Improving Chest X-ray Report Generation by Leveraging Text of Similar Images

Saeed Niksaz[a], Fahimeh Ghasemian[a*]
[a]Department of Computer Engineering, Faculty of Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract
Automatic medical report generation is the production of reports from radiology images that are grammatically correct and coherent. Encoder-decoder is the most common architecture for report generation, which has not achieved to a satisfactory performance because of the complexity of this task. This paper presents an approach to improve the performance of report generation that can be easily added to any encoder-decoder. In this approach, in addition to the features extracted from the image, the text related to the most similar image in the training data set is also provided as the input to the decoder. So, the decoder acquires additional knowledge for text production which helps to improve the performance and produce better reports. To demonstrate the efficiency of the proposed method, this technique was added to several different models for producing text from chest images. The results of evaluation demonstrated that the performance of all models improved.

Keywords
Report generation; Chest X-ray; Similar image; Bert; CNN; Encoder-decoder

1) Introduction

Medical imaging refers to various technologies used to observe the human body to diagnose, monitor, or treat medical conditions. Each type of technology provides different information about the area of the body being studied or treated for possible disease, injury, or the effectiveness of medical treatment [1].

There are different types of medical imaging technologies [2]:

- Radiography: An image is recorded for subsequent evaluation. Mammography is a type of radiography to image the internal structure of breasts.
- Fluoroscopy: A continuous X-ray image is displayed on a monitor that allows real-time monitoring.
- CT: Many X-ray images are captured by the detector moving around the patient's body. A computer reconstructs all single images into slices of internal organs and tissues.

By exploring these images, doctors and specialists write a text (Figure 1) describing the abnormalities and essential points of a person's illness. This is error-prone for inexperienced specialists, and time-consuming for experienced doctors. Automated support for this task can ease clinical workflows and improve the quality and standardization of care [3].
A typical radiology report includes below information [4]:

- Type of exam
- History or reason for exam
- Comparison: if any previous exams are available, the radiologist compares them in this section.
- Techniques: This section explains how the exam is done and whether contrast is injected into the patient's vein.
- Findings: This section lists what radiologists observed in each part of the body. The radiologist will determine if the area is normal, abnormal, or potentially abnormal.
- Impression: In this section, the radiologist summarizes and reports the most critical findings and possible causes for those findings. This section provides the most essential information for

* Corresponding author: ghasemianfahime@uk.ac.ir

decision-making. Sometimes the report does not answer the clinical question and more exams may be needed.

-

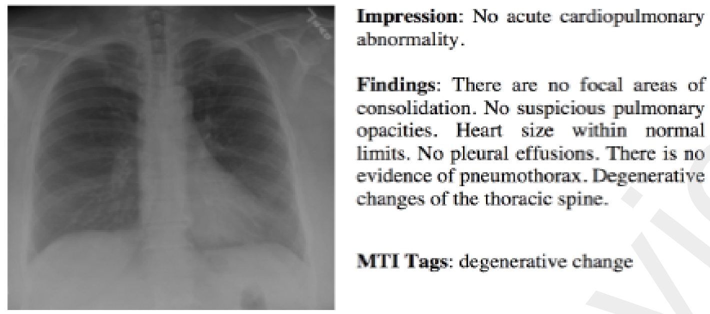

Figure1. A sample of a medical image and its report [28]

Automatic medical report generation is the production of reports from radiology images that are grammatically correct, and coherent. The report must include accurate information to diagnose, treat, and track patient progress in the report. Figure 2 shows the parts of a standard automated report generation system (encoder-decoder architecture [5]). In this system, first, the crucial areas of the image and the relationships between them must be identified using image processing techniques. Next, based on these findings, a text is produced as a report, which should be syntactically, semantically, and medically correct. Each of these parts is described in more detail below.

- Image Processing: In this section, a convolutional neural network is usually used to diagnose medical abnormalities. In this type of network, images are used as input, and the output is the features that are obtained from the image. These features are taken out of pixel mode and represented in a more meaningful format. Pretrained models like ResNeXt-101[6], and Resnet50[7] are good choices for this part.
- Text Generation: Once the visual features of the image have been obtained, a decoder is needed to produce sentences. These sentences can describe the content of the image. For this part RNN, LSTM, and transformers like GPT2 are appropriate. Also, for some decoders, a word embedding approach like Glove should be used for converting words to vectors.

