UMass at ImageCLEF Caption Prediction 2018 Task

Yupeng Su[1], Feifan Liu[2], and Max P. Rosen[2]

[1] Worcester Polytechnic Institute, Worcester MA 01609, USA ysu4@wpi.edu
[2] University of Massachusetts Medical School, Worcester MA 01655, USA
feifan.liu@umassmed.edu, max.rosen@umassmemorial.org

Abstract. This paper describes the details of our participation in the image caption prediction task at CLEF 2018. We explored and implemented an encoder-decoder framework to generate caption given a medical image. For the encoder, we compared two variations of convolutional neural networks (CNN) architectures: ResNet-152 and VGG-19, and for the decoder we used the long short term memory (LSTM) recurrent neural network. The attention mechanism was also experimented on a smaller sample to evaluate its impact on the model fitting and prediction performance. We submitted 4 valid runs and the best result achieved 0.18 BLEU score, which ranked second among all participant teams.

Keywords: Image Caption · Encoder-Decoder · LSTM · Convolutional Neural Network

1 Introduction

Automatically making computer understand the content of an image and offering reasonable description in natural language has gained more and more attentions from the computer vision and natural language processing community. In clinical practice, medical specialist and medical researchers usually write diagnosis reports to record microscopic findings from images, so automatic captioning on medical images will benefit healthcare providers with valuable insights and reduce their burden across the overall clinical workflow.

Due to its importance, ImageCLEF 2018 [5] continued the medical image captioning challenge [6] with the aim of advancing methodological development in mapping visual information from medical images to condensed textual descriptions. The main stream of Recent methods [2–4, 10] was combining the recurrent neural networks (RNN) for modeling natural language, with deep convoluational neural networks(CNN) for extracting image features. Within that framework, RNN based natural language generation was conditioned on CNN based image contextual features through encoder-decoder framework which was originally proposed for neural machine translation [1]. In the meantime, different deep architectures were also explored, including Deep Residual Networks [9] which created shortcuts between different layers to enable neural network to model the

identity mapping, and Wide Residual Networks [8] which shows that the widening of residual networks blocks can lead to improved performance compared to increasing their depth. More recently, Zhang et al. applied MDNet [7] on medical image captioning, where the image model utilizes an enhanced multi-scale feature ensemble method and the language model adopts an improved attention mechanism, outperforming other comparative baselines. It is worth noting that the caption in their clinical dataset was generated through a special guideline which requires additional manual efforts.

In participating the CLEF caption prediction task, we adopted the proven successful Encoder-Decoder architecture of neural networks, where we investigated two state-of-the-art image encoder variants (VGG-19 and ResNet-152) integrated with LSTM recurrent neural networks. Most existing studies use image encoders which were pre-trained on ImageNet, a very large dataset containing daily images from general domain. However, this pre-trained model is not available in medical domain. Therefore, we performed finetuning during end-to-end training of our model.

2 Dataset

The whole training dataset of Image CLEF caption prediction contains 222,314 image-caption pairs, and the length of captions varies from 1 word to 816 words. The average caption length is 20 words. Images are mainly from medical domain, but some of them are generic pictures such as animals or geography pictures. Caption itself is also noisy, e.g. some copy right or author information is also included. The official test data contains 9,938 images which are used to evaluate the system's performance.

3 Methods

In the encoder-decoder framework, the encoder extracts effective image features, based on which the decoder generates a sequence of textual terms for summarizing the image's content. As image understanding is crucial for caption prediction, we built two systems: VGG Attention Model using VGG-19 architecture as encoders integrated with the attention mechanism, and ResNet Model using residual networks as encoders. Those two encoders have shown promising results on image classification and are expected to work well for caption prediction. Both systems depend on LSTM model for decoding.

3.1 Preprocessing

To improve generality of model, we used random crop method to augment more data into original dataset. As we found that only 0.28%(628 instances) of the

training data contain captions with length larger than 145, we removed these data to reduce the sequence length required in LSTM model which can speed up the training process without losing much data representativeness.

