Paper summaries

October 6, 2020

1 On the Automatic Generation of Medical Imaging Reports (Jing, Xie, and Xing, 2018)

1.1 Introduction

The reading and interpretation of medical images are usually conducted by specialized medical professionals. Report writing can be error-prone for inexperienced physicians, and time-consuming and tedious for experienced physicians. Several challenges need to be addressed:

1. A complete report consists of multiple heterogeneous sources of information
2. Localize image regions and attach the right description to them
3. Descriptions in reports are usually long, with multiple sentences

The proposed solutions are:

1. A **multi-task learning framework** for simultaneous prediction of tags and text generation
2. A **Co-attention mechanism:** simultaneous attention to images and predicted tags; explores synergistic effects of visual and semantic information
3. A **Hierarchical LSTM:** Leverages compositional nature of reports: first generates high-level topics, then fine-grained descriptions from each one

1.2 Methods and Architecture

An image is divided into regions, and a CNN encoder is used to learn visual features for these patches. These features are fed into a *multi-label classifier*, from which tags are predicted. These tags are transformed into *semantic feature vectors* by a custom embedding. Both visual and semantic features are fed into the co-attention module, which produces a combined *context vector,* which **simultaneously captures the visual and semantic information of this image.**

The decoding and caption generation process is performed by the hierarchical LSTM, which leverages the compositional structure of a medical report (each sentence focusing on one specific topic). The *sentence LSTM*, using the context vector, first generates a sequence of high-level topic vectors representing sentences. Each one is passed to the *word LSTM,* which then generates a sentence for each topic vector. The number of sentences or topic vectors to be generated is regulated by the *stop control*.

1.2.1 Tag prediction

This is treated as a multi-label classification task. Given an image I, visual features $\{\boldsymbol{v}_n\}_{n=1}^N \in \mathbb{R}^D$ are extracted from the CNN encoder, and fed to a *multi-label classification* (MLC) network, which then generates a probability distribution over the L tags

$$\boldsymbol{p}_{I,\text{pred}}\left(\boldsymbol{l}_i = 1 \mid \{\boldsymbol{v}_n\}_{n=1}^N\right) \propto \exp\left(\text{MLC}_i\left(\{\boldsymbol{v}_n\}_{n=1}^N\right)\right)$$

where $\boldsymbol{l} \in \mathbb{R}^L$ is a binary tag vector, each component representing the presence or absence of the corresponding tag. Finally, the embeddings of the M most likely tags $\{\boldsymbol{a}_m\}_{m=1}^M$ are used as **semantic features.**

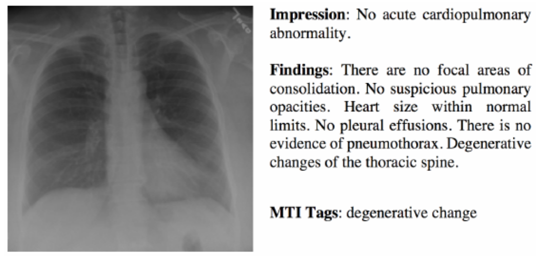

Figure 1: An exemplar chest x-ray report. In the *impression* section, the radiologist provides a diagnosis. The *findings* section lists the radiology observations regarding each area of the body examined in the imaging study. The *tags* section lists the keywords which represent the critical information in the findings. These keywords are identified using the Medical Text Indexer (MTI).