Although much research has been done in the field of report generation from medical images, a significant result has not been achieved. It's because of the complexity of this task. This paper presents a comprehensive way to improve the performance of these models that can be easily added to any encoder-decoder model used to generate a report. In this method, in addition to the features extracted from the image, the text related to the most similar image in the training data set is also given as input to the decoder. In this way, the decoder acquires additional knowledge for text production, which helps to improve the performance and produce better text. To demonstrate the efficiency of the proposed method, this technique was added to several different models for making text from chest images. The results of training and evaluation on the Indiana University Chest Ray Collection [8] showed that the performance of all models improved.

Another factor affecting the performance of report generation systems is the method used to convert words to vectors in the decoder. We used BioBert [9] in this part for converting words to equivalent vectors. BioBert is a pre-trained language model for biomedical text mining. Evaluation results showed that it has higher performance for reporting than Glove [10].

The contributions of our work are summarized as follows:

- We adopted a transformer-based model called Bio-Bert for converting each word into a vector to feed into the model as the input.
- We proposed an approach that uses the report of the most similar image with the input image as another input to provide more information to the decoder part.

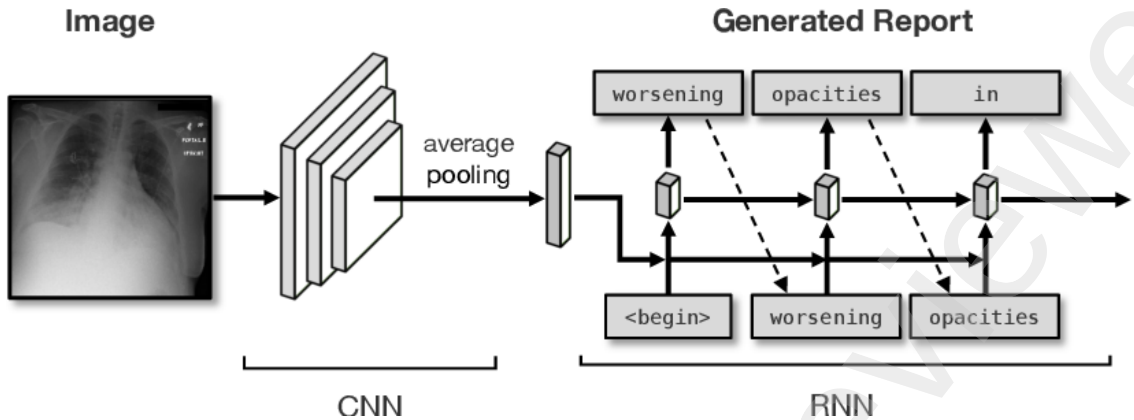

Figure 2. The architecture of an encoder-decoder model for report generation from medical images [5]

2) Related Works:

Many researchers consider automatic report generation as an image captioning task [11-14]. Nowadays, several studies have tried to use novel deep learning techniques to generate high-quality medical reports. Convolution-recurrent architectures (CNN-RNN) is the typical approach for automatic report generation tasks [15-17]. This architecture is composed of two main parts. An encoder based on convolutional neural networks (CNN) and a decoder based on recurrent neural networks (RNN) that are widely used in natural language processing, which capture the temporal information of textual sequences [18].

Different types of neural networks can be used to generate the caption for images. The methods can be grouped from different points of view: visual space versus multimode space, caption for dense areas versus caption for the whole image, supervised learning versus other deep learning Like reinforcement learning, and unsupervised learning, and encoder-decoder architecture versus hybrid architecture. New methods of caption generation use methods such as attention-based, concept-oriented, image style and object-based [19].

In many studies, semantic attention is added to visual attention [20]. A hierarchical LSTM decoder is used to manage generation at the paragraph level [21]. Also, some works adopted transformer-based models as decoders which are trained faster and provide advanced results in most NLP problems [22].

Medical report generation is considered a kind of image captioning task with extended captions. Several studies have been conducted to annotate medical images. Kisilf et al. [23] developed a pipeline to predict the features of medical images. Sheen et al. [24] adopted a CNN-RNN-based framework for predicting chest X-ray image tags (e.g., location and intensity). In the research of Zhang et al., [16] the aim is to create semi-structured pathology reports whose contents are limited to 5 pre-defined topics. Wang et al. proposed a novel Text-Image Embedding network (TieNet) for extracting the distinctive image and text representations. They used it in a reporting system and achieved high performance compared to some baseline methods [25].

