CS484/684 Group Project: Trustworthy Medical Report Training AI System

Ashi Garg/agarg22
Tianyi Ye/tye5
Yiqing Shen/yshen92

1 Introduction

To reduce the burdens in medical diagnosis, particularly in the clinical report context, broad applications of deep learning algorithms have been developed based on convolutional neural networks or Transformers. For example, DeepOpht [2] is a CNN-based report generator for retinal images. Interestingly, they incorporate the Gradient-weighted Class Activation Mapping (GradCAM) technique to improve the transparency of the modal, which verifies the importance of model explainability to the end users in medical report generation context [7]. However, one drawback of their work is the absence of user studies, which are neglected in most interpretable clinical report generation methods. Beside to DeepOpt [3], plenty of parallel works have demonstrated solutions for different modalities of medical images, including but not limited to X-ray, MRI, CT scan, etc [6].

However, the absence of users' trust from clinicians marginally restricted a broad application of those report generation systems to real clinical practice. Moreover, the heavy regulations and ethical approval also make it hard to turn to a commercial product. Instead, rather than developing a clinical report AI generation system, we head for another direction in this project. To be more specific, we focus on applying the report generation model to help achieve the medical training purpose. In this case, interpretability is still paid attention to by the end user i.e., those who attempt to improve their medical reporting skills. To this end, we make the first attempt at an innovative application of DL to the field of clinical report generation, namely interpretable medical AI training system, that scores the reports made by medical students or young fellows and helps them improve their report writing and diagnosis skills. Afterwards, a user study is carried out to evaluate how the interpretation can contribute to improving the user's trust.

2 Methods

Problem Formulation. From the machine learning perspective, the problem is formalized as follows. Our system distinguishes itself from the existing work by a new problem formulation, namely, we take into account both the medical images and the reports and output the score to the end user's clinical reports. Subsequently, the output grade is accompanied by suggestions for improvement and also the interpretations of the score. The score aims to measure the semantic similarity between the input clinical report and the report by the senior domain expert, which is not available and thus generated by an accurate pre-trained neural network. The major goal of our framework is to help our targeted end users, namely the medical students, junior fellows, and all those who attempt to practice their report writing skills or diagnosis ability, learn from the scores along with the improvement suggestion provided.

Machine Learning Framework. The overall framework is depicted in Fig. 1. Firstly, we leverage a state-of-the-art pre-trained X-ray clinical report generator network *CvT2DistilGPT2*[5] to generate the clinical report[1]. Then, the generated report along with the user's input report will be fed a BioBERT [4], i.e. a biomedical language representation model pre-trained on an X-ray dataset, to embed the features and use a linear regressor to get the semantic similarity.

Techniques to Provide Interpretations. We provide three interpretations techniques for the score, namely (1) Text heatmap w.r.t. the score. For the explainability, we first use integrated gradients, a model-agnostic method similar to SHAP to get the attributions of each word of the combination of input text and generated text (named actual input)by computing the integrated gradient of the output w.r.t. the actual input and compare with the preset baseline input. In our case the baseline input are two exactly the same texts, thus the attributions of the actual input indicate how each word is related to a high semantic similarity between the user's report and the reference report. (2) Top-k similar reports associated with the input image, using the similarity score as the metric (k=5 in our work). Rather than synthesizing the new report which is both inefficient and less accurate, we use the retrieval techniques to lookup the similar reports from the whole IU X-ray dataset. (3) Top-k similar images associated with the input image (k=5 in our work), these are the images paired with the top-k similar reports.

Dataset. For the method implementation and evaluation, we leveraged the widely used public clinical report generation dataset, i.e. IU X-ray dataset [1][2] to train and evaluate the machine learning model. We follow the dataset partition in previous work [5]. As we use the pre-trained checkpoints from the previous works, we are not going to retrain the model, hence we only employ the test set to perform the user study.

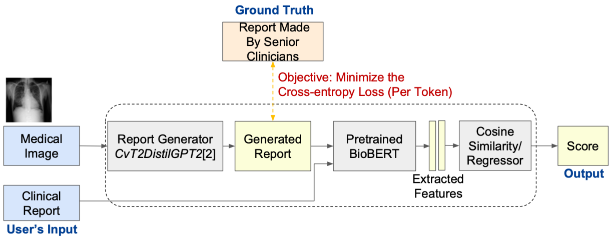
Ground Truth
Report Made By Senior Clinicians
Objective: Minimize the Cross-entropy Loss (Per Token)
Medical Image
Report Generator *CvT2DistilGPT2*[2]
Generated Report
Pretrained BioBERT
Extracted Features
Cosine Similarity/ Regressor
Score
Output
Clinical Report
User's Input

Figure 1: The overall framework of our proposed scoring system. The blue box indicates the user's input, and the yellow box highlights the system's output.

