in the classification module is suboptimal in this situation. This is because the outputs of all ensembles share classifiers such that the importance of their individual features are undermined. To address this issue, we propose to decouple the ensemble outputs and apply classifiers to them individually by using

$$p^c = \sum_{i,j} \Big(\boldsymbol{w}_1^c \cdot y_1 + \sum_{m=1}^{L-1} \boldsymbol{w}_{m+1}^c \cdot \mathcal{F}_m\Big). \quad (5)$$

Compared with Eq. (4), this equation assigns individual weight $\boldsymbol{w}_1^c$ to $\boldsymbol{w}_L^c$ for each ensemble output, which enables the classification module to independently decide the information importance from different residual blocks.

We propose a "redesign" of the ResNet architecture to realize the above idea, i.e., a new way to skip-connect a residual block, defined as follows:

$$y_{l+1} = \mathcal{F}_l(y_l) \otimes y_l, \quad (6)$$

where $\otimes$ is the concatenation operation. We define this skip-connection scheme as *ensemble-connection*. It allows outputs from residual blocks to flow through concatenated feature maps directly to the classification layer in parallel

---

[1]We omit the spatial coordinate (i,j) and feature map dimension changes from y_1 to $\mathcal{F}_L$ for brevity.

(see Figure 2), such that the classification module assigns weights to all network ensemble outputs and map them to the label space. It is straightforward to see that our design also ensures unimpeded information flow [7] to overcome the gradient vanishing effect.

We apply *ensemble-connection* between residual blocks connecting block groups where the feature map dimension changes (see Appendix A) and maintain the identity mapping for blocks inside a group[2]. *ensemble-connection* in nature integrates multi-scale representations in the last convolution layer. This multi-scaling scheme is essentially different from the skip output schemes used by [35, 1].

4. Language modeling and network training

4.1. Language model

For language modeling, we use LSTM [8] to model the diagnostic reports by maximizing the joint probability over sentences:

$$\log p(\boldsymbol{x}_{0:T}|I;\theta_L) = \sum_{t=0}^{T} \log p(\boldsymbol{x}_t|I, \boldsymbol{x}_{0:t-1};\theta_L), \quad (7)$$

where $\{\boldsymbol{x}_0, ..., \boldsymbol{x}_T\}$ are sentence words (encoded as one-hot vectors). The LSTM parameters θ_L are used to compute several LSTM internal states [8, 33]. According to [36], we integrate the "soft" attention mechanism into LSTM through a context vector $\boldsymbol{z}_t$ (defined as follows) to capture localized visual information. To make prediction, LSTM takes the output of last time step $\boldsymbol{x}_{t-1}$ along with hidden state $\boldsymbol{h}_{t-1}$ and $\boldsymbol{z}_t$ as inputs, and computes the probability of next word $\boldsymbol{x}_t$ as follows:

---

[2]Later on, we notice a new network, DenseNet [10], which ends up with an analogous solution (concatenation replacing addition). We argue that our solution is based on a different motivation and results in a different architecture. Nevertheless, this network can be viewed as a successfully validation of our ensemble analysis.

$$\boldsymbol{h}_t = LSTM(E\boldsymbol{x}_{t-1}, \boldsymbol{h}_{t-1}, \boldsymbol{z}_t),$$
$$p(\boldsymbol{x}_t | I, \boldsymbol{x}_{0:t-1}; \theta_L) \propto \exp(G_h \boldsymbol{h}_t), \quad (8)$$

where E is the word embedding matrix. G_h decodes $\boldsymbol{h}_t$ to the output space.

The attention mechanism dynamically computes a weight vector to extract partial image features supporting the word prediction, which is interpreted as an attention map indicting where networks capture visual information. Attention is the main component supporting the visual interpretability of our network. In practice, we observe that the original attention mechanism [36] is more difficult to train, which often generates attention maps that smoothly highlight the majority of image area.