3.2 VGG Attention Model

We adopted the architecture from the "show,attend, and tell" [10] as our VGG attention model, which is one of the state-of-art models for general domain image captioning task. It consists of a VGG-19 encoder and an attention based LSTM decoder where the soft attention mechanism was implemented. The architecture is illustrated in Fig. 1.

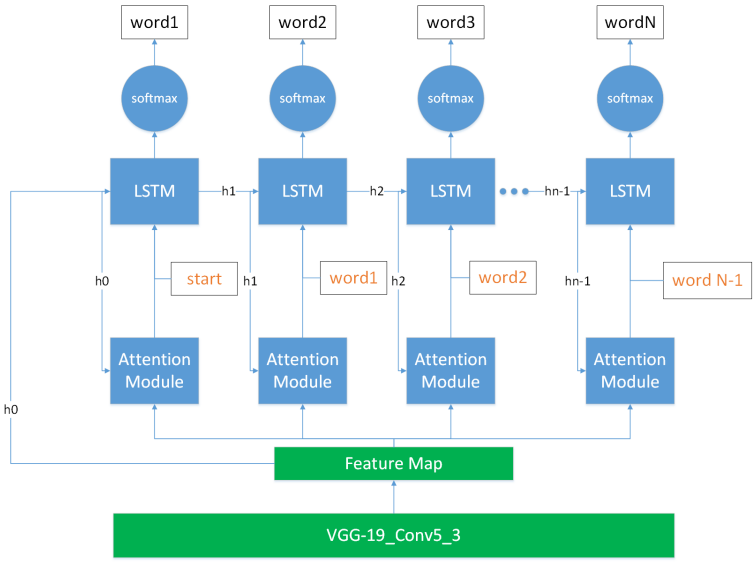

Fig. 1. Illustration of VGG Attention Model. We use VGG-19-conv5-3 to extract features from images and the initialization of hidden state (h0 in picture) is the average feature map. LSTM receives the current word embedding, and weighed feature map to predict the next word.

Encoder: VGG-19 Attention mechanism can guide the LSTM decoder to dynamically focus on specific parts of the image when generating the next caption term at each time step. Therefore we choose the output from the relatively lower convolution layer of the VGG net as image feature representation, which enables the model to attend to salient regions of the image. Specifically, we extracted features from the conv5_3 layer, which provides the 14 x 14 x 512 feature map (512 is the number of feature maps). The flattened 196 x 512 encoding will be fed in the subsequent LSTM decoder.

Decoder: Attention Based LSTM As shown in Fig. 1, the LSTM encoder receives the current word embedding and a dynamic image representation at each time step through the attention module. The attention module calculates different weights for feature map from different image locations, so that the LSTM encoder will know which location should be focused more for producing the next word. The weights are based on the hidden state of the LSTM at previous time step and the feature map extracted from the aforementioned VGG encoder(Fig. 1).

3.3 ResNet Model

Another model we explored is to use ResNet(residual neural networks) architecture in place of VGG network as image encoders. The motivation to swap VGG with ResNet is that the latter achieved state-of-the-art results in ImageNet classification tasks in general domain. It indicates that ResNet has the potential to produce a better encoding on image features. By adding shortcut of each block, residual network architecture also accounts for the vanishing gradients problem when training very deep convolution neural networks.

As shown in Fig 2, we adopted ResNet-152 encoder which directly connects to a standard LSTM decoder. Due to the depth of the architecture, we didn't add attention in this model as it is tricky to pick up an appropriate layer to attend to, which we will investigate in the future work. For implementation, we used pre-trained models' parameters to initialize ResNet-152 and performed end-to-end finetuning. Note that for both systems, the LSTM decoder doesn't use any pre-trained model.

