CS482/682 Final Project Report Group 16

Biomedical Image Captioning

Weiting Tan, Chenyu Zhang, Jingyu Zhang, Jia Pan
{wtan12, czhan105, jzhan237, jpan26}@jhu.edu

1 Introduction

Background Medical imaging is widely used in clinical practice for diagnosis and treatment. Report-writing can be error-prone for inexperienced physicians, and time consuming and tedious for experienced physicians. To address these issues, automatic medical image captioning technology is very necessary. But still, to complete this task, there will be many challenges. For example, a complete report contains multiple heterogeneous forms of information. Abnormal regions in medical images are difficult to identify. Last, the reports are typically long, containing multiple sentences. For our project, we will develop our architecture to solve or alleviate such challenges.

Related Work MDNet [6] is a Resnet + LSTM encoder-decoder model, and was the first to use attention mechanism in biomedical image captioning. [1] also uses CNN-RNN network. For the CNN part, it first consists of a VGG19 to extract embedding visual representations. Then there is extra MLP layer in the encoder to predict the tag distribution. The decoder is hierarchical LSTM which consists of a word LSTM and a sentence LSTM and stop control.

2 Methods

Dataset Although labeled datasets are easy to obtain for general image captioning, publicly available datasets for biomedical image captioning are of smaller size. Currently, there are three main public available datasets for that we could utilize. The IU X-RAY dataset contains 7,470 X-Ray images and 3,955 associated reports, along with Medical Text Indexer (MTI) [1] extracted terms for each report. The PEIR Gross dataset and ICLEF-CAPTION dataset are elaborated in the Appendix. We use the IU X-Ray dataset, and obtain it using the script provided by [2]. The ”findings” and ”impressions” sections will be used for captioning.

Setup, Training and Evaluation Since previous work in biomedical image captioning did not have a unified evaluation metric, we conduct ablation study on a variety of features that have been used and studied various approaches to the problem.

First, we adopted the standard CNN-RNN encoder-decoder architecture. For the CNN encoder, we used VGG, Resnet, and the pre-trained CheXNet [4] (based on Densenet, originally developed for pneumonia detection and classification). For the RNN decoder, we used LSTM. One feature we tested is attention, where an attention layer is implemented between the encoder and decoder. Another feature is to add a MLP that predicts explicit tags from the CNN encoder as semantic features, and feeding those together with the image representations to the decoder via a visual-semantic co-attention mechanism [1]. This paper also proposed a hierarchical LSTM method, where a sentence LSTM is used to predict sentential semantics and a word LSTM is then used to generate tokens.

We also implemented a visual transformer, that breaks images into small patches and feed them into the transformer blocks as input, and get captions as

[1] https://ii.nlm.nih.gov/MTI/

output.

For evaluation, we used the BLEU-4 score offered by the sacrebleu package [3] to evaluate the generated captions. As proof of concept, we also tested a couple of our models on the COCO image captioning dataset.

A baseline model is the 1-NN image retrieval model that simply returns the caption of the most similar training image. [2].

3 Results

All of our models beat the nearest-neighbor baseline (see table below in the appendix). All of the models are able to generate coherent outputs (see sample output below). For the basic CNN-LSTM models, changing the CNN doesn't cause much difference in performance. Even with the pre-trained CheXNet encoder, the performance barely improved [2]. We think the reason is that without attention and semantic feature extractor, it is hard for the encoder model to learn features from the X-Ray images which all look alike.

We conducted experiments on the COCO dataset with the Resnet+LSTM model, both with and without attention. The Resnet+LSTM model with attention outperforms the one without attention, proving that the attention implementation was correct and useful. The BLEU-4 score for the model with attention is very high, although caution should be taken when interpreting the result, since we adapted this attention model from a public repository [3] and the metric calculation may not be exactly the same as our other models.

The model from [1] is reported as very effective, although we were unable to replicate the results exactly (following [4]). However, with qualitative analysis, the tagging MLP and hierarchical LSTM help with output diversity and produces reasonable captions.

Further parameter tuning could be done to get the best parameters for each model architecture.

Example output from ResNet+LSTM+Attn model: `<bos> no acute cardiopulmonary abnormality the lungs are clear bilaterally specifically no evidence of focal consolidation pneumothorax or pleural effusion cardio mediastinal silhouette is unremarkable visualized osseous structures of the thorax are without acute abnormality <eos>`

Example output tags: `effusion, lymphatic diseases, pulmonary disease, osteophytes`

4 Discussion

The fact that our encoder-decoder model is able to produce coherent and relevant sentences proves that it is the right architecture for image captioning tasks. However, the output is often repeating the same sentence for different images, showing that the model has not fully captured the abnormalities in the images. This is most likely due to that the IU X-RAY dataset is highly imbalanced: most images are normal chest X-Rays, so there is not much representation of abnormality. The captions are also quite repetitive, with only around 2000 unique captions among all 7,470 images. Therefore, it is difficult for the model to learn the association between the X-Rays and the captions.

In addition, the quantitative results of our visual transformer is not very promising. The reason may be that transformer models are too complicated and our IU X-RAY dataset is too small to fully train all its parameters.

References

[1] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. *CoRR*, abs/1711.08195, 2017.

[2] Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. A survey on biomedical image captioning. *CoRR*, abs/1905.13302, 2019.

[2] Part of the reason may be that the pre-trained version we downloaded is a bit corrupted.

[3] https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

[4] https://github.com/ZexinYan/Medical-Report-Generation

[3] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[4] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.

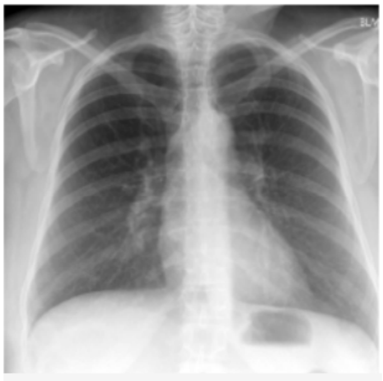
[5] Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4, 2016.

[6] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. *CoRR*, abs/1707.02485, 2017.

A Appendices

Two other datasets The PEIR Gross dataset contains 7,443 teaching images with associated sentences, along with top TF-IDF caption words. A larger dataset is the ICLEF-CAPTION dataset, which contains 232,305 samples with Unified Medical Language System [5] identified tags. Although this dataset is significantly larger than the previous two, it is obtained using an automatic extraction process and thus contains noise. While we didn't have time for this project, one of the future work could be to experiment with these two datasets as well.

Arch	BLUE-4 (COCO)	BLUE-4 (X-ray)
Baseline	-	5.7
VGG+LSTM w/o attention	-	15.2
ResNet+LSTM w/o attention	25.7	15.2
ResNet+LSTM w/ attention	27.9	39.9
CheXNet+LSTM w/o attention	-	15.4
CheXNet+LSTM w/ attention	-	39.1
Jing et. al. (co-attn, hier LSTM)	-	24.7
Visual Transformer	18.9	15.4



Ground Truth	**Our Model**
No active disease. The heart and lungs have in the interval. Both lungs are clear and expanded. Heart and mediastinum normal	No acute cardiopulmonary abnormality. The lungs are clear bilaterally. Specifically no evidence of focal consolidation pneumothorax or pleural effusion. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality