PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection Tasks

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[1], and Oladimeji Farri[1]

[1] Artificial Intelligence Laboratory, 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

Abstract. In this paper, we describe our caption prediction and concept detection systems submitted for the ImageCLEF 2017 challenge. We submitted four runs for the caption prediction task and three runs for the concept detection task by using an attention-based image caption generation framework. The attention mechanism automatically learns to emphasize on salient parts of the medical image while generating corresponding words in the output for the caption prediction task and corresponding clinical concepts for the concept detection task. Our system was ranked first in the caption prediction task while showed a decent performance in the concept detection task. We present the evaluation results with detailed comparison and analysis of our different runs.

Keywords: Caption Prediction, Concept Detection, Encoder-Decoder Framework, Attention Mechanism

1 Introduction

Automatically understanding the content of an image and describing in natural language is a challenging task which has gained a lot of attention from computer vision and natural language processing researchers in recent years through various challenges for visual recognition and caption generation [1, 2]. Due to the ever-increasing number of images in the medical domain that are generated across the clinical diagnostic pipeline, automated understanding of the image content could especially be beneficial for clinicians to provide useful insights and reduce the significant burden on the overall workflow across the care continuum. Motivated by this need for automated image understanding methods in the healthcare domain, ImageCLEF[3] organized its first caption prediction and concept detection tasks in 2017 [3, 4]. The main objective of the concept detection task was to retrieve the relevant clinical concepts that are reflected in a medical image whereas in the caption prediction task, participants were supposed to

[3] http://www.imageclef.org/2017/caption

leverage the clinical concept vocabulary created in the concept detection task towards generating a coherent caption for each medical image.

The recent advances in deep neural networks have been shown to work well for large scale image processing, classification and captioning tasks. Specifically, the combined use of deep convolutional neural networks (CNN) with recurrent neural networks (RNN) has demonstrated superior performance for these tasks [5–11] based on the use of sequence to sequence learning and encoder-decoder-based frameworks for neural machine translation [12–14].

Motivated by the success of such prior works, we use an encoder-decoder based deep neural network architecture for the caption prediction task [9], where the encoder uses a deep CNN [5] 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 follow a similar approach to address the concept detection task by treating it as a text generation problem. Our system was ranked first in the caption prediction task while showed a decent performance in the concept detection task. In the next sections, we discuss the experimental settings, present the evaluation results with analysis, and conclude the paper.

2 Experimental Setup

2.1 Data

The training data contains 164,614 biomedical images with associated clinical concepts or captions extracted from PubMed Central[4]. Furthermore, 10K images per task are provided as the validation set while 10K additional images are provided as the test set for both tasks.

2.2 Training

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 [9]. We use the VGGnet-19 [5] deep CNN model pre-trained on the ImageNet dataset [6] with fine tuning on the given ImageCLEF training dataset to extract the image feature representation from a lower convolution layer such that the decoder can focus on the salient aspects of the image via an attention mechanism.

The decoder uses a long short-term memory (LSTM) network [15] with a soft attention mechanism [12, 9] 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.

[4] https://www.ncbi.nlm.nih.gov/pmc/

Our models are trained with stochastic gradient descent using Adam [16] as the adaptive learning rate algorithm and dropout [17] as the regularization mechanism. Our models were trained with two NVIDIA Tesla M40 GPUs.

2.3 Run Description

For the caption prediction task, we submitted four runs as follows:

- **Run1:** This run does not consider any semantic pre-processing of the captions; the entire training and validation set are used to train the model as described in Section 2.2.
- **Run2:** This run considers semantic pre-processing of captions using MetaMap [18] and the Unified Medical Language System (UMLS) metathesaurus [19], 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 as described in Section 2.2 with a random subset of 24K training images and 2K validation images to minimize time and computational complexity.
- **Run3:** This run is similar to Run1 with automatic generation of UMLS concept unique identifiers (CUIs) using the training dataset for the concept detection task, instead of the captions from the caption prediction task, and then replacing the CUIs (generated for the test set) with the longest relevant clinical terms from the UMLS metathesaurus as the caption.
- **Run4:** This run is similar to Run3 where we replace the CUIs (generated for the test set) with all relevant clinical terms (including synonyms) from the UMLS metathesaurus as the caption.