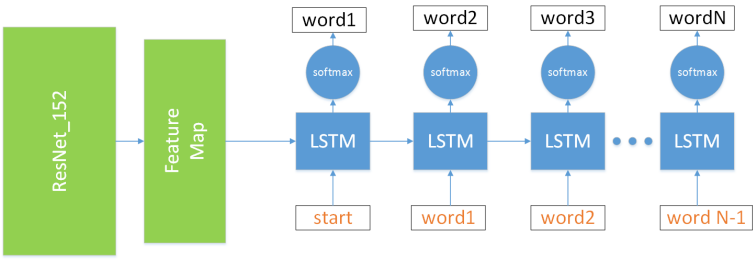

Fig. 2. Illustration of ResNet-based Model.

3.4 Over-fit a Small Dataset

One of the challenges for this caption prediction task is that the caption length varies largely, and the number of words per caption ranges from 1 to 145 even

with our preprocessing step. In last year's challenge, ISIA [11] implemented separate deep learning models to handle different lengths of captions, while the other team PRNA [12] developed one deep learning model to handle all of them. It shows that a single model may work well as long as it is properly parameterized. As the Hochreiter pointed out([13]), LSTM has capacity to handle the length over 1000. In order to determine some hyper-parameters, we sampled 10 instances from the training data whose captions contains words from 140 to 145. Using this small data, we tried to over-fit our deep learning model to empirically choose hyper-parameters by measuring its capacity of modeling this small data with long captions. Two hyper-parameters we are interested are the hidden size of the LSTM and size of word embedding, which plays an important role in mapping visual context features with textual terms.

We use the BLEU score as a metric to measure whether the model can overfit the small data properly. All the experiments are using Adam gradient descent method, and the learning rate is set as 0.0001. As we can see from Table 1, the capacity of the model to handle long length captions is more sensitive to hidden size than word embedding size, i.e. when the hidden size is not large enough(256), increasing the word embedding size doesn't help much(run1 vs. run2); when the hidden size is increased to 512, the model can fit well with embedding size of 256 or 512(run3 or run4). As expected, when embedding size is further reduced to 128, the capacity is adversely affected(run4 vs. run5). Based on this result, we choose the hidden size of 512 and the word embedding size of 256 in our systems. We also did this experiment on the attention LSTM model with VGG encoder and similar results were observed.

Table 1. Over-fit results of ResNet using a small data. Maximum number of epoch is 500.

Runs	Hidden and Word Dimension	Model	BLEU
1	256 x 256	ResNet_LSTM	0.132
2	256 x 512	ResNet_LSTM	0.124
3	512 x 512	ResNet_LSTM	0.908
4	512 x 256	ResNet_LSTM	0.923
5	512 x 128	ResNet_LSTM	0.324

3.5 Training

We used Adam stochastic gradient descent method for model training, with learning rate set as 0.0001, the maximum epoch set as 20, and the batch size set as 16. All our models were trained on two Tesla M60 GPUs in an Amazon EC2 server. However, attention based model training process is extremely slow, so we have no time to get a usable model before the challenge deadline. Therefore, our submitted runs were based on ResNet model.

4 Submitted Runs

We delivered totally 6 runs, and 4 of them are valid, so we only describe those 4 runs. As mentioned before, those 4 runs are all based ResNet encoder with standard LSTM decoder with different epochs of training. Table 2 shows our official test results.

- best_results_among_teams: the best results among all participants for caption prediction.
- run1_epoch_6: the run with 6 epochs of training
- run2_epoch_13: the run with 13 epochs of training
- run3_epoch_13_remove_UNK: unknown words (UNK) were removed from run2
- run4_epoch_13_remove_UNK_finetune: same setting with run3, except for allowing whole ResNet architecture to be finetuned in the training (other runs only finetuned the top 2 layers)

We can see that run1 delivered an under-fitted model obtaining the worst performance. With the same training epochs, no differences were found in terms of removing UNK or not, which may be due to the pre-processing step embedded in the evaluation script. Run4 achieved the best result of 0.18 demonstrating fully finetuning is beneficial.

