A Diagnostic Report Generator from CT Volumes on Liver Tumor with Semi-supervised Attention Mechanism

Jiang Tian, Cong Li, Zhongchao Shi, and Feiyu Xu(✉)

AI Lab, Lenovo Research, Beijing, China
{tianjiang1,licong17,shizc2,fxu}@lenovo.com

Abstract. Automatically generating diagnostic reports with interpretability for computed tomography (CT) volumes is a new challenge for the computer-aided diagnosis (CAD). In this paper, we propose a novel multimodal data and knowledge linking framework between CT volumes and textual reports with a semi-supervised attention mechanism. This multimodal framework includes a CT slices segmentation model and a language model. Semi-supervised attention mechanism paves the way for visually interpreting the underlying reasons that support the diagnosis results. This multi-task deep neural network is trained end-to-end. We not only quantitatively evaluate our system performance (76.6% in terms of BLEU@4), but also qualitatively visualize the attention heat map for this framework on a liver tumor dataset.

1 Introduction

Every working day, a Chinese radiologist in Beijing in a regular hospital has to review more than 100 CT volumes[1] and to write many corresponding diagnostic reports, which is time-consuming and prone to inter- and intra-rater variations. There have been continuous efforts and progresses in the automatic recognition and localization of specific diseases and organs, mostly on radiology images. Nonetheless, generating a description about the content of a medical image automatically like a report written by a human radiologist might have a big impact for countries like China where doctors have a very big work load. The full power of this field has vast potentials to renovate medical CAD, but very little related work has been done [1–3]. Recent joint visual content and language modeling by deep learning enables the generation of semantic descriptions, which provide more intelligent predictions [4,11].

The principles of deep neural networks during training and testing are difficult to interpret and justify. In some scenarios, predictions and metrics calculated on these predictions do not suffice to characterize the model. On the other hand, interpretability of a CAD system is highly needed, due to medical diagnosis mission critical application nature. In comparison, human decision-makers are

[1] Typically tens to hundreds of slices per volume.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 702–710, 2018.
https://doi.org/10.1007/978-3-030-00934-2_78

themselves interpretable because they can explain their actions. Consequently, doctors must feel confident in the reasoning behind the program, and it is difficult to trust systems that do not explain or justify the conclusions.

Inspired by this fact, in this paper, we propose a framework which can train a fully convolutional neural network (FCN) to segment CT slices, and a separate Long-Short Term Memory (LSTM) language model, to generate captions, which might be regarded as interpretations that accompany segmentation. Meanwhile, ability to visualize what the model "sees" may determine qualitatively what a model has learned. Attention is such a main component supporting the visual interpretability of deep neural networks. We integrate two attention mechanisms into this framework. Segmentation may be regarded as a supervised approach to let the network capture visual information on "targeted" regions of interest. Another attention mechanism dynamically computes a weight vector along the axial direction to extract partial visual features supporting word prediction. While these interpretations may not elucidate precisely how a model works, they may nonetheless confer useful information for doctors.

More specifically, we apply this interpretable diagnostic report generator to CT volumes of liver. The liver is a common site of primary or secondary tumor development [12]. Until now, only interactive methods achieved acceptable results on segmenting liver tumors. In comparison, in this work, our framework provides explanations and justifications for its diagnosis prediction and makes the process interpretable to radiologists.

2 Model

In this section, as shown in Fig. 1(c), we describe a general approach, based on the encoder-decoder framework [7], to generate diagnostic report. Throughout the paper, we will denote by $LSTM(\mathbf{h}_{t-1}, \mathbf{y}_t)$ the LSTM operation on vectors $\mathbf{h}_{t-1}$ and $\mathbf{y}_t$ to achieve $\mathbf{h}_t$, wherein, $\mathbf{h}_t$ denotes hidden states from LSTM, $\mathbf{y}_t$ denotes word embedding at time t.

Encoder-Decoder Framework. The encoder network encodes the CT slices into a set of vectors as $a = \{\mathbf{a}_0, \cdots, \mathbf{a}_{\Pi-1}\}$, where Π is the total number of slices in a CT volume. Automatic liver and tumor segmentation from the contrast-enhanced CT volumes is a very challenging task due to the low intensity contrast between the liver and other neighboring organs. To tackle these difficulties, strong ability for extracting visual features is required. Recently, FCNs have achieved remarkable success on segmentation tasks [9]. We build the encoder upon this elegant architecture. The architecture of FCN is illustrated in Fig. 1(a). The final layer consists of two branches for predicting segmentation masks for liver and tumor separately. A 1×1 convolution is used to map each 128-component feature vector to the desired number of classes in each branch.