To address this issue, we propose an auxiliary attention sharpening (AAS) module to improve its learning effectiveness. The attention mechanism can be viewed as a type of alignment between image space and language space. As indicted by [20], improving such alignment can be achieved by adding supervision on attention maps by using region-level labels (e.g. bounding boxes). In order to deal with datasets that do not have any region-level labels, a new method needs to be developed. In our approach, rather than putting direct supervision on the weight vector $\boldsymbol{a}_t$, we propose to tackle this problem by utilizing the implicit class-specific localization property of global average pooling [40] to support image-language alignment. Overall, $\boldsymbol{z}_t$ can be computed as follows:

$$\boldsymbol{a}_t = softmax(W_{att} \tanh(W_h \boldsymbol{h}_{t-1} + \boldsymbol{c})),$$
$$\boldsymbol{c} = (\boldsymbol{w}^c)^T \mathcal{C}(I), \quad (9)$$
$$\boldsymbol{z}_t = \boldsymbol{a}_t \mathcal{C}(I)^T,$$

where W_{att} and W_h are learned embedding matrices. $\mathcal{C}(I)$ denotes Conv feature maps with dimension $512 \times (14 \cdot 14)$ generated by the image model. $\boldsymbol{c}$ denotes a 196-dimensional Conv feature embedding through $\boldsymbol{w}^c$.

The original attention mechanism learns $\boldsymbol{w}^c$ inside LSTM implicitly. In contract, AAS adds an extra supervision (defined in Section 4.2) to explicitly learn to provide more effective attention model training. Specifically, the formulation of this supervision is a revisit of Eq. (4) ($\mathcal{C}(I)$ stands for y_L; we use different notations for consistence). $\boldsymbol{w}^c$ is a 512-dimensional vector corresponding to the c-th column of the fully connected weight matrix, selected by assigned class c (see Figure 2); when applied to $\mathcal{C}(I)$, the obtained $\boldsymbol{c}$ that carries class-specific and localized region information is used to learn the alignment with $\boldsymbol{h}_{t-1}$ and compute a (14×14)-dimensional $\boldsymbol{a}_t$ and a 512-dimensional context vector $\boldsymbol{z}_t$. Figure 3 compares the qualitative results between the original method and our proposed method.

4.2. Effective gradient flow

In the well-known image captioning scheme [14, 13], a CNN provides an encoded image feature $F(I)$ as the LSTM

Figure 3: The attention maps of the original method (middle) and our method (right). Our method generates more focal attention on informative (urothelial) regions.

input $\boldsymbol{x}_0$. Then a special START token is used as $\boldsymbol{x}_1$ to inform the start of prediction. Generating effective gradients w.r.t $F(I)$ is the key for the image model optimization.

A complete medical diagnostic report describes multiple symptoms of observing images, followed by the diagnostic conclusion about either one or multiple type of diseases. For example, radiology images have multiple disease labels [25]. Each symptom description specifically describes one type of image (symptom) feature. Effectively utilizing the semantic information in different descriptions is critical to generate effective gradient w.r.t $F(I)$ by LSTM.

In our method, we let one LSTM focus on mining discriminative information from a specific description. All description modeling shares LSTM. In this way, the modeling of each image feature description becomes a function of the complete report generation. We denote the number of functions as K. In the training stage, given a mini-batch with B pairs of image and reports, after forwarding the mini-batch to the image model, we duplicate each sample inside, resulting in a $K \times B$ mini-batch as the input of LSTM. Each duplication takes shared image features and one of K types particular feature description extracted from the report (see Figure 2). The LSTM inputs of $\boldsymbol{x}_0^e$ and $\boldsymbol{x}_1^e$ are defined as

$$\boldsymbol{x}_0^e = W_F F(I), \quad \boldsymbol{x}_1^e = ES(e), \quad (10)$$

where W_F is a learned image feature embedding matrix. $S(e)$, $e = \{1, ..., K\}$ is the one-hot representation of the e-th image feature type. In this way, we use particular $\boldsymbol{x}_1^e$ to inform LSTM the start of a targeting task. During back-propagation, the gradients w.r.t $F(I)$ from duplications are merged. All the operations are end-to-end trainable.

To train AAS, we use the diagnostic conclusion as labels. The motivation are two-fold. First, the Conv feature embedding generated by AAS is specific to conclusion labels. Since all symptom descriptions support the inference of conclusion labels, it in nature contains necessary visual information to support different types of symptom descriptions and thereby can facilitate better alignment with description words in the attention model. Second, AAS serves as an extra supervision on the image model, which makes sure the image model training towards to optimal diagnostic conclusion.

