Attention-based Medical Caption Generation with Image Modality Classification and Clinical Concept Mapping

Sadid A. Hasan[1], Yuan Ling[1], Joey Liu[1], Rithesh Sreenivasan[2], Shreya Anand[2], Tilak Raj Arora[2], Vivek Datla[1], Kathy Lee[1], Ashequl Qadir[1], Christine Swisher[3⋆], and Oladimeji Farri[1]

[1] Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA
{firstname.lastname,kathy.lee_1,dimeji.farri}@philips.com
[2] Philips Innovation Campus, Bengaluru, India
{firstname.lastname}@philips.com
[3] Human Longevity, Inc., San Diego, CA, USA
{christinelswisher}@gmail.com

Abstract. This paper proposes an attention-based deep learning framework for caption generation from medical images. We also propose to utilize the same framework for clinical concept prediction to improve caption generation by formulating the task as a case of sequence-to-sequence learning. The predicted concept IDs are then mapped to corresponding terms in a clinical ontology to generate an image caption. We also investigate if learning to classify images based on the modality e.g. CT scan, MRI etc. can aid in generating precise captions.

Keywords: Caption Prediction, Concept Detection, Attention

1 Introduction

Automatically describing the content of an image is a key challenge in artificial intelligence at the intersection of computer vision and natural language processing. This could especially be beneficial to clinicians for useful insights and reduction of the significant burden on the overall workflow in patient care. The recent advances in deep neural networks have been shown to work well for large scale image analysis tasks [2, 4]. Hence, we use an encoder-decoder based deep neural network architecture [4] to address the task of medical image caption generation, where the encoder uses a deep CNN [2] to encode a raw medical image to a feature representation, which is in turn decoded using an attention-based RNN to generate the most relevant caption for the given image. We also utilize the same framework for clinical concept prediction to improve caption generation. Additionally, we investigate if learning to classify image modalities can aid in generating precise captions by efficiently capturing the specific characteristics of an image modality. Our experiments are conducted on an open access biomedical image corpus. The results show the effectiveness of our approach.

⋆ The author was affiliated with Philips Research at the time of this work.

2 Approach

We use an encoder-decoder-based framework that uses a CNN-based architecture to extract the image feature representation and a RNN-based architecture with an attention-based mechanism to translate the image feature representation to relevant captions [4] (Figure 1).

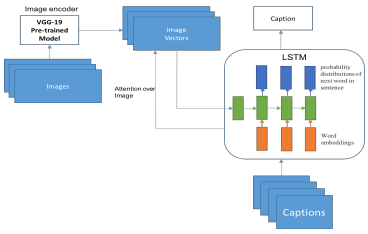

Fig. 1. The overall framework for medical image caption generation.

2.1 Image Encoder

We encode image features in two ways. First, we use the VGGnet-19 [2] deep CNN model (Figure 1) pre-trained on the ImageNet dataset [3] with fine tuning on the open access PubMed Central biomedical image corpus to extract the image feature representation from a lower convolution layer. Second, we modify the VGG-19 network architecture by including an additional softmax layer at the end for classifying medical images into N imaging modality classes including CT, MR, Ultrasound, X-ray, Pathology, Endoscopy etc. (Figure 2). The results of imaging modality classes are combined with other image features (directly learned using the pre-trained VGG-19 model from the medical images) into an image vector representation.

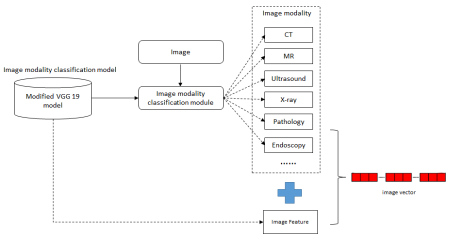

Fig. 2. Image vector representation with image modality classification.

