Zhang, Yazhou; Rong, Lu; Li, Xiang; Tiwari, Prayag; Zheng, Qian; Liang, Hui

MedSeq2Seq: A Medical Knowledge Enriched Sequence to Sequence Learning Model for COVID-19 Diagnosis

Published in:
Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021

DOI:
10.1109/BIBM52615.2021.9669854

Published: 01/01/2021

Please cite the original version:
MedSeq2Seq: A Medical Knowledge Enriched Sequence to Sequence Learning Model for COVID-19 Diagnosis

1st Yazhou Zhang
Software Engineering College
Zhengzhou University of Light Industry
Zhengzhou, China
yzzhang@zzuli.edu.cn

3rd Xiang Li
Shandong Computer Science Center (National Supercomputing Center in Jinan)
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, China
xiangli@sdas.org

5th Qian Zheng
Software Engineering College
Zhengzhou University of Light Industry
Zhengzhou, China
zhengqian@zzuli.edu.cn

2nd Lu Rong†
Software Engineering College
Zhengzhou University of Light Industry
Zhengzhou, China
lurong2013@outlook.com

4th Prayag Tiwari
Department of Computer Science
Aalto University
Helsinki, Finland
prayag.tiwari@aalto.fi

6th Hui Liang†
Software Engineering College
Zhengzhou University of Light Industry
Zhengzhou, China
hliang@zzuli.edu.cn

Abstract—The COVID-19 pandemic has had a severe impact on humans’ lives and healthcare systems worldwide. How to early, fastly and accurately diagnose infected patients via multi-modal learning is now a research focus. The central challenges in this task mainly lie on multi-modal data representation and multi-modal feature fusion. To solve such challenges, we propose a medical knowledge enriched multi-modal sequence to sequence learning model, termed MedSeq2Seq. The key components include two attention mechanisms, viz., intra-modal (I_a) and inter-modal (I_e) attentions, and a medical knowledge augmentation mechanism. The former two mechanisms are to learn multi-modal refined representation, while the latter aims to incorporate external medical knowledge into the proposed model. The experimental results show the effectiveness of the proposed MedSeq2Seq framework over state-of-the-art baselines with a significant improvement of 1%-2%.

Index Terms—COVID-19 diagnose, coronavirus epidemic, Seq2Seq learning, attention mechanism, deep learning

I. Introduction

Novel coronavirus (COVID-19) has swept across the globe and had a severe impact on humans’ lives and healthcare systems worldwide. It is considered as one of the greatest medical challenges of the 21st century [1]. In view of its quick transmissibility and strong virulence, sensitive and accessible diagnosis approaches are of great significance. However, the main clinical treatment, e.g., reverse transcription polymerase chain reaction (RT-PCR), has been suffering from weak timeliness and expensive cost. Hence, diagnosing COVID-19 with the help of artificial intelligence (AI) approaches is a research focus in AI and Medical communities [2]–[4].

Artificial intelligence and deep learning has recently achieved noticeable success in a wide range of classification tasks, e.g., emotion recognition [5], [6], sentiment classification [7], [8], image classification [9], [10], etc. The reason is that the feasibility of big data and the strong learning ability of deep neural networks make them obtain dazzling performance. So, the use of deep learning and artificial intelligence in medical analysis and COVID-19 detection is a natural way to improve its performance. Further, the lack of medical resources also leads to the development of AI based COVID-19 diagnosing.

In this paper, we extend this problem to multi-modal scenario since rich records of chest CT and textual symptoms have been produced. Such multi-modal records could provide vivid and accurate descriptions of COVID-19 symptoms, leading to a fact that multi-modal COVID-19 diagnosis has gained popularity [11]. The central challenges in this task mainly lie on multi-modal data representation and multi-modal feature fusion [12]. Most existing approaches have paid full attention to the former, while how to capture the correlations across multi-modal features remains an open problem to be solved. Researchers fail to reach an agreement on the optimal manner of merging multi-modal information, due to the heterogeneities and augmented diversities across modalities.
To generically solve such challenges, we propose a medical knowledge enriched multi-modal sequence to sequence learning model, termed MedSeq2Seq. The key components include two attention mechanisms, viz. intra-modal (I_t) and inter-model (I_e) attentions, and a medical knowledge augmentation layer. We place I_t and I_e attentions in the encoder for learning multi-modal refined representation. The knowledge augmentation layer is designed to incorporate external medical knowledge into multi-modal representation. The final multi-modal representation is fed into the long-short term (LSTM) decoder for COVID-19 diagnosis.