As shown in Fig. 1(b), the feature maps corresponding to liver and tumor, which represent categorical probability distributions, will be forwarded to a convolutional neural network (CNN) to embed the segmentation results. With prior knowledge that a liver tumor is inside the liver, an element-wise multiplication

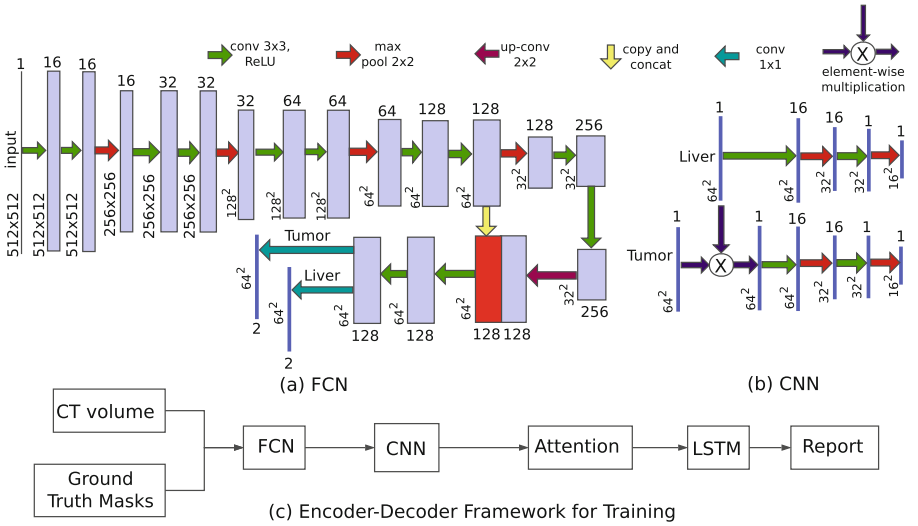

Fig. 1. FCN (a), CNN (b), and overall illustration (c) of the proposed encoder-decoder framework for training (best viewed in color). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted either on top of or below the box. The x-y size is provided at the lower left edge of the box. Red box represents copied feature maps. The arrows denote different operations. Based on input CT slices and resized ground truth masks, the FCN model generates a segmentation mask to pass to CNN to embed the visual features, attention mechanism combines these features with LSTM's hidden state, the resulting vectors are then fed into LSTM in the form of task tuple. LSTM executes prediction tasks according to a specified start token.

of liver mask and tumor one is conducted to eliminate the tumor pixels outside the liver. The resulting two 16×16 feature maps will be the visual features for each input slice. Note that $\mathbf{a}_i$ is defined as reshaping only one 16×16 feature map, either for liver or for tumor depending on a specific task, into a vector.

Visual features are used as the input into *LSTM* through attention mechanism for subsequently predicting words sequentially within a sentence. The prediction stops when *#end*, which designates the end of a sentence, is emitted. The decoding process is summarized as follows.

$$\mathbf{h}_t = LSTM(\mathbf{h}_{t-1}, \mathbf{y}_t, \mathbf{z}_t), \qquad p(word|\mathbf{h}_t) \propto exp(\mathbf{h}_t, \mathbf{y}_t, \mathbf{z}_t), \tag{1}$$

where $\mathbf{z}_t = \sum_{i=0}^{\Pi-1} \alpha_{ti}\mathbf{a}_i$, it is a dynamic vector that represents the relevant part of image feature at time step t, α_{ti} is a scalar weighting of visual vector $\mathbf{a}_i$ at time step t, defined as follows.

$$\alpha_{ti} = exp(e_{ti}) / \sum_{k=0}^{\Pi-1} exp(e_{tk}), \tag{2}$$

$$e_{ti} = f_{attention}(\mathbf{a}_i, \mathbf{h}_{t-1}). \tag{3}$$

$f_{attention}$ is a function that determines the amount of attention allocated to image feature $\mathbf{a}_i$, conditioned on the LSTM hidden state $\mathbf{h}_{t-1}$. This function is implemented as a multilayer perceptron as follows.

$$f_{attention} = \mathbf{w}^T tanh(U_a \mathbf{h}_{t-1} + W_a \mathbf{a}_i + \mathbf{b}_a). \tag{4}$$

Note that by construction $\sum_{i=0}^{I-1} \alpha_{ti} = 1$.