Recently, natural language processing (NLP) models have changed from recurrent models to attention-based models, known as transformers, like Bert [26], because of their performance in different NLP tasks. With the widespread use of transformer-based models, NLP entered the transition learning phase, which also entered computer vision after the ImageNet data set came into being [27]. Most NLP research achieves superior results by fine-tuning a pre-trained transformer model on a large dataset. In another study, a transformer is pre-trained based on visual and semantic features for the production of medical reports. It includes semantic similarity metrics in the quantitative analysis of generated reports. A recent paper presents a conditional transformer-based model called CDGPT2 [28]. This model consists of an encoder that gives visual and semantic features from the image and a decoder to generate words. The encoder in the Chexnet model was fine-tuned to predict multiple tags from the image. The predicted scores of each tag are then multiplied by the corresponding pre-trained word2vec embeddings. The decoder is a pre-trained distilGPT2 [29] conditioned to generate reports based on visual features and tags.

3) Methods

As shown in Fig. 3, the model architecture consists of three major parts, the visual model for extracting features from medical images, the function to find the most similar image to the input image, and the decoder for generating a report. Below each part is explained.

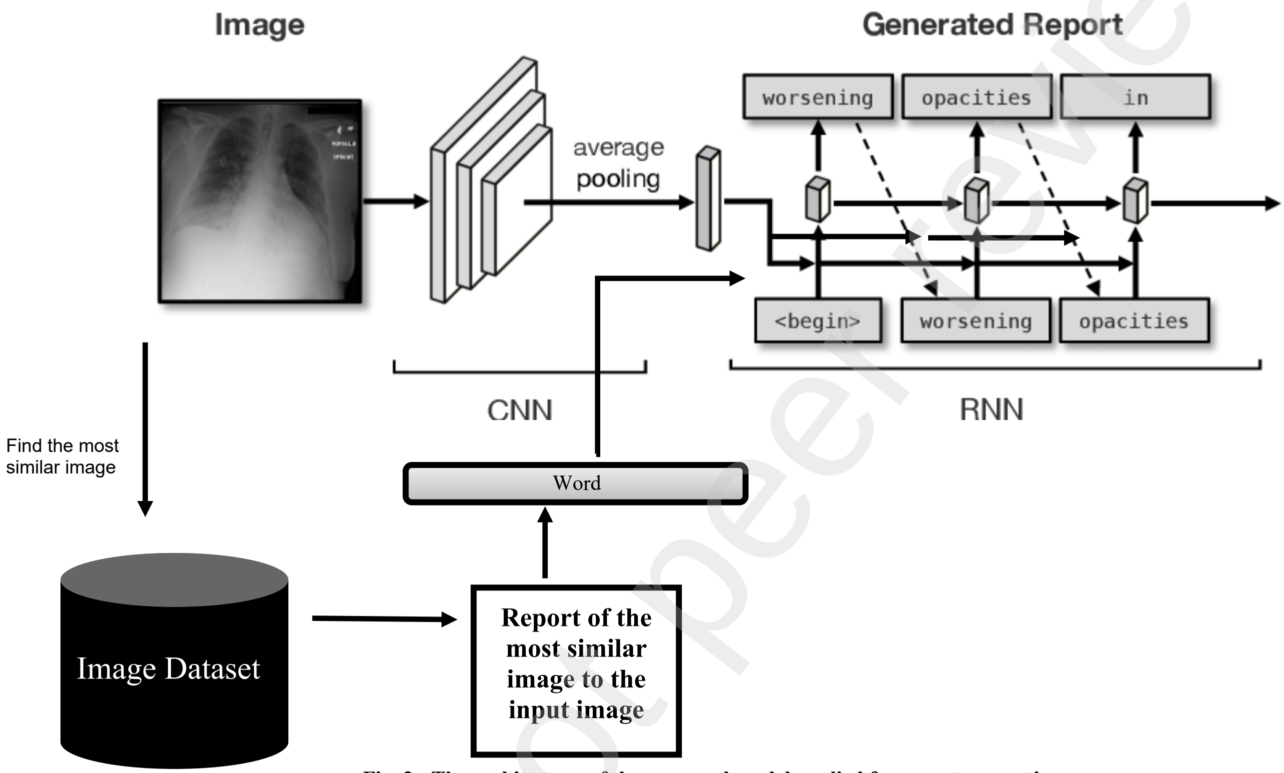

Fig. 3: The architecture of the proposed model applied for report generation