Figure 1: Sample report from IU X-ray

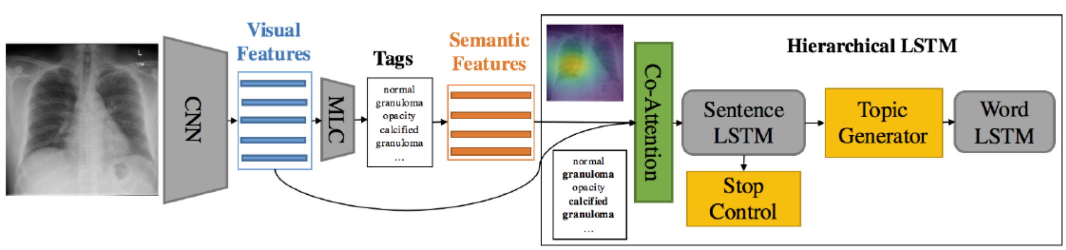

Figure 2: Illustration of the proposed model. MLC denotes a *multi-label classification* network. Semantic features are the word embeddings of the predicted tags. The boldfaced tags "calcified granuloma" and "granuloma" are attended by the co-attention network.

Figure 2: Architecture

1.2.2 Co-Attention

Visual attention does not provide sufficient high-level semantic information, which the tags can always provide. A co-attention mechanism can simultaneously attend to visual and semantic modalities.

In the sentence LSTM at time step s, the joint context vector $\mathbf{ctx}^{(s)} \in \mathbb{R}^C$ is generated by a co-attention network $f_{\text{co-att}}\left(\{\boldsymbol{v}_n\}_{n=1}^N, \{\boldsymbol{a}_m\}_{m=1}^M, \boldsymbol{h}_{\text{sent}}^{(s-1)}\right)$, with $\boldsymbol{h}_{\text{sent}}^{(s-1)} \in \mathbb{R}^H$ being the previous hidden state. The co-attention network $f_{\text{co-att}}$ uses a single feedforward layer to compute separate soft visual and semantic attentions

$$\alpha_{\mathbf{v},n} \propto \exp\left(\mathbf{W}_{\mathbf{v}_{\text{att}}}\mathbf{v}_n + \mathbf{W}_{\mathbf{v},\mathbf{h}}\mathbf{h}_{\text{sent}}^{(s-1)}\right)$$
$$\alpha_{\mathbf{a},m} \propto \exp\left(\mathbf{W}_{\mathbf{a}_{\text{att}}}\mathbf{a}_m + \mathbf{W}_{\mathbf{a},\mathbf{h}}\mathbf{h}_{\text{sent}}^{(s-1)}\right)$$

The visual and semantic context vectors

$$\mathbf{v}_{\text{att}}^{(s)} = \sum_{n=1}^{N} \alpha_{\mathbf{v},n}\mathbf{v}_n$$
$$\mathbf{a}_{\text{att}}^{(s)} = \sum_{m=1}^{M} \alpha_{\mathbf{a},m}\mathbf{a}_m$$

These context vector may be combined by concatenation followed by a fully connected layer:

$$\mathbf{ctx}^{(s)} = \mathbf{W}_{\text{fc}}\left[\mathbf{v}_{\text{att}}^{(s)}; \mathbf{a}_{\text{att}}^{(s)}\right]$$

1.2.3 Sentence LSTM

A single LSTM layer whose input is the joint context vector $\mathbf{ctx}^{(s)}$ and generates a topic vector $\mathbf{t} \in \mathbb{R}^K$ as long as the stop control allows it to.

Topic generator Deep output layer (LSTM + multi-layer feedforward)

$$\mathbf{t}^{(s)} = \tanh\left(\mathbf{W}_{\mathbf{t},\mathbf{h}}\mathbf{h}_{\text{sent}}^{(s)} + \mathbf{W}_{\mathbf{t},\mathbf{ctx}}\mathbf{ctx}^{(s)}\right)$$

Stop control Deep output layer for the continuation of the sentence LSTM. The layer takes the previous and current hidden states and produces a distribution over $\{\text{STOP} = 1, \text{CONTINUE} = 0\}$, $p_{\text{stop}}^{(s)}$

$$p\left(\text{STOP} \mid \mathbf{h}_{\text{sent}}^{(s-1)}, \mathbf{h}_{\text{sent}}^{(s)}\right) \propto \exp\left\{\mathbf{W}_{\text{stop}}\tanh\left(\mathbf{W}_{\text{stop},s-1}\mathbf{h}_{\text{sent}}^{(s-1)} + \mathbf{W}_{\text{stop},s-1}\mathbf{h}_{\text{sent}}^{(s)}\right)\right\}$$

This stopping probability is then compared with a predefined threshold.