Rather than compress an entire CT volume into a static representation, attention allows for salient features to dynamically come to the forefront as needed. It can be viewed as a type of alignment between image space and language one. Improving such alignment can be achieved by adding supervision on attention maps [2,8]. In our approach, we propose to tackle this problem by utilizing segmentation networks to support image-language alignment. Specifically, we perform two FCN based segmentation tasks (one for liver, the other for tumor). The purpose of the segmentation task here is two-fold. First, it is a supervised attention mechanism. Second, it helps the gradients back-propagation during training.

Report Generation. A radiologist narrates findings in a report from observations of a liver CT imaging study, wherein, an observation is a distinctive area compared to background liver at imaging.[2] The task of report generation is traditionally solved for X-ray or pathology images by learning to map a set of diverse image measurements to a set of semantic descriptor values (or class labels). In comparison, our model is an end-to-end one which avoids managing a complex sequential pipeline.

In this work, a report is structured as *shape*, *contour*, and *intensity* sections to communicate information (e.g., volumetric, morphometric) with regard to region of interests. Take the descriptions of tumor symptoms on the right lobe of liver as an example, we have English translations as follows. *There are multiple low density regions with large area on the right lobe of liver. There are multiple low density regions on the right lobe of liver. There is one low density region on the right lobe of liver. There is one low density region with large area on the right lobe of liver. There is no abnormal region on the right lobe of liver.* Note that our method localizes the right lobe of liver and produces free-text descriptions.

A report describes multiple symptoms of observing CT volumes. It has more regular linguistic structure than natural image captions [2]. We let one LSTM focus on mining discriminative information for a specific description. All description modelings share LSTM weights. In this way, complete report generation becomes a function of modeling of each specific feature description.

In the training stage, given a mini-batch of CT slices and report, after forwarding the CT slices to the image model, we duplicate each sample inside, resulting in a size of five mini-batch as the input of LSTM. Each duplication takes one of the image features either for liver or for tumor, and one particular symptom description extracted from the report. Note that we denote by *#start* a special start word and *#end* a special stop one which designate the start and

[2] https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS.

end of a sentence. Each particular feature description has unique start and end tokens. In this way, we use particular signal to inform LSTM the start and end of a targeting task.

In testing, the first element of every input sequence is always a special start token $\#start$, the network therefore makes a prediction corresponding to a special aspect of the visual features based on the LSTM state and the previous word. If $\#end$ token is emitted, the prediction will stop.

3 Experiments

To the best of our knowledge, there is no benchmark for generating diagnostic report from CT volumes on liver tumor. In this section, we describe our experimental methodology, run a number of ablations to analyze the framework, and show the corresponding results.

Dataset. We use a publicly available dataset of MICCAI 2017 LiTS Challenge [12], which contains 131 and 70 contrast-enhanced abdominal CT scans for training and testing, respectively. The diagnostic reports (with segmentation masks as reference) for the training dataset were collected in collaboration between two experts on medical imaging and a doctor focused on abdominal surgery in a top hospital in Beijing. A paragraph in Chinese[3] is provided to describe observations to address three types of appearance features, namely the shape, contour, and intensity. We clip all CT scans with a window $[-75, 175]$ HU (Hounsfield scale) according to radiologist's advice to ignore irrelevant abdominal organs for both training and testing. We perform a 3-fold cross validation on the training dataset.

Implementation. All models are trained and tested with Tensorflow on NVIDIA Tesla P100 (with $16276M$ memory) GPUs. Due to the GPU memory constraint, each CT volume has been pooled along the axial direction into 96 slices, if the number of its slices is smaller than 96, they will be zero padded. We implement this average pooling in a sliding window manner as $win = \lceil \Pi/96 \rceil$ and $str = \lfloor \Pi/96 \rfloor$, where win is the window size, str is the stride. $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote ceiling and floor operations, respectively. In comparison, we take the union rather than average of the ground truth segmentation masks in the win scope for not missing information. We train the full model end-to-end by stochastic gradient descent (SGD) with momentum. We use a mini-batch size of 1 CT volume and 5 sentences, and a fixed learning rate of 10^{-2}. We use a momentum of 0.9. The model is trained with random initialized weights.