3-1) Feature extraction

Each medical image is fed into a CNN to extract features. In this paper, ResNeXt-101 and ResNet50, which both are based on the ResNet model, are used. The model's output is a vector of size 2048. ResNeXt repeats a block that aggregates a group of transformations with the same structure. It uses the same appraoch of VGG/ResNet to repeat layers. It is also similar to the Inception [30] module, but Inception has a different filter and size for every single block, while ResNeXt shares hyper-parameters among blocks. It is possible to configure the size of transformations dimension called cardinality. Increasing cardinality is a more efficient method to gain accuracy than going deeper or wider.

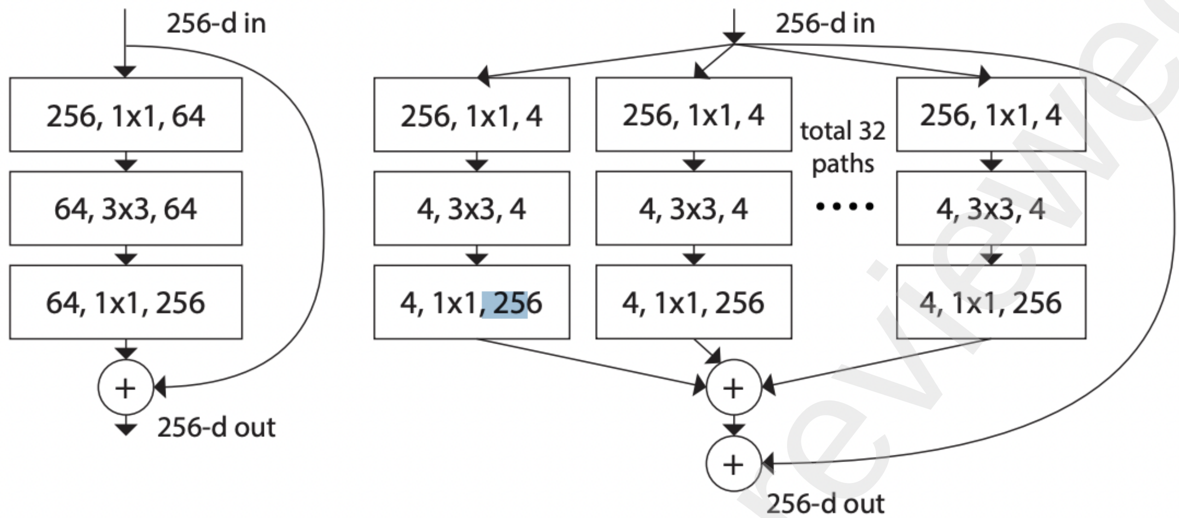

Figure 1. A block of ResNeXt with the cardinality of 32 [6]

Another model is Resnet which was proposed in 2015 by researchers at Microsoft Research. Before the introduction of this model, the use of neural networks with many layers was problematic. With the increase in the number of layers, the network would suffer from the vanishing gradient problem; the Resenet network was able to almost solve this problem by providing a solution. For this reason, this network can even have up to 152 layers. The technique was used in Resnet called skip connections, which connect activations of one layer to other layers by ignoring some layers in between. So, it causes to form residual block. The result of this approach is that if a layer decreases the performance of architecture, then the network skips this layer by regularization technique.

