IMAGE CAPTIONING USING TRANSFER LEARNING OF CNNs

Authors:
Y.Siva kumar, P.Nishanth, B.Gopi
Co-Author: Himanshu Sharma
Electronic and Communication Engineering Department
Lovely Professional University, Punjab

***Abstract*—**

Generating a caption from an image has attracted interests recently because of its importance in practical applications and also it connects two major artificial intelligence fields i.e., computer vision and natural language processing. Existing techniques on image captioning are top-down and bottom-up approaches. The top-down approach starts from gist of image and converts it into words. The bottom-up approach first come up with words that describe the image and combine them. In this paper, we proposed a model using transfer learning that generates a caption for an input image. This task involves computer vision and natural language processing concepts to recognize the context of an image and describe them in a natural language like English. To train our model, we have used Flickr8k dataset that contains around 8000 image (6000 training images, 1000 test images and 1000 validation images) and also each image contains 5 captions. Before training the model, first we extract all the features from the images. This feature extracting will be done through transfer learning. Our model works good not only at Flickr8k dataset but also at other datasets.

***Keywords*—Neural Networks, Imag, Caption, RNN, LSTM, Deep Learning.**

1 Introduction:

Automatically generating a natural language description of an image, a problem known as image captioning, has recently received a lot of attention in Computer Vision. The problem is interesting not only because it has important practical applications, such as helping visually impaired people see, but also because it is regarded as a grand challenge for image understanding which is a core problem in Computer Vision. Generating a meaningful natural language description of an image requires a level of image understanding that goes well beyond image classification and object detection. The problem is also interesting in that it connects Computer Vision with Natural Language Pro- cessing which are two major fields in Artificial Intelligence.

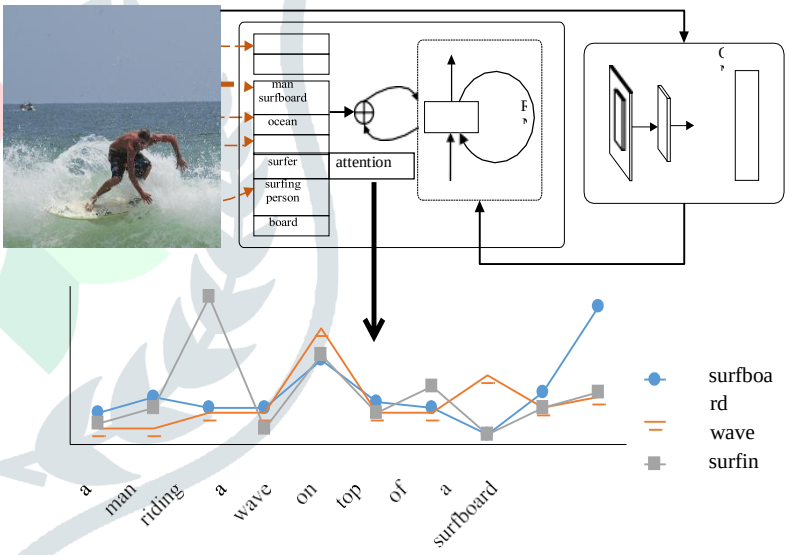

Figure 1. **Top**: an overview of the proposed framework. Given an image, we use a convolutional neural network to extract a top- down visual feature and at the same time detect visual concepts (regions, objects, attributes, etc.). We employ a semantic attention model to combine the visual feature with visual concepts in a re- current neural network that generates the image caption. **Bottom**: We show the changes of the attention weights for several candidate concepts with respect to the recurrent neural network iterations.

There are two general paradigms in existing image cap- tioning approaches: top-down and bottom-up. The top- down paradigm [4,35,26,16,8,36,25] starts from a "gist"

of an image and converts it into words, while the bottom-up one [12,19,23,9,20,11,22] first comes up with words describing various aspects of an image and

then combines them. Language models are employed in both paradigms to form coherent sentences. The state-of-the-art is the top- down paradigm where there is an end-to-end formulation from an image to a sentence based on recurrent neural networks and all the parameters of the recurrent network can be learned from training data. One of the limitations of the top-down paradigm is that it is hard to attend to fine details which may be important in terms of describing the image. Bottom-up approaches do not suffer from this problem as they are free to operate on any image resolution. However, they suffer from other problems such as there lacks an end-to-end formulation for the process going from individual aspects to sentences. There leaves an interesting question: Is it possible to combine the advantages of these two paradigms? This naturally leads to *feedback* which is the key to combine top-down and bottom-up information.

Visual attention [17,30] is an important mechanism in the visual system of primates and humans. It is a feed- back process that selectively maps a representation from the early stages in the visual cortex into a more central non-topographic representation that contains the properties of only particular regions or objects in the scene. This selective mapping allows the brain to focus computational resources on an object at a time, guided by low-level image properties. The visual attention mechanism also plays an important role in natural language descriptions of images biased towards semantics.In particular, people do not describe everything in an image. Instead, they tend to talk more about semantically more important regions and objects in an image.