For the concept detection task, we submitted three runs as follows:

- **Run1:** In this run, 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; the entire training and validation set are used to train the model as described in Section 2.2.
- **Run2:** This run is created by simply transforming the generated captions (for the test set) from Run1 of the caption prediction task by replacing clinical terms with the best possible CUIs from the UMLS metathesaurus.
- **Run3:** This run is created by simply transforming the generated captions (for the test set) from Run2 of the caption prediction task by replacing clinical terms with the best possible CUIs from the UMLS metathesaurus.

2.4 Evaluation and Analysis

The evaluation for the caption prediction task is conducted using the well-known metric for machine translation, BLEU [20] whereas F1 score is used to evaluate the concept detection systems. Table 1 and Table 2 show the evaluation results.

We can see that for the caption prediction task, Run4 and Run1 achieved high scores denoting the effectiveness of our approach. Overall, our system was

ranked first in the caption prediction task. Run4 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 caption and the ground truth caption. Run2 likely suffered from the limited training data whereas Run3 has a lower score as it accepts only the longest possible clinical term as a replacement for a CUI in the caption.

For the concept detection task, Run1 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. Run2 and Run3 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.

Runs	**Mean BLEU score**
Run1	**0.2638**
Run2	0.1107
Run3	0.1801
Run4	**0.3211**

Table 1. Evaluation of caption prediction runs

Runs	**Mean F1 score**
Run1	**0.1208**
Run2	0.0234
Run3	0.0215

Table 2. Evaluation of concept detection runs

3 Conclusion

We presented the details of our participation in the caption prediction and concept detection tasks of the ImageCLEF 2017 challenge. Our system was ranked first in the caption prediction task and showed decent performance in the concept detection task. Overall, evaluation results showed the effectiveness of our approach. We highlighted potential reasons for errors in our submissions and discussed future work to consider for improved results.

References

1. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3): 211-252 (2015).

2. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. IEEE Trans. Pattern Anal. Mach. Intell. 39(4): 652-663 (2017).
3. Bogdan Ionescu, Henning Mller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba Garcia Seco de Herrera, Cathal Gurrin, Bayzidul Islam, Vassili Kovalev, Vitali Liauchuk, Josiane Mothe, Luca Piras, Michael Riegler, and Immanuel Schwall. Overview of ImageCLEF 2017: Information extraction from images. Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017, Springer LNCS 10456, 2017.
4. Carsten Eickhoff, Immanuel Schwall, Alba Garca Seco de Herrera, and Henning Mller. Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images, CLEF 2017 Labs Working Notes, CEUR Workshop Proceedings, 2017.
5. Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556, 2014.
6. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012: 1106-1114.
7. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CVPR 2015: 3156-3164.
8. Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. CVPR 2015: 2625-2634.
9. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML, 2015.
10. Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple Object Recognition with Visual Attention. ICLR, 2015.
11. Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of Visual Attention. NIPS, 2014.
12. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. ICLR, 2015.
13. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. NIPS 2014: 3104-3112.
14. Kyunghyun Cho, Bart van Merrienboer, aglar Glehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP 2014: 1724-1734.
15. Sepp Hochreiter, and Jrgen Schmidhuber. Long Short-Term Memory. Neural Computation 9(8): 1735-1780 (1997).
16. Diederik P. Kingma, and Jimmy Ba. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980 (2014).
17. Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1): 1929-1958 (2014).
18. Alan R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA 2001.
19. Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl 1):D267-D270, 2004.
20. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. ACL 2002: 311-318.