4.3. Network optimization

The overall model has three sets of parameters: θ_D in the image model D, θ_L in the language model L, and θ_M in the AAS module M. The overall optimization problem in MDNet is defined as

$$\max_{\theta_L, \theta_D, \theta_M} \mathcal{L}_M(l_c, M(D(I; \theta_D); \theta_M)) \\ + \mathcal{L}_L(l_s, L(D(I; \theta_D); \theta_L)), \quad (11)$$

where $\{I, l_c, l_s\}$ is a training tuple: input image I, label l_c and groundtruth report sentence l_s. Modules M and L are supervised by two negative log-likelihood losses $\mathcal{L}_M$ and $\mathcal{L}_L$, respectively.

The updating processes of θ_M and θ_L are independent and straightforward using gradient descent. Updating θ_D involves the gradients from both modules. We develop a backpropagation scheme to allow their composite gradients co-adapted mutually. Compared with [5], the gradients in our method is calculated based on a mixture of a recurrent generative network and a multilayer perceptron. Specifically, θ_D is updated as follows:

$$\theta_D \leftarrow \theta_D - \lambda \cdot \left((1 - \beta) \cdot \frac{\partial \mathcal{L}_M}{\partial \theta_D} + \beta \cdot \eta \frac{\partial \mathcal{L}_L}{\partial \theta_D} \right), \quad (12)$$

where λ is the learning rate, and β dynamically regulates two gradients during the training process. We also introduce another factor η to control the scale of $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$, because $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$ often has smaller magnitude than $\frac{\partial \mathcal{L}_M}{\partial \theta_D}$. We will analyze the detailed configuration of these two hyperparameters and demonstrate the advantages of our proposed strategy.

5. Experimental Results

In this section, we start by validating the proposed image model (denoted as EcNet and explained in Section 3) of MDNet on the two CIFAR datasets that is specific for image recognition, with the purpose to show its superior performance against several other CNNs. Then, we conduct sufficient experiments to validate the proposed full MDNet for medical image and diagnostic report modeling on the BCIDR dataset. Our implementation is based on Torch7 [3]. Please refer to Appendix for complete details.

5.1. Image recognition on CIFAR

We use well-known CIFAR-10 and CIFAR-100 [18] to validate our proposed EcNet. We follow the common way [7] to process data and adopt the learning policy suggested by wide-ResNet (WRN) [39]. To choose baseline ResNet architectures, we consider depth as well as width to trade-off the memory usage and training efficiency [39]. We adopt the bottleneck residual block design instead of the "tubby"-like block with two 3×3 convolution layers used by WRN, since we observe the former offers consistent improvement. We hypothesize that it is because the bottleneck
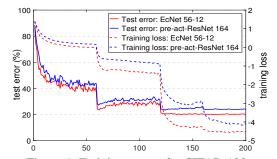

Figure 4: Training curves for CIFAR-100.

Method	D-W	Params	C-10	C-100
NIN [19]	-	-	8.81	35.67
Highway [29]	-	-	7.72	32.39
ResNet [6]	110	1.7M	6.43	25.16
ResNet$^+$ [7]	164	1.7M	5.46	24.33
ResNet$^+$ [7]	1001	10.2M	4.92	22.71
WRN [39]	40-4	8.7M	4.53	21.18
EcNet	110-4	1.8M	4.91	22.53
EcNet	56-12	8.0M	**4.43**	**19.94**

Table 1: The error rate (%) on CIFAR-10 (C-10) and CIFAR-100 (C-100). ResNet$^+$ denotes pre-act-ResNet. The second column indicates network Depth-Width. Our result is tested on one trial.

design compacts information of feature maps doubled by *ensemble-connection* (due to its concatenation operation), which promotes more efficient feature usage. Detailed architecture illustration is provided in Appendix A.