2.2 LSTM-based Decoder

The decoder uses a long short-term memory (LSTM) network with a soft attention mechanism [4] that generates a caption by predicting one word at every time step based on a context vector (which represents the important parts of

the image to focus on), the previous hidden state, and the previously generated words. In particular, during training of the caption generation module, the image features are given as input to the first LSTM cell along with the first caption word, and the sequence of words are similarly passed along to the subsequent LSTM cells. Image weights are shared across all LSTM steps during the decoding stage to learn the association between image features and caption words. We use an attention mechanism over the image features in the decoder such that the caption words can learn the inherent alignments for important image regions without explicitly relying on segmentation information. Ultimately, the series of LSTM cells learns the probabilities of the next word given an input word and a medical image. The resulting model is able to generate a caption given a medical image.

2.3 Concept Mapping

To generate clinically relevant text, the training data should contain relevant clinical concepts embedded as part of captions. Because, biomedical images generally indicate certain anatomies, findings, diagnoses, location descriptors etc., which are usually available as clinical terms in a comprehensive ontology. Hence, it could be interesting to see if clinical concepts can be identified from the captions using a clinical NLP engine [7, 5] to prepare a dataset of biomedical images and their corresponding clinical concepts per image. Such a dataset can be utilized to formulate a clinical concept prediction task from images. We cast this task as a sequence-to-sequence learning problem. The predicted clinical concept IDs are later replaced by all possible terms from a clinical ontology such as UMLS metathesaurus to generate a caption of an image (Figure 3).

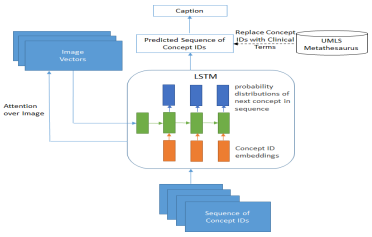

Fig. 3. Biomedical image caption generation with clinical concept mapping.

3 Experimental Setup

3.1 Corpus

We use the 2017 ImageCLEF caption prediction and concept detection task datasets [1] for our experiments. For the caption prediction task, the training data contained 164,614 biomedical images along with their captions extracted from PubMed Central. Furthermore, 10K images with captions were provided as

the validation set while 10K additional images were provided as the test set. The same collection was used for the concept detection task, except a set of clinical concepts is associated with each biomedical image instead of the caption.

3.2 Training

We use a one-hot vector approach to represent the words or clinical concept IDs in all models. Each LSTM in the decoder is built with 1024 hidden units. Our models are trained with stochastic gradient descent (SGD) using Adam as the adaptive learning rate algorithm and dropout as the regularization mechanism. The update direction of the SGD algorithm is computed using a mini batch size of 32 image-caption pairs. We use TensorFlow and a publicly available repository of encoder-decoder templates[4] for our experiments. Our models are trained with two NVIDIA Tesla M40 GPUs for approximately one month.

3.3 Models for Comparison

For comparison and analysis, we propose four models for biomedical image caption generation as follows: **Model1:** The entire training and validation sets are used to train this model without considering any semantic pre-processing of the captions, **Model2:** This model considers semantic pre-processing of captions using MetaMap [5] and the UMLS metathesaurus [6], initially trained on the modified VGG19 model with a randomly selected subset of 20K ImageCLEF training images to automatically generate image features and classify the imaging modality, and then finally trained with a random subset of 24K training images and 2K validation images to minimize time and computational complexity, **Model3:** This model is similar to Model1 with automatic generation of UMLS CUIs using the training dataset for the concept detection task, and then replacing the CUIs (generated for the test set) with the longest relevant clinical terms from the UMLS metathesaurus as the caption, and **Model4:** This model is similar to Model3 except we replace the CUIs with all relevant clinical terms (including synonyms) from the UMLS metathesaurus to generate a possible caption.

For the concept detection task, we prepared three models as follows: **Concept-Model1:** In this model, we consider the task as a sequence-to-sequence generation problem similar to caption generation, where the CUIs associated with an image are simply treated as a sequence of concepts, **Concept-Model2:** This model is created by simply transforming the generated captions (for the test set) from Model1 of the caption prediction task by replacing clinical terms with the best possible CUIs from the UMLS metathesaurus, and **Concept-Model3:** This model is created by simply transforming the generated captions (for the test set) from Model2 of the caption prediction task by replacing clinical terms with the best possible CUIs from the UMLS metathesaurus.