Extensive experiments are conducted the COVID-CT-MD dataset, in comparison with a wide range of strong baselines, including three strong deep learning approaches, i.e., hyper deep neural networks (HDNN), multi-modal deep convolutional neural network (DCNN), bidirectional gated recurrent unit (BiGRU), ResNet-34 and VGG-16. Our model obtains significant improvements in terms of F1 and accuracy scores. The results show the effectiveness of the proposed framework.

The major innovations of the work can be written as follows.
- We propose a novel multi-modal medical knowledge learning framework.
- We design intra-modal and inter-modal attentions to learn effective multi-modal fused representation.
- We verify the effectiveness of our model. Empirical experimental results show that our model outperforms strong baselines.

The rest of this paper is organized as follows. Section II outlines the related work. In Section III, we describe the proposed framework in detail. In Section IV, we report the empirical experiments and analyze the results. Section V concludes the paper and points out future research directions.

II. Related Work

Now, we introduce a brief review of the related work, i.e., machine and deep learning based COVID-19 detection approaches.

Machine learning based approach. Such methods mainly leverage many kinds of machine learning methods, such as naive bayes (NB), support vector machine(SVM), random forest (RF), and shallow-layer neural networks. They often involve training classifiers from labeled samples, which belongs to a supervised classification task. Hu et al. [13] noticed the importance of multi-modal information, and proposed a multi-modal SVM for early COVID-19 detection based on chest CT dataset. Zoabi et al. [14] designed eight binary features: sex, age, known contact with an infected individual, etc., as the judgment rule, and achieved good results. Barbosa et al. [15] used 41 hematological parameters from common blood tests as features and fed them into a random forest for predicting COVID-19. They set the number of tree is 90. They proved that using a simple machine learning approach could also achieve high performance. Sun et al. [16] presented a adaptive feature selection method to select the most important features and used four machine learning methods to obtain the experimental results. Wu et al. [17] presented a hyper machine learning framework including RF and SVM classifiers, and used a slime mould algorithm to optimize the loss function.

Deep learning based approach. Since deep neural networks based frameworks have recently dominated many AI and medical analysis related tasks, more and more researchers apply deep learning technologies to COVID-19 recognition as well. For example, Sethy et al. [18] combined deep learning with machine learn approaches for COVID-19 detection, which used CNN to extract deep features and fed them into a SVM classifier to perform prediction. He et al. [19] created a publicly-available dataset containing hundreds of CT scans, and thus designed a deep learning and transfer learning based COVID-19 detection approach. Song et al. [20] used chest CT scans of 88 patients diagnosed with the COVID-19 as dataset, and developed a deep learning-based CT diagnosis system. Their model performed very well. Serte et al. [21] used the ResNet-50, which is a strong image analysis pre-trained model, to detect COVID-19 infection from CT scans. In view of the popularity of deep learning, Alakus et al. [22] aimed to evaluate the predictive performance of several typical deep learning models, and their results showed that six deep learning methods overcome other machine learning methods. From the perspective of pre-trained language mode, Liu et al. [23] proposed to use the medical visual language BERT (Medical-VLBERT) model to identify COVID-19 and generate the medical report automatically.

Remarkable progress has been made in the current state-of-the-art. However, there is yet lack of mechanisms to capture the intra-modal and inter-modal interactions in multimodal medical records for COVID-19 detection.

III. Methodology

In this section, we detail the proposed MeqSeq2Seq model by presenting the key components.

A. Attention based Multi-Modal Representation

1) Textual Representation: For text, each word $w_i$ is initialized with pre-trained BERT embeddings. We thus feed them into a bidirectional Gated Recurrent Unit (BiGRU) to learn the contextual relationship between the words and the hidden states $H = [h^1, h^2, ..., h^n]$. To measure the contribution of the words, we use the attention mechanism and produce a weighted representation $h^t$, which can be formulated as:

$$
\alpha = \text{softmax} \left( W^T \tanh (W_d H + b_d) \right)
$$

$$
\hat{h}^t = \alpha H
$$

2) Visual Representation: For Chest CT, its feature vectors are extracted by the pre-trained EfficientNet network. We thus feed each image clip into the attention based GRU unit to produce its weighted visual representation $h^v$.