Since this experiment is not the main focus of this paper, we left full architecture exploration for future work. We present two variants having similar number of parameters with compared variants of ResNet and WRN. The first one has depth 110 and width 4 and the second has depth 56 and width 12. Table 1 compares the error rate on two datasets and Figure 4 compares the training curves. Our EcNet-56-12 achieves obviously better error rate (4.43% in CIFAR-10 and 19.94% in CIFAR-100) with only 8M parameters compared with WRN-40-4 with 8.7M parameters or ResNet$^+$-1001 with 10.2M parameters. The results demonstrate that our *ensemble-connection*, which enables the classification module to assign independent weights to network ensemble outputs, substantially improves network ensembling effectiveness and, consequently, leads to higher efficiency of feature and parameter utilization. As mentioned in Section 1, these properties are favorable to medical images.

5.2. Image-language evaluation on BCIDR

We evaluate our MDNet for two tasks: report generation and symptom based image retrieval. We follow common evaluation methods [22] but also suggest complementary evaluation metrics specially designed for medical images. To validate our method, we use 5-fold cross validation. Ap-
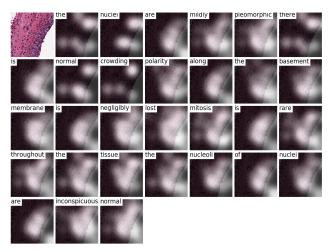
Figure 5: The image model predicts diagnostic reports (left-up corner) associated with sentence-guided attention maps. The language model attends to specific regions per predicted word. The attention is most sharp on urothelial neoplasms, which are used to diagnose the type of carcinoma.
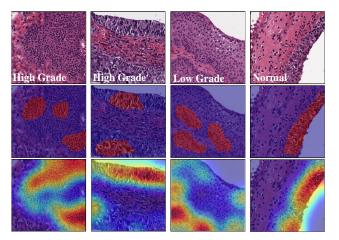

Figure 6: The illustration of class-specific attention. From top to bottom, test images, pathologist annotations, and class attention maps. Like the pathologist annotations, the attention maps are most activated in urothelial regions, largely ignoring stromal or background regions. Best viewed in color.

pendix B discusses training details.

Dataset The bladder cancer image and diagnostic report (BCIDR) dataset was collected in collaboration with a pathologist. Whole-slide images were taken using a 20X objective from hematoxylin and eosin (H&E) stained sections of bladder tissue extracted from a cohort of 32 patients at risk of a papillary urothelial neoplasm. From these slides, 1000 500x500 RGB images were randomly extracted close to urothelial neoplasms (each slide yields a slightly different number of images). We used a web interface to show each image (without diagnostic information of patient slides) and the pathologist then provided a paragraph describing observations to address five types of cell appearance features (Figure 2 shows an example), namely the state of nuclear pleomorphism, cell crowding, cell polarity, mitosis, and prominence of nucleoli followed by a diagnostic conclusion. The conclusion is comprised of four classes, i.e., normal, papillary urothelial neoplasm of low malignant potential (PUNLMP)/low-grade carcinoma, high-grade carcinoma, and insufficient information. Following this procedure, four doctors (non-experts in bladder cancer) wrote an additional four descriptions in their own free words but referring to the pathologist's description to guarantee accuracy. Thus there are five ground-truth reports per image in total. Each report varies in length between 30 and 59 words.

We randomly select 20% (6/32) of patients including 200 images as testing data and the remaining 80% of patients including 800 images for training and cross-validation. For data processing, the input image is resized to 224×224. We subtract the RGB mean from each image and augment the training data through clip, mirror and rotation operations. According to this dataset, the five descriptions and one conclusion are treated as $K{=}6$ separate tasks (defined in Sec-

tion 4.2) for LSTM training to support complete report generation. The conclusion is used as (4-way) labels for CNN training in all comparison experiments.

Baseline We choose the well-known image captioning scheme [14, 33] (the source code of [14]) as the baseline, which is to first train a CNN to represent images, followed by training an LSTM to generate descriptions. We use GoogLeNet instead of its originally used VGG [28], since the former performs better on BCIDR. We also train a small version of our EcNet, which has depth 38 and width 8, including 2.3M parameters (our purpose here is not to compare EcNet and GoogLeNet). Pre-trained GoogleNet and EcNet per validation fold are shared by all comparative models. When training LSTM, we test the cases with and without fine-tuning CNNs.