1.2.4 Word LSTM

For each topic vector, the words in the sentence are generated by the *word LSTM*. Its first and second inputs are the topic vector $\mathbf{t}$ and a START token, then followed by the rest of the words (Krause et al., 2016). Each hidden state $\mathbf{h}_{\text{word}} \in \mathbb{R}^H$ is directly used to predict the distribution over words:

$$p(\text{word} \mid \mathbf{h}_{\text{word}}) \propto \exp(\mathbf{W}_{\text{out}}\mathbf{h}_{\text{word}})$$

1.2.5 Parameter learning

Each training example is a tuple $(I, \mathbf{l}, \mathbf{w})$, with $\mathbf{w}$ being the paragraph, with S sentences, each with T_S words (ground truth).

1. The MLC predicts the tag distribution $\mathbf{p}_{I,\text{pred}}$. Its ground truth may be computed by $\mathbf{p}_I = \mathbf{l}/\left\|\mathbf{l}\right\|_1$. Thus, the training loss for this task is the cross-entropy between both distributions ℓ_{tag}.

2. The sentence LSTM is unrolled for S steps, producing that number of topic vectors $\mathbf{t}^{(s)}$ and stop distributions $p_{\text{stop}}^{(s)}$. For each sentence, this stop probability is compared with the indicator $I\{s = S\}$, which evaluates to 0 [CONTINUE], until $s = S$, when it evaluates to 1 [STOP] (that is, a kronecker delta δ_{sS}). The cross entropy between them is the loss ℓ_{sent}.

3. The S topic vectors are fed to the word LSTM to generate $\mathbf{w}_{s,t}$ words. The training loss for each word is the cross entropy ℓ_{word} between the ground truth word $w_{s,t}$ and the predicted word distribution $p_{s,t}$.

Thus, the overall training loss is

$$\begin{aligned}\ell\left(I, \mathbf{l}, \mathbf{w}\right) = \lambda_{\text{tag}} \ell_{\text{tag}} \\ + \lambda_{\text{sent}} \sum_{s=1}^{S} \ell_{\text{sent}} \left(p_{\text{stop}}^{(s)}, I\{s = S\}\right) \\ + \lambda_{\text{word}} \sum_{s=1}^{S} \sum_{t=1}^{T_S} \ell_{\text{word}} \left(p_{s,t}, w_{s,t}\right)\end{aligned}$$

Furthermore, and attention regularization loss (Xu et al., 2015) for both visual $\boldsymbol{\alpha} \in \mathbb{R}^{N \times S}$ and semantic $\boldsymbol{\beta} \in \mathbb{R}^{M \times S}$ attention coefficients. This regularization encourages the model to pay equal attention over different image regions and tags.

$$\ell_{\text{reg}} = \lambda_{\text{reg}} \left[\sum_{n=1}^{N} \left(1 - \sum_{s=1}^{S} \alpha_{n,s} \right)^2 + \sum_{m=1}^{M} \left(1 - \sum_{s=1}^{S} \beta_{m,s} \right)^2 \right]$$

References

Jing, Baoyu, Pengtao Xie, and Eric Xing (July 2018). "On the Automatic Generation of Medical Imaging Reports". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2577–2586. DOI: 10.18653/v1/P18-1240. URL: https://www.aclweb.org/anthology/P18-1240.

Krause, Jonathan, Justin Johnson, Ranjay Krishna, and Li Fei-Fei (2016). *A Hierarchical Approach for Generating Descriptive Image Paragraphs*. arXiv: 1611.06607 [cs.CV].

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio (2015). *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. arXiv: 1502.03044 [cs.LG].