[4] https://github.com/yunjey/show-attend-and-tell

3.4 Evaluation and Analysis

The evaluation for the caption prediction task is conducted using BLEU whereas F1 score is used to evaluate the concept detection task. Table 1 and Table 2 show the evaluation results.

Caption Generation Models	Mean BLEU score
Model1	0.2638
Model2	0.1107
Model3	0.1801
Model4	**0.3211**

Table 1. Evaluation of caption prediction models

Concept Prediction Models	Mean F1 score
Concept-Model1	**0.1208**
Concept-Model2	0.0234
Concept-Model3	0.0215

Table 2. Evaluation of concept detection models

Ground Truth:	Pouchogram of the patient with entero-pouch fistula.
Model1	barium enema showing a distended rectum .
Model2	x ray showed radiolucent area
Model3	loops large intestine colon gastrointestinal tract small intestine
Model4	intestinum intestine colons colonic tenue intestinal large small colon intestines tract loops gastrointestinal bowel structure nos

Fig. 4. Example outputs of caption prediction from different models.

For the caption prediction task (Table 1), Model4 and Model1 achieved high scores denoting the effectiveness of our approach. Overall, our system was ranked first in the caption prediction task in ImageCLEF 2017 [8, 1]. Model4 is better as it includes all possible terms from the ontologies in the generated caption but trades-off the coherence of the caption. Hence, this approach increases the BLEU scores, which essentially computes exact word overlaps between the generated and the ground truth captions. Model2 likely suffered from the limited training data whereas Model3 has a lower score as it accepts only the longest possible clinical term as a replacement for a CUI in the caption. As evident from the example in Figure 4, Model4 generates the longest caption while compromising with the coherence aspect; however, we find its effectiveness in improving the BLEU scores justifying our hypothesis that concept mapping can indeed increase the coverage of words in a caption to improve its potential overlap with

the ground truth caption. Model2 is the only successful model to predict that Pouchogram is a type of X-ray test, showing the usefulness of image modality classification in generating a precise caption. However, Model2 states *a radiolucent area*, while the large intestine shown is radio-opaque. For Model1 we see that *barium enema* is a likely differential diagnosis. For the concept detection task (Table 2), Concept-Model1 performed reasonably well, but shows that there is still room for improvement. We may consider treating the task as a multi-label classification problem to achieve possible improvements. Concept-Model2 and Concept-Model3 were limited due to the 2-step translation of clinical terms to CUIs from the generated captions of the other task, which potentially indicates propagation of errors in learning the captions to the downstream task.

4 Conclusion

We presented an attention-based deep learning framework for caption generation from medical images. We also proposed to utilize the same framework for clinical concept prediction to improve caption generation. Our experiments conducted on an open access PubMed Central biomedical image corpus demonstrated that generating medical image captions by first predicting clinical concept IDs and then mapping them to all possible clinical terms in the ontology helps to improve the overall coverage of words in predicted captions compared to ground truth captions. Our experiments also revealed the usefulness of image modality classification in generating precise captions. In the future, we would extend this work by leveraging advanced deep learning algorithms and larger datasets.

References

1. C. Eickhoff et al. (2017). Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images, CLEF Labs Working Notes.
2. K. Simonyan and A. Zisserman (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556.
3. A. Krizhevsky et al. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS: 1106-1114.
4. K. Xu et al. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML.
5. A. R. Aronson (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA.
6. O. Bodenreider (2004). The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl 1):D267-D270.
7. V. Datla et al. (2017). Automated Clinical Diagnosis: The Role of Content in Various Sections of a Clinical Document. IEEE BIBM-BHI (pp. 1004-1011).
8. S. A. Hasan et al. (2017). PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection Tasks. Working Notes of CLEF.