Ablation study MDNet is jointly trained which needs no pre-training or fine-tunning. For detailed comparison with the baseline, we also test two cases which training MDNet using the baseline strategies. In these cases, our optimization is not applied, so the differences from the baseline are task-separated LSTM and the integrated attention model.

5.2.1 Interpret model prediction

We start by qualitatively demonstrating the diagnosis process of MDNet: generating reports and showing image attention to interpret how the network uses visual information to support its diagnostic prediction. Two kinds of attention maps are demonstrated.

Sentence-guided attention is computed by our attention model, where each attention map corresponds to a predicted word to show the relevant part of image that the network attend. According to pathologists' observations, our computed attention maps are fairly encouraging, which intend to attend on informative regions and avoid less useful re-

Model	CNN	P?	F?	J?	B1	B2	B3	B4	M	R	C	DCA(%)±std
Baseline	GN	✓			90.6	81.8	73.9	66.6	39.3	69.5	2.05	72.6±1.8
	GN	✓	✓		90.7	82.0	74.3	66.9	39.5	69.9	**2.09**	74.2±3.8
	EN	✓			90.1	81.1	73.2	65.8	39.3	69.7	2.01	73.7±2.4
	EN	✓	✓		90.3	81.9	74.1	66.8	39.6	69.8	2.02	74.4±4.8
Ours	EN	✓			90.4	81.9	74.1	66.6	39.3	69.8	1.95	72.7±4.2
	EN	✓	✓		90.4	81.5	73.4	65.9	39.0	69.5	1.92	71.6±4.2
	EN			✓	**91.2**	**82.9**	**75.0**	**67.7**	**39.6**	**70.1**	2.04	**78.4**±1.5

Table 2: Quantitative evaluation of generated description quality and the DCA score. See text for metric notations. P, F, and J denote whether a pre-trained CNN is used, whether fine-tuning pre-trained CNNs when training LSTM, and whether using our proposed joint training approach (i.e. our proposed MDNet), respectively. The 5th and 6th rows are for the ablation study. GN and EN denote GoogleNet and EcNet.

	CNN	P?	F?	J?	Cr@1	Cr@5	Cr@10
Baseline	GN	✓			71.7±2.5	71.9±5.2	72.9±4.1
	GN	✓	✓		70.1±8.3	72.5±5.9	72.8±5.3
	EN	✓			64.4±2.4	70.8±0.9	72.5±1.6
	EN	✓	✓		68.3±2.0	71.8±1.5	73.4±1.9
Ours	EN	✓			68.7±5.5	73.1±2.8	74.3±1.7
	EN	✓	✓		71.6±5.5	75.7±3.9	75.8±2.7
	EN			✓	**78.6**±4.0	**79.5**±3.6	**79.4**±3.1

Table 3: Quantitative evaluation (mean±std) of report to image retrieval. See text for explanation of the metric Cr@k. The last row is our proposed MDNet.

gions. Figure 5 shows sample results. Please see the supplementary material for more results.

The conclusion-specific attention map is computed by AAS (i.e. the 14×14 Conv feature embedding). Recall that it has the implicit localization ability on image parts relate to the predicted label. To evaluate this attention qualitatively, we ask the pathologist to draw regions of interest of some test images that is necessary to infer conclusion based on his experience. Figure 6 shows the results. There is fairly strong correspondence between the pathologist annotations and regions with the sharpest attention. Recall that the training stage does not have region level annotations. These results demonstrate that MDNet has learned to discover useful information to support its prediction.

5.2.2 Diagnostic report generation

Evaluation metrics We report commonly used image captioning evaluation metric scores [31], including BLEU(B), METEOR(M), Rouge-L(R), and CIDEr(C). The diagnostic reports have more regular linguistic structure than natural image captions. Our experiments show that standard LSTM can capture the general structure, resulting in similar metric scores. Nevertheless, we care more about whether the trained models accurately express pathologically meaningful keywords. To make more definitive evaluation, we report the predicted diagnostic conclusion accuracy (DCA) extracted from generated report sentences.