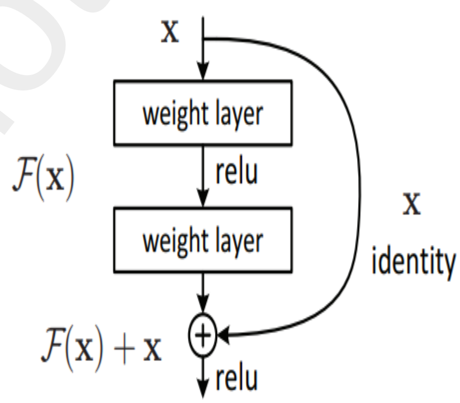

Figure 2. Skip connection [31]

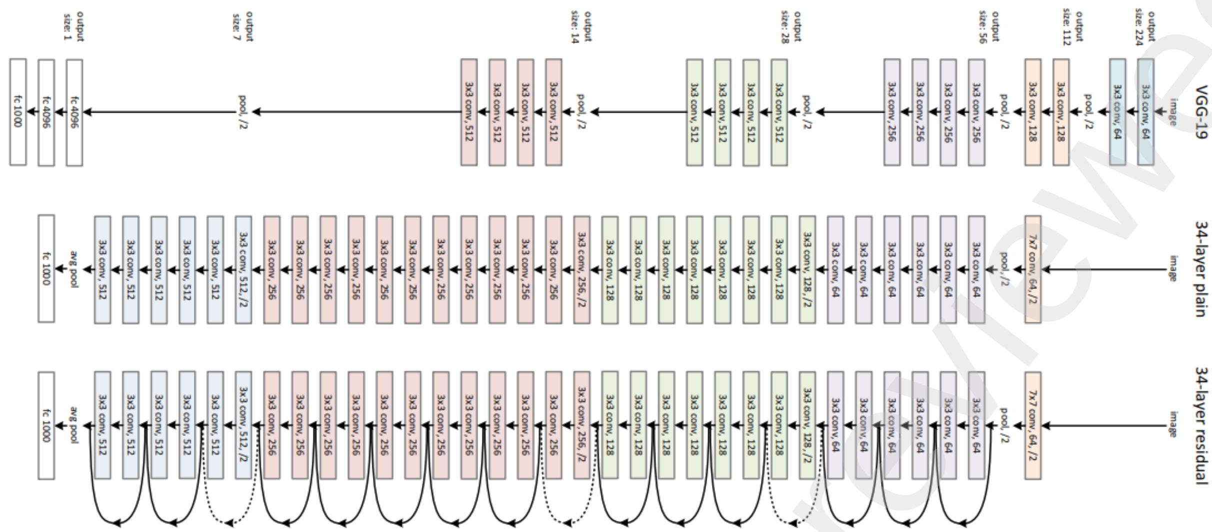

Figure 3. An example of adding skip connections to a 34-layer plain network architecture [31]

3-2) Finding the Most Similar Image

For every input image, a mechanism is used to find the most similar image in the train set. After that, we fetched its impression text and considered it the second input. Euclidean distance was used to determine the distances between image feature vectors from pre-trained models to vectors related to other images in the train set. Then, the impression text of the nearest image is selected.

$$distance = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \tag{1}$$

Where p and q are the vectors and I = 0, …, n indicates the number of elements in each vector.

3-3) Word Embedding

Also, to convert each sentence in impressions to integer sequences. First, a tokenizer is adopted. Then, integer sequences are converted into a fixed length. A weighted matrix of the embedding layer maps each word to a high dimensional space by using a pre-trained language model or a word embedding algorithm. Word embedding is used to convert the impression of the most similar image and the impression of the input image. Two different word embedding approaches are used in this paper:

- BioBert: Bidirectional Encoder Representations from Transformers for Biomedical text mining, is the first domain-specific language model trained on large-scale biomedical corpora (English Wikipedia, BooksCorpus, PubMed Abstracts, and PMC full-text articles). It has the same architecture as Bert and outperforms BERT and previous state-of-the-art models in some biomedical text mining tasks. BioBert, like Bert, converts each token into a 768-dimension vector. To the best of our knowledge, BioBert wasn't used for the task of medical report generation.
- Glove: is a model for word representation that captures the global corpus statistics. It is an unsupervised learning algorithm for converting words to vectors. It combines global matrix

factorization and local context window methods. The main insight of the model is the observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning. In this project, the Glove model embedded each token in a 300-length vector.

EXPERIMENTS AND RESULTS

Dataset

The Indiana University Chest X-ray Collection (IU X-Ray) is a set of chest x-ray images and their diagnostic reports. The dataset includes 7,470 pairs of images and reports. Each report consists of the following sections: impression, findings, comparison, and indication. In this paper, we treat the contents in impressions as the target captions to be generated (Figure 1 provides an example).