B. Multi-Head Inter-Modal Attention Fusion

We aim to fuse multi-modal information by learning a latent adaptation across modalities. Given textual and visual modalities $t$ and $v$ with their vectors $h^t$ and $h^v$, we treat textual modality as $Q$ query, i.e., $Q^t = W Q^t h^t$, and visual
modality as as Keys and Values, i.e., $K^v_{\mu} = W^v_{\mu} h^v$ and $V^v_{\mu} = W^v_{\mu} h^v$, where $\mu \in \{1, 2, \ldots, H\}$, $H$ is the number of heads. The mappings from $t$ to $v$ and $v$ to $t$ are defined as:

$$M^{t\rightarrow v}_\mu = \text{softmax} \left( \frac{Q^t_{\mu} K^v_{\mu}}{\sqrt{d_k}} \right) V^v_{\mu}$$
$$M^{v\rightarrow t}_\mu = \text{softmax} \left( \frac{Q^v_{\mu} K^t_{\mu}}{\sqrt{d_k}} \right) V^t_{\mu}$$

(2)

Eq. [2] will yield $H$ output values respectively. The reason to use multi-head attention is that different modalities may focus on different words/ clips. Then, we merge them together to obtain the multi-modal representation.

$$M^{(m)}_k = [M^{t\rightarrow v}; M^{v\rightarrow t}]$$

(3)

C. Medical Knowledge Augmentation

The participant gender, age, weight and medical history are the necessary knowledge for COVID-19 diagnosis. We represent such information using the pre-trained BERT embedding $h^{gen}, h^{age}, h^{his}$ and thus merge them with the multi-modal representation, which is defined as:

$$M^{(m)} = M^{(m)}_k \oplus [h^{gen}; h^{age}; h^{his}]$$

(4)

D. Classification. The fused representation $M^{(m)}$ is forwarded through the softmax function to yield COVID-19 label. We use the backpropagation method to compute the gradients and update all the parameters $\Theta$ by:

$$\Theta = \Theta - \lambda_t \frac{\partial J(\Theta)}{\partial \Theta}$$

(5)

where $\lambda_t$ is the learning rate. To avoid overfitting, we use a dropout strategy to randomly omit half of the feature detectors on each training case.

IV. Experimental Setup

Datasets. There is currently a lack of rich datasets for COVID-19 analysis. We choose COVID-CT-MD [1] to evaluate the proposed model in this work.

Evaluation metrics. We first remove the stop words using a standard stopword list from Python’s NLTK package. We do not filter out the punctuations since some punctuations, such as question marks and exclamation points, tend to carry medical information. We adopt precision (P), recall (R) and micro-F1 (M1-F1) as evaluation metrics.

Hyper-parameter Setup. All weight matrices are given their initial values by sampling from a uniform distribution $U(-0.1, 0.1)$, and all biases are set to zero. We use the Adam algorithm to train the network, and the number of epochs is set to 100. The coefficient of $L2$ normalization in the objective function is set to $10^{-5}$, and the dropout rate is set to 0.5.

Baselines. We compare our model with several strong baselines, including:

**HDNN** [24]: Hyper neural networks (HDNN) is a deep learning method that assigns two labels to each medical record.

**Multi-modal DCNN** [25]: We use two different CNNs to extract the textual and visual features respectively, and merge them together. Then a fully-connected layer is used to perform multi-modal detection.

**Bidirectional GRU** [26]: We implement a standard bidirectional GRU. The bidirectional GRU takes the sentence as input to obtain the hidden representation of each word.

**ResNet-34** [27]: We use a ResNet-34 which were originally trained on the ImageNet data set consisting of 3.2 million images for COVID-19 detection.

**VGG-16** [28]: We use an appropriate convolution layer (4th pooling layer) of the VGG-16 model to evaluate the performance.