The results are shown in Table 2. Our proposed MD-

Net (last row) outperforms all comparative baseline models by demonstrating significantly improved DCA (also smaller std) and most of other metrics. For the baseline methods in the first block of the table, the models using EcNet (3rd and 4th rows) achieve slightly better results than the models using GoogLeNet. We also observe that fine-tuning the pre-trained CNNs (either EcNet and GoogleNet) is generally beneficial but more unstable (i.e. higher std). The following image retrieval experiments provide more quantitative evaluation of the sentence-image mapping quality.
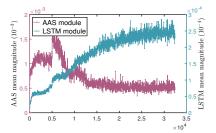
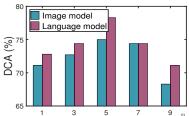### 5.2.3 Symptom description based image retrieval

We evaluate all trained models in Table 2 for the symptom description based image retrieval task shown in Table 3.

Evaluation metric Natural image captioning methods evaluate the groundtruth image recall at top k positions based on the ranking of images given an query sentence [9, 22]. However, in the medical image domain, this metric is not necessarily valid because images with close symptoms could share similar descriptions. Thus, low recall does not exactly indicate poor models. Instead, we evaluate the ability of the model to retrieve images with correct diagnostic conclusion given a query report. But for all query reports, we remove the words related to conclusion and only keep image feature descriptions. The intuition behind this metric is that doctors have clinical needs to query images with specified symptoms. Given some diseased image descriptions, it should be a failure if the model retrieves a healthy image. This metric is an exact measurement of sentence-image mapping quality because a mistake in a single symptom description could result in retrieval errors. We report the correct conclusion recall rate, denoted as Cr@k, $k = \{1, 5, 10\}$, of top k retrieved images corresponding to the query report.

Table 3 shows the mean (std) scores over 5 folds. As can be observed, fine-tuning EcNet results in noticeable improvement generally, especially for the two experimental cases on our network (5th and 6th rows), though they do not reach the results of our proposed MDNet (last row). Based on present results, we observe:

1. In general, fine-tuning pre-trained EcNet gives rise to larger improvement than fine-tuning GoogLeNet.
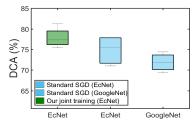
Figure 7: **Left**: The mean gradient magnitude. **Middle**: The DCA scores of the image model and language model in a MDNet respect to different η in x-axis. **Right**: The DCA (over 5 folds) scores of EcNet (stands for the image model of MDNet) and pre-trained EcNet and GoogLeNet.

2. MDNet that separates the modeling of overall reports as functions of independent image descriptions is more accurate to capture fine discrimination in descriptions, while fine-tuning (6th row against 5th row) further improves the mapping quality thanks to the design in Section 4.2.

3. Our proposed MDNet significantly outperforms baseline models, which indicates much better sentence-image mapping quality. One reason is because our joint training method prevents overfitting effectively.

6. Discussion

Optimization The weight of composite gradients are shifting during training. The basic rule is to assign large weight to $\frac{\partial \mathcal{L}_M}{\partial \theta_D}$ to allow AAS to dominate the image model training for a while, and gradually increase the scale of $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$ to introduce semantic knowledge and facilitate two models co-adapt mutually. We use a sigmoid-like function to change β from 0 to 1 gradually during the entire training process.

Balancing the scale of the two gradients is critical. We observe that simply scaling up $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$ without scaling down $\frac{\partial \mathcal{L}_M}{\partial \theta_D}$ (i.e. remove $1-\beta$) has negative effects in our practice, probably because the totally summed gradient w.r.t θ_D will grow larger and increase instability in model training. We observe $\sim 4\%$ DCA score decrease of the language model without averaging. Thus, we argue that using weighted averaging is necessary. However, simply averaging two gradients (using β) will make $\frac{\partial \mathcal{L}_M}{\partial \theta_D}$ overwhelm $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$ since they have different magnitudes. A heuristic way to observe this fact is to visualize their mean gradient magnitudes. As can be observed in Figure 7(left), the gradient magnitude of $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$ is much smaller than that of $\frac{\partial \mathcal{L}_M}{\partial \theta_D}$. We cross-validated η (see Figure 7(middle)) and set $\eta = 5$ throughout.