User Study Design. Our hypothesis to test via the user study is three-fold:

(1) Each individual interpretable component contributes to improving the user's truth to our grading system.

(2) Given all the interpretation techniques, the end users will gain more trust than receiving a sole interpretation.

(3) The more user truth about our grading system, the more confident they will feel to get benefits and improve their clinical reporting skills.

We measure the user's truth to our grading system at different conditions using the following user behavior metrics: (a) The subjective trust level (1-5), (b) How does each explanation improve the users' understanding of the prediction (1-5), we also collect the user's suggestions.

We collect the users' feedback under the following three conditions to the users: (a) No interpretation is provided, i.e. only the score itself. (b1-b3) one kind of interpretation is provided. The three techniques are tested with different sampled images sequentially and considered as three different conditions, i.e., each user is tested three times at this condition. (c) All interpretations are provided. The user is asked to respond to the questionnaire[3] after they are tested with one condition (a, b1, b2, b3, c). The overall pipeline for the user study is shown in Fig. 2

[1] We implement the score system use their public released codes and checkpoints at github.

[2] The dataset is publicly available at kaggle.

[3] The questionnaire is available at here

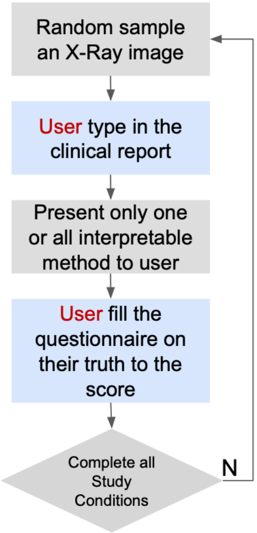
Random sample an X-Ray image
User type in the clinical report
Present only one or all interpretable method to user
User fill the questionnaire on their truth to the score
Complete all Study Conditions
N

Figure 2: The pipeline for the user study.

3 Results

Our user study is performed on a total number of 17 Johns Hopkins University affiliated medical students, who claimed they have learned the fundamental knowledge about X-ray diagnosis and clinical reporting. We discard the data collected from 1 student who claimed to have no background in image reporting All the involved 16 medical students are identified as our targeted users to perform the user's study.

As suggested in Fig. 3, the averaged truth level for baseline is 2.938, lower than those with one interpretation strategy i.e., 4.063, 3.563, 3.500 for improvement suggestion & text heatmap, similar image, and similar report. When given all the interpretations, the truth is further boosted to 4.375, which verifies our hypotheses 1 and 2. The questionnaire indicates that all the users claim to trust our system with either one of the interpretations presented or a combination of all the interpretations. The score for the contributions of each explanation indicates that the users rely on the reports of different scores to trust the system while the improved suggestion is more important than other components. We observe a highly correlated trend between the user's truth and their confidence in the model to help them improve their understanding of the clinical reporting, which thus verifies our hypothesis 3.

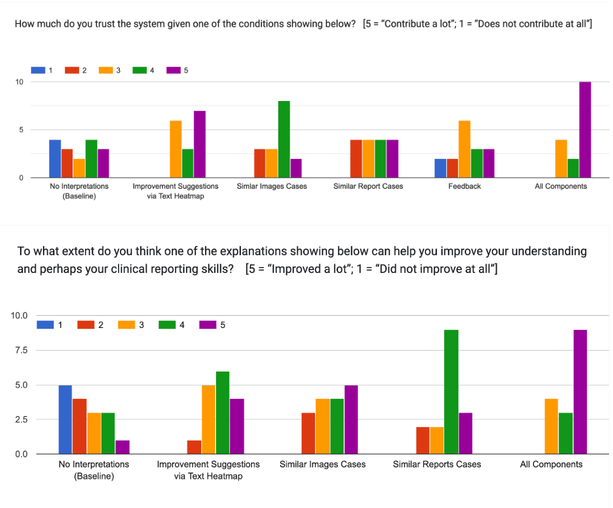

Figure 3: The result for the user study.

4 Discussion

Our user study result implies that all the aforementioned interpretable methods can help improve the users' trust in our grading system, and a joint application of all the methods can further boost the users' trust. Our results and finding suggest a promising potential for interpretable ML in the education field beyond the decision-making process.

References

[1] Dina Demner-Fushman et al. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[2] Jia-Hong Huang et al. Deepopht: medical report generation for retinal images via deep models and visual explanation. In *Proceedings of the*

IEEE/CVF winter conference on applications of computer vision, pages 2442–2452, 2021.

[3] Jia-Hong Huang et al. Non-local attention improves description generation for retinal images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1606–1615, 2022.

[4] Jinhyuk Lee et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[5] Aaron Nicolson et al. Improving chest x-ray report generation by leveraging warm-starting. *arXiv preprint arXiv:2201.09405*, 2022.

[6] Jarrel CY Seah et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health*, 3(8):e496–e506, 2021.

[7] Ramprasaath R Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.