We preprocessed the data by converting all tokens to lowercases, removing all the non-alpha tokens (like '%', '$', '#', etc.) and erroneous tokens ("XXXX", "X-XXX)). On average, each image is associated with 2.2 tags, 5.7 sentences, and each sentence contains 6.5 words. All images are resized to 299*299 dimensions. Also, data augmentation is used to increase the number of images. Therefore, 22410 images were used. Finally, we randomly selected 21754 images for training, 250 images for validation, and 250 images for testing.

Implementation Details

We used the Adam (Kingma and Ba, 2014) [32] optimizer for parameter learning. Early stopping was used to prevent over-fitting. The model was implemented with python 3. For training the model, an NVIDIA Tesla K80 GPU provided by Google Colab, 13 GB of RAM, and 68 GB of hard disk was used. The models were trained for 30 epochs and 256 as batch size.

Visual Model: We used the pre-trained models, which are ResNext and Resnet, to extract features of each image. These models expect an image with 224*244 dimensions and output vectors. This vector is of size 2048. Deciding on which layer to extract from is a bit of a science. Early layers in the model usually learn low-level features, while higher layer learns more abstract features specific to the training data. Because the last layer of these models is a dense layer for detection of the input image class, this layer is ignored and the features of conv5_block3 located before the last layer were used.

Decoder: Our method provides three inputs to the decoder part. The first input is used to extract features from images using the ResNeXt pre-trained network. The next two inputs, which are the report of the corresponding image and the report of the most similar image to the current image. First, the word |startseq| is given to the model, then the successive word is predicted. In the next step, the two generated words are given along with the previous inputs to generate the third word. This process continues until the model generates the word |endseq|.

Results

The accurate reports are the ones that include most of the essential information and contain no false information. These reports could pass as reports written by experts. We use the BLEU metric to compare the word embedding models, and also, to evaluate the effect of our proposed model. The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for assessing a generated sentence to a reference sentence [33]. The score is used for evaluating the predictions made by the automatic machine. This measure is calculated by counting matching n-grams in the generated text to n-grams in the reference text, where a unigram compares each token, and a bigram compares each word pair. The comparison is made regardless of word order. We evaluated models with BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores. A best match results can be scored 1, while the results of a complete mismatch are scored 0.

Table 1. Comparison of different methods for the medical report generation task

Model	Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ResNet50 + GloVe	Without the report of most similar image	0.1469116	0.1137501	0.1006066	0.066392
ResNet50 + GloVe	**With the report of most similar image**	**0.1748103**	**0.1426615**	**0.1304903**	**0.08185959**
ResNet50 + BioBert	Without the report of most similar image	0.1650674	0.1406434	0.1303598	0.0926241
ResNet50 + BioBert	**With the report of most similar image**	**0.1923267**	**0.1575279**	**0.1429462**	**0.0903338**
ResNeXt+ GloVe	Without the report of most similar image	0.132935	0.096687	0.084244	0.048116
ResNeXt+ GloVe	**With the report of most similar image**	**0.154874**	**0.119754**	**0.107933**	**0.062116**
ResNeXt+ BioBert	Without the report of most similar image	0.178479	0.146880	0.135813	0.102407
ResNeXt+ BioBert	**With the report of most similar image**	**0.182991**	**0.149515**	**0.138966**	**0.102529**

For comparing the models, test data which consisted of 50 images and 50 reports related to each image, was used. As the result shown in Table 1, for automatic medical report generation, the results of with report of the most similar image outperformed the other approaches. We compared three methods and found that in all of them using the report of the most similar image as an input, had better outcomes. These results showed that the proposed method is effective in improving report generation and because it introduces more information into the decoder section, it can help improve the performance. Also, our results demonstrated that BioBert is a better embedding approach than GloVe. BioB mechanism, and has shown its superiority over other methods in text processing. the from seen be can As .information contextual encoding at better performed model language BioBERT the using models the ,table before come that words the from word given the of embedding the learns BioBERT because is result This words considers just Glove which text a in it after and located before the target wordcontex the as..