Ablation Study. We are more interested in whether supervised attention model has better captioning performance. In this section we give experimental support. Figure 2 denotes a baseline structure incorporating neither liver nor tumor segmentation structures for visual embedding. In order to make fair comparison, we only make minor modifications of the visual embedding structure shown in

[3] It has the same processing pipeline as English report once embedded.

Fig. 1(a). The implicit attention refers to the alignment between the visual feature space and the textual one, it is still in effect for the baseline architecture shown in Fig. 2. Semi-supervised attention refers to segmentation supervision on the visual feature for the aforementioned implicit attention.

Table 1. BLEU@n and ROUGE(R) scores of our methods on liver tumor dataset. All values are reported as percentage (%).

Method	B1	B2	B3	B4	R
S1024	88.3	83.7	79.8	75.6	89.2
I1024	88.1	83.0	78.7	74.0	88.6
S512	88.9	84.5	80.7	76.6	89.8
I512	88.8	83.9	79.7	75.2	89.4
S256	88.8	84.1	80.1	75.9	89.2
S128	86.9	81.5	76.9	71.0	87.9

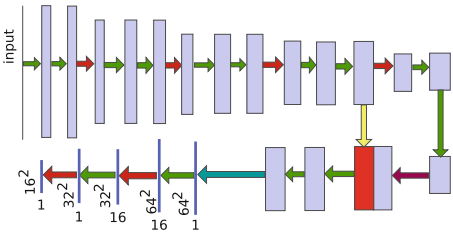

Fig. 2. Baseline architecture for visual embedding. The arrows denote same operations as shown in Fig. 1. The number of channels and x-y size are omitted if they are the same with Fig. 1(a).

We quantitatively evaluate the report generation performance on BLEU(B) [5] and ROUGE(R) [6] scores. The results are given in Table 1. The first column in the table lists the LSTM hidden units number, wherein, S stands for semi-supervised attention, and I depicts implicit attention, the number after S or I lists the corresponding LSTM hidden state size.

The unigram BLEU scores are to account for how much information is retained. The longer n-gram BLEU scores account for to what extent it reads like "diagnostic report". The comparison between S# and I# indicates not only that model using semi-supervised attention improves over the baseline model with only implicit attention mechanism, but also that the longer the n-gram, the bigger the performance difference between S# and I#. It suggests that our framework provides better alignment from natural language words to provided visual features, and semi-supervised mechanism becomes particular necessary.

At the same time, increasing the hidden layer size can generally lead to performance improvements except hidden state size of 1024, which means overfitting occurs here. Meanwhile, the number of parameters increase exponentially. Therefore, in our experiments, the hidden layer size is empirically set to 512 as it has a better tradeoff between performance and model complexity.

Generating Visual Interpretations. One common approach to generate interpretations is to render visualizations in the hope of determining qualitatively what a model has learned. As shown in Fig. 3, by visualizing the attention component learned by the model, we are able to add an extra layer of interpretability to the output of the model. In the computer vision community, these approaches have been explored to investigate what information is retained at various layers of a neural network [10].