Small dataset and regularization The size of BCIDR is much smaller than common natural image datasets. This situation yields higher possibilities to end up with overfitted models, though we use regularization techniques and cross-validation. However, small dataset size is a common issue in the medical image domain; these large networks are still widely used [25, 26]. Figuring out effective regularization is extremely necessary. Both pre-trained CNNs and the image model of MDNet (i.e. AAS outputs) predict diagnostic conclusion labels. We can utilize this definite DCA score for more detailed analysis and comparison.

For all trained models, we observe the DCA of the language model strongly relies on that of corresponding image model (see Figure 7(middle)), which motivates us to analyze more about CNN training itself. According to Eq. (12), module M provides a standard CNN loss. If we interpret $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$ from module L as "noise" added onto the gradient $\frac{\partial \mathcal{L}_M}{\partial \theta_D}$, this "noise" disturbs the loss of module M and overall CNN training. In fact, moderate disturbance on the loss layer has regularization effects [34]. Therefore, our optimization behaves particular regularization on CNN to overcome overfitting. As compared in Figure 7(right), the image model of MDNet trained using our optimization approach outperforms pre-trained CNN models using stochastic gradient descent (SGD).

Multimodal mapping for knowledge fusion Image feature descriptions in diagnostic reports contain strong underlying supports for diagnostic conclusion inference. According to our results, our proposed MDNet for multimodal mapping learning effectively utilizes these semantic information to encourage sufficient multimodal knowledge sharing between image and language models, resulting in better mapping quality and more accurate prediction.

7. Conclusion and Future Work

This paper presents a novel unified network, namely MDNet, to establish the direct multimodal mapping from medical images and diagnostic reports. Our method provides a novel perspective to perform medical image diagnosis: generating diagnostic reports and corresponding network attention, making the network diagnosis and decision-making process semantically and visually interpretable. Sufficient experiments validate our proposed method.

Based on this work, limitations and open questions are drawn: building and testing large-scale pathology image-report datasets; generating finer [27] attention for small biomarker localization; applying to whole slide diagnosis. We expect to address them in the future work.

References

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016. 2, 3

[2] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *MICCAI*, 2013. 1

[3] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011. 5

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1, 2

[5] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 5

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 5

[7] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 2, 3, 5

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[9] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 7

[10] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CVPR*, 2017. 2, 3

[11] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 2

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 2

[13] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 2, 4

[14] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2, 4, 6

[15] A. Karpathy, A. Joulin, and F.-F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 2

[16] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Multimodal neural language models. In *ICML*, 2014. 2

[17] P. Kisilev, E. Walach, S. Hashoul, E. Barkan, B. Ophir, and S. Alpert. Semantic description of medical image findings: Structured learning approach. In *BMVC*. 2

[18] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. 2, 5

[19] M. Lin, Q. Chen, and S. Yan. Network in network. In *ICLR*, 2014. 2, 5

[20] C. Liu, J. Mao, F. Sha, and A. Yuille. Attention correctness in neural image captioning. *AAAI*, 2017. 4

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[22] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015. 1, 2, 5, 7

[23] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *ICML*, 2013. 2

[24] S. Reed, Z. Akata, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 2

[25] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *CVPR*, 2016. 2, 4, 8

[26] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016. 8

[27] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014. 2, 8

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 6

[29] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015. 2, 5

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[31] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 7

[32] A. Veit, M. Wilber, and S. Belongie. Residual networks are exponential ensembles of relatively shallow networks. *arXiv preprint arXiv:1605.06431*, 2016. 2

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2, 3, 6

[34] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian. Disturblabel: Regularizing cnn on the loss layer. In *CVPR*, 2016. 8

[35] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015. 2, 3

[36] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2, 3, 4

[37] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2

[38] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2

[39] S. Zagoruyko and N. Komodakis. Wide residual networks. In *BMVC*, 2016. 2, 5

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 4

[41] L. M. Zintgraf, T. S. Cohen, and M. Welling. A new method to visualize deep neural networks. *arXiv preprint arXiv:1603.02518*, 2016. 2