A. Comparative Analysis

The experimental results are summarized in Table I. We observe that HDNN performs the worst among all baselines, which shows that deep fully-connected networks are insufficient to represent complex medical records. Multi-modal DCNN, BiGRU perform better than HDNN. The possible reasons are: (1) multi-modal DCNN could learn effective visual features; (2) BiGRU is able to learn contextual features for each word. However, their performance is worse than ResNet and VGG-16, because that we do not make much effort in tuning the parameters of CNN and BiGRU while finetuning the parameters of a neural network is an important process. ResNet and VGG-16 perform very well, which overcome the above mentioned approaches by a large margin. Text-MedSeq2Seq and Image-Seq2Seq perform not very well against VGG and ResNet, demonstrating that text or visual modalities cannot be treated independently for multi-modal COVID-19 detection.

The proposed MedSeq2Seq model achieves the best micro-F1 of 92.1% as compared to micro-F1 of 91.3% of the state-of-the-art system. This shows that the proposed MedSeq2Seq framework successfully leverages the advantages of intra-modal and inter-modal attentions in modeling multi-modal fusion.

Ablation Test. We perform an ablation study to further study the effectiveness of different components: (1) No $I_a$ Attention that removes the intra-modal attention; (2) No $I_e$ Attention that replaces the cross-modal attentive fusion with multi-modal feature concatenation; (3) No Attention that removes both $I_e$ and $I_t$ attentions from the whole framework.

---

TABLE I: Comparison of different models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>( P )</th>
<th>( R )</th>
<th>( M_{1-F1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDNN</td>
<td>83.3</td>
<td>81.7</td>
<td>82.1</td>
<td></td>
</tr>
<tr>
<td>Multi-modal DCNN</td>
<td>86.1</td>
<td>87.6</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td>BiGRU</td>
<td>84.9</td>
<td>83.5</td>
<td>84.2</td>
<td></td>
</tr>
<tr>
<td>ResNet</td>
<td>90.2</td>
<td>91.8</td>
<td>91.3</td>
<td></td>
</tr>
<tr>
<td>VGG-16</td>
<td>89.3</td>
<td>90.9</td>
<td>90.1</td>
<td></td>
</tr>
<tr>
<td>Text-MedSeq2Seq</td>
<td>88.7</td>
<td>89.2</td>
<td>89.2</td>
<td></td>
</tr>
<tr>
<td>Image-MedSeq2Seq</td>
<td>88.6</td>
<td>89.0</td>
<td>88.6</td>
<td></td>
</tr>
<tr>
<td>MedSeq2Seq</td>
<td>92.4</td>
<td>91.6</td>
<td>92.1</td>
<td></td>
</tr>
<tr>
<td>( \Delta )SOTA</td>
<td>(+1.8%)</td>
<td>(+1.4%)</td>
<td>(+1.3%)</td>
<td></td>
</tr>
</tbody>
</table>
The results in Table II show that the inter-modal $I_e$ attention contributes the most to overall performance.

V. Conclusions

Multi-modal COVID-19 diagnosis is an important and challenging medical AI task. We propose a medical knowledge enriched multi-modal sequence to sequence learning model. The key components include two attention mechanisms, viz. intra-modal ($I_a$) and inter-modal ($I_e$) attentions, and a medical knowledge augmentation layer. The experimental results show the effectiveness of the proposed MedSeq2Seq framework. Our future works will focus on designing an unified multi-task learning model to capture the correlation among triple or more tasks, e.g., headache, pneumonia, etc. How to design a better multi-modal fusion strategy to deal with text, image and audio is also left to our future work.

Acknowledgment

This work is supported by National Science Foundation of China under grant No. 62006212, the fund of State Key Lab. for Novel Software Technology in Nanjing University under grant No.KFKT2021B41.

References


TABLE II: Ablation experiment results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Models</th>
<th>Metrics</th>
<th>M, F1</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-CT-MD</td>
<td>No $I_e$, Attention</td>
<td>90.1</td>
<td>90.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No $I_a$, Attention</td>
<td>89.6</td>
<td>90.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No Attention</td>
<td>88.4</td>
<td>89.1</td>
<td></td>
</tr>
<tr>
<td>MeqSeq2Seq</td>
<td>92.1</td>
<td>91.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