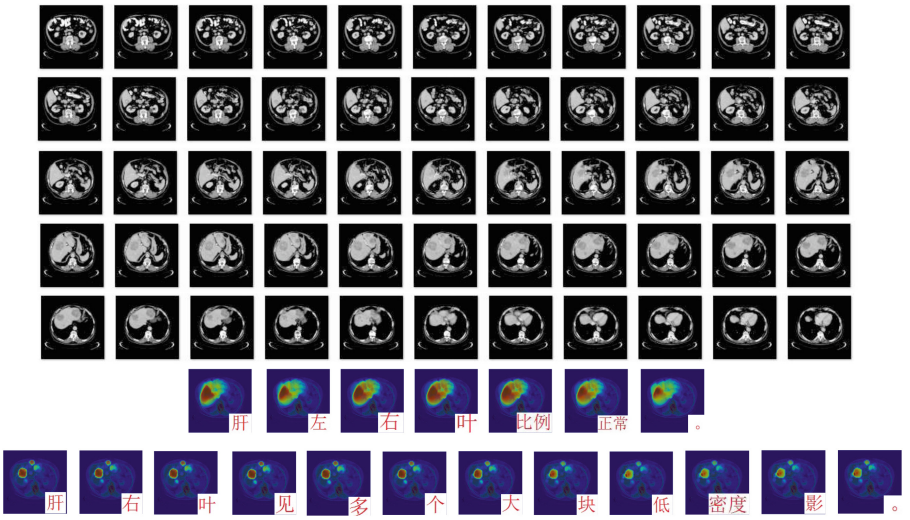

Fig. 3. An example of generating interpretations. The first five rows list a sequence of CT slices, the patient has symptoms of liver cancer as indicated by low density regions on a series of slices. The sixth row shows the attention heat maps for liver superposed onto the averaged CT image. English translation for the description is "The ratio between left lobe of liver and right lobe of liver is normal". The seventh row shows the attention heat maps for tumor superposed onto the averaged CT image. English translation is "There are multiple low density regions with large area on the right lobe of liver", whereas that for visual embedding defined in Fig. 2 is "There are multiple low density regions on the right lobe of liver". The semi-supervised attention captures information about the size (*large area*) compared with the implicit one.

In order to visualize the attention weights for the model, we record down each α_{ti}, a scalar weighting of visual vector $\mathbf{a}_i$ at time step t, in Eq. (2). Heat map for attention at 64×64 scale is calculated as $hm = \sum_{i=0}^{H-1} \alpha_{ti} S_c(i)$, where $S_c(i)$ corresponds to the c-th class (either liver or tumor) of the segmentation output for the i-th input. We up-sample the hm by a factor of 8, which is then superposed onto the averaged image through all CT slices in a volume. As shown in Fig. 3, the model learns alignments that correspond very strongly to the "targeted" regions. We see that it is possible to exploit such visualizations to get an intuition as to why the text explanations were made.

Evaluation on Test Set. Based on BLEU@4, we select the best model from cross validation for the test set. We let the state-of-the-art results from the challenge[4] be ground truth for the segmentation task in our framework. We pre-process the ground truth masks using the aforementioned pooling along the axial direction, and resize them to the dimension of 64×64. Finally, for test set, the Dice per case (an average Dice per volume score) for liver is 0.942, and for tumor is 0.549.

[4] We are the MICCAI 2017 LiTS Challenge first place team.

In addition to the automated evaluation metrics, we present human evaluations on the test results. We randomly select 30 report predictions from the test set, and conduct a paid doctor evaluation. A rating scale (1: definite accept, 5: definite reject) is adopted here for each report. Average score is 2.33.

4 Conclusion

This paper investigates multimodal knowledge sharing between CT volumes and diagnostic reports with semi-supervised attention mechanisms. It focuses on introducing a system to assist doctors in medical imaging interpretation. Textual descriptions communicate facts in CT images to doctors. These descriptions are regarded as interpretations. The second objective is to visually interpret the underlying reasons that support the diagnosis results. Showing image attention to interpret how the network uses visual information will support its diagnostic prediction. Furthermore, through experiments, we show that using segmentation supervision on the visual feature for the implicit attention improves the captioning performance.

References

1. Xu, T., Zhang, H., Huang, X., Zhang, S., Metaxas, D.N.: Multimodal deep learning for cervical dysplasia diagnosis. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 115–123. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_14
2. Zhang, Z., Xie, Y., Xing, F., Mcgough, M., Yang, L.: Mdnet: a semantically and visually interpretable medical image diagnosis network. In: CVPR, pp. 6428–6436 (2017)
3. Zhang, Z., Chen, P., Sapkota, M., Yang, L.: TandemNet: distilling knowledge from medical images using diagnostic reports as optional semantic references. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 320–328. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_37
4. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
5. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL, pp. 311–318 (2002)
6. Lin, C.-Y.: Rouge: a package for automatic evaluation of summaries, pp. 74–81. In: ACL Workshop (2004)
7. Yao, L., et al.: Describing videos by exploiting temporal structure. In: ICCV, pp. 4507–4515 (2015)
8. Liu, C., Mao, J., Sha, F., Yuille, A.: Attention correctness in neural image captioning. In: AAAI, pp. 4176–4182 (2017)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

10. Zintgraf, L.M., Cohen, T.S., Welling, M.: A new method to visualize deep neural networks. arXiv preprint arXiv:1603.02518 (2016)
11. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML, pp. 2048–2057 (2015)
12. MICCAI 2017 LiTS Challenge. https://competitions.codalab.org/competitions/17094