Conclusion

In this paper, an attempt been made to introduce a general method for improving the performance of CNN-RNN architecture for automatic medical report generation. In this method, in addition to the features extracted from the image, the text related to the most similar image in the training data set is also given as input to the decoder. In this way, the decoder acquires additional knowledge for text production which helps to improve the performance and produce better text. The results of training and evaluation on the Indiana University Chest Ray Collection showed that the performance of all models significantly enhanced. Also, different approaches for word embedding, including BioBert, and GloVe, were evaluated. Our result showed that BioBert, which is a language model based on the transformer, is a better approach for this task.

References

[1] https://www.fda.gov/radiation-emitting-products/medical-imaging/medical-x-ray-imaging

[2] https://www.radiologyinfo.org/en/info/article-read-radiology-report

[3] Monshi, Maram Mahmoud A., Josiah Poon, and Vera Chung. "Deep learning in generating radiology reports: A survey." *Artificial Intelligence in Medicine* 106 (2020): 101878.

[4] https://www.radiologyinfo.org/en/info/article-read-radiology-report

[5] Sun, Li, Weipeng Wang, Jiyun Li, and Jingsheng Lin. "Study on medical image report generation based on improved encoding-decoding method." In *International Conference on Intelligent Computing*, pp. 686-696. Springer, Cham, 2019. http://dx.doi.org/10.1007/978-3-030-26763-6_66

[6] Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492-1500. 2017. https://doi.org/10.1109/CVPR.2017.634

[7] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016. https://doi.org/10.1109/CVPR.2016.90

[8] https://openi.nlm.nih.gov/faq

[9] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36, no. 4 (2020): 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[10] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014. http://dx.doi.org/10.3115/v1/D14-1162

[11] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156-3164. 2015. https://doi.org/10.48550/arXiv.1411.4555

[12] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In *International conference on machine learning*, pp. 2048-2057. PMLR, 2015. https://doi.org/10.48550/arXiv.1502.03044

[13] Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. "Self-critical sequence training for image captioning." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008-7024. 2017. https://doi.org/10.1109/CVPR.2017.131

[14] You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651-4659. 2016.

[15] Shin, Hoo-Chang, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2497-2506. 2016. https://doi.org/10.48550/arXiv.1603.08486

[16] Zhang, Zizhao, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. "Mdnet: A semantically and visually interpretable medical image diagnosis network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6428-6436. 2017. https://doi.org/10.48550/arXiv.1707.02485

[17] Jing, Baoyu, Pengtao Xie, and Eric Xing. "On the automatic generation of medical imaging reports." *arXiv preprint arXiv:1711.08195* (2017). http://dx.doi.org/10.18653/v1/P18-1240

[18] Yuan, Jianbo, Haofu Liao, Rui Luo, and Jiebo Luo. "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 721-729. Springer, Cham, 2019. http://dx.doi.org/10.1007/978-3-030-32226-7_80

[19] Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. "A comprehensive survey of deep learning for image captioning." *ACM Computing Surveys (CsUR)* 51, no. 6 (2019): 1-36. https://doi.org/10.1145/3295748

[20] You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651-4659. 2016.

[21] Yin, Changchang, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network." In *2019 IEEE international conference on data mining (ICDM)*, pp. 728-737. IEEE, 2019. https://doi.org/10.1109/ICDM.2019.00083

[22] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

[23] Kisilev, Pavel, Eugene Walach, Ella Barkan, Boaz Ophir, Sharon Alpert, and Sharbell Y. Hashoul. "From medical image to automatic medical report generation." *IBM Journal of Research and Development* 59, no. 2/3 (2015): 2-1. https://doi.org/10.1147/JRD.2015.2393193

[24] Shin, Hoo-Chang, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2497-2506. 2016. https://doi.org/10.48550/arXiv.1603.08486

[25] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049-9058. 2018. https://doi.org/10.48550/arXiv.1801.04334

[26] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018). https://doi.org/10.48550/arXiv.1810.04805

[27] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

[28] Alfarghaly, Omar, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. "Automated radiology report generation using conditioned transformers." *Informatics in Medicine Unlocked* 24 (2021): 100557. https://doi.org/10.1016/j.imu.2021.100557

[29] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019). https://doi.org/10.48550/arXiv.1910.01108

[30] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826. 2016. https://doi.org/10.1109/CVPR.2016.308

[31] https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/

[32] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014). https://doi.org/10.48550/arXiv.1412.6980

[33] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318. 2002. http://dx.doi.org/10.3115/1073083.1073135