Table 2. Official Test Results

Submitted Runs	BLEU	Rank
best_results_among_teams	**0.25**	1
run1_epoch_6	0.132	4
run2_epoch_13	0.176	2
run3_epoch_13_remove_UNK	0.176	2
run4_epoch_13_remove_UNK_finetune	**0.180**	2

5 Discussion and Conclusions

CLEF image caption prediction dataset is very different with COCO or Flickr 10k dataset. Some captions of this dataset are extraordinary long. Fig. 3 and Fig. 4 show two examples, one is from CLEF image caption prediction dataset, which has five sentences, and the other one is from COCO dataset, which only has one sentence.

The caption in Fig. 3 is very complicated, covering different semantics: a summary at the beginning followed by other descriptions as well as reasoning/inferring which can't be seen in the picture at all. In contrast, the caption from COCO dataset(Fig. 4), it only describes what can be seen from the picture. Thus, compared with the general domain VQA task, CLEF caption prediction is

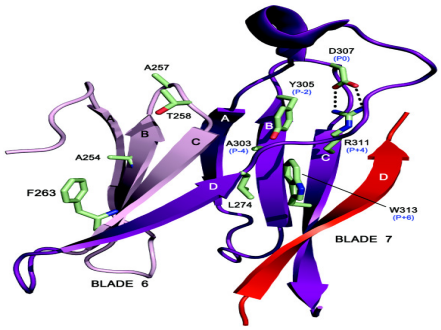

Fig. 3. Example: image of CLEF training dataset, **caption:** (1)Summary of key interactions in 'atypical' WD-repeats 6 and 7 of RACK1, illustrated with the structure of yeast Asc1p (PDB: 3FRX). (2)RACK1 proteins are characterised by a significant sequence extension between blades 6 and 7, leading to a knob-like protrusion from the upper face of the propeller. (3)The P0 Asp in blade 7 forms a salt bridge to an Arg at the P+4 position and this is packed against a Tyr (or in some orthologues Phe) in the P-2 position. (4) The P0 Asp is absent in blade 6. (5)In principle, these features together with the absence of the GH signature dipeptide on the loops between blades 5 and 6 and between blades 6 and 7 may facilitate structural transitions in this region of the protein that are important for function.

Fig. 4. Example: image of COCO dataset, caption:The man at bat readies to swing at the pitch while the umpire looks on

more challenging, not to mention the additional added complexity due to difficult medical terms. To address this, the model needs to be more intelligent with adequate reasoning ability, which may require more complex and hierarchical text modeling structure with support of background knowledge.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs, stat]. (2014).
2. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description. arXiv:1411.4389 [cs]. (2014).
3. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: Fully Convolutional Localization Networks for Dense Captioning. arXiv:1511.07571 [cs]. (2015).
4. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and Tell: A Neural Image Caption Generator. arXiv:1411.4555 [cs]. (2014).
5. Ionescu, B., Müller, H., Villegas, M., Herrera, A.G.S. de, Eickhoff, C., Vincent Andrearczyk, Cid, Y.D., Liauchuk, V., Vassili Kovalev, Hasan, S.A., Ling, Y., Farri, O., Joey Liu, Lungren, M., Dang-Nguyen, D.-T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation, Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), 2018.
6. García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 Caption Prediction tasks. In: CLEF2018 Working Notes. CEUR-WS.org <http://ceur-ws.org>, Avignon, France (2018).
7. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. arXiv:1707.02485 [cs]. (2017).
8. Zagoruyko, S., Komodakis, N.: Wide Residual Networks. arXiv:1605.07146 [cs]. (2016).
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs]. (2015).
10. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv:1502.03044 [cs]. (2015).
11. Liang, S., Li, X., Zhu, Y., Li, X., Jiang, S.: ISIA at the ImageCLEF 2017 Image Caption Task. 9.
12. Hasan, S.A., Ling, Y., Liu, J., Sreenivasan, R.: PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection Tasks. 5.
13. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. 9, 17351780 (1997).