2 Related work Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. Earlier methods first generate annotations (i.e., nouns and adjectives) from images (Sermanet et al., 2013; Russakovsky et al., 2015), then generate a sentence from the annotations (Gupta and Mannem,). Donahue et al. (Donahue et al.,) developed a recurrent convolutional architecture suitable for large-scale visual learning, and demonstrated the value of the models on three different tasks: video recognition, image description and video description. In these models, long-term dependencies are incorporated into the network state updates and are end-to-end trainable. The limitation is the difficulty of understanding the intermediate result. The LRCN method is further developed to text generation from videos (Venugopalan et al.,). Instead of one architecture for three tasks in LRCN, Vinyals et al. (Vinyals et al.,) proposed a neural image caption (NIC) model only for the image caption generation. Combining the GoogLeNet and single layer of LSTM, this model is trained to maximize the likelihood of the target description sentence given the training images. The performance of the model is evaluated qualitatively and quantitatively. This method was ranked first in the MS COCO Captioning Challenge (2015) in which the result was judged by humans. Comparing LRCN with NIC, we find three differences that may indicate the performance differences. First, NIC uses GoogLeNet while LRCN uses VGGNet. Second, NIC inputs visual feature only into the first unit of LSTM while LRCN inputs the visual feature into every LSTM unit. Third, NIC has simpler RNN architecture (single layer LSTM) than LRCN (two factored LSTM layers). We verified that the mathematical models of LRCN and NIC are exactly the same for image captioning. The performance difference lies in the implementation and LRCN has to trade off between simplicity and generality, as it is designed for three different tasks. Instead of end-to-end learning, Fang et al. (Fang et al.,) presented a visual concepts based method. First, they used multiple instance learning to train visual detectors of words that commonly occur in captions such as nouns, verbs, and adjectives. Then, they trained a language model with a set of over 400,000 image descriptions to capture the statistics of word usage. Finally, they re-ranked caption candidates using sentence-level features and a deep multi-modal similarity model. Their captions have equal or better quality 34% of the time than those written by human beings. The limitation of the method is that it has more human controlled parameters which make the system less re-producible. We believe the web application captionbot (Microsoft,) is based on this method.

Figure 2: The visual-semantic alignment method can generate descriptions of image regions. Figure from (Karpathy and Fei-Fei,).

Karpathy *et al.* (Karpathy and Fei-Fei,) proposed a visual-semantic alignment (VSA) method. The method generates descriptions of different regions of an image in the form of words or sentences (see Fig. 2). Technically, the method replaces the CNN with Region-based convolutional Networks (RCNN) so that the extracted visual features are aligned to particular regions of the image. The experiment shows that the generated descriptions significantly outperform retrieval baselines on both full images and on a new dataset of region-level annotations.

3. Approach In this section we describe our approach using recurrent neural networks. Our goals are twofold. First, we want to be able to generate sentences given a set of visual observations or features. Specifically, we want to compute the probability of a word wt being generated at time t given the set of previously generated words Wt−1 = w1, . . . , wt−1 and the observed visual features V . Second, we want to enable the capability of computing the likelihood of the visual features V given a set of spoken or read words Wt for generating visual representations of the scene or for performing image search. To accomplish both of these tasks we introduce aset of latent variables Ut−1 that encodes the visual interpretation of the previously generated or read words Wt−1. As we demonstrate later, the latent variables U play the critical role of acting as a long-term visual memory of the words that have been previously generated or read. Using U, our goal is to compute P(wt|V, Wt−1, Ut−1) and P(V |Wt−1, Ut−1). Combining these two likelihoods together our global objective is to maximize, P(wt, V |Wt−1, Ut−1) = P(wt|V, Wt−1, Ut−1)P(V |Wt−1, Ut−1). (1) That is, we want to maximize the likelihood of the word wt and the observed visual features V given the previous words and their visual interpretation. Note that in previous papers [27, 23] the objective was only to compute P(wt|V, Wt−1) and not P(V |Wt−1). rt is made as a filter through which we can sales of any country for a particular year. For example in the year 1999 North american sales are highest with 126.1 million dollors

4. EXPERIMENTS In this section, we design several groups of experiments to accomplish following objectives: • Qualitatively analyze and understand how bidirectional multimodal LSTM learns to generate sentence conditioned by visual context information over time. • Measure the benefits and performance of proposed bidirectional model and its deeper variant models that we increase their nonlinearity depth from different ways. • Compare our approach with state-of-the-art methods in terms of sentence generation and image-sentence retrieval tasks on popular benchmark datasets.

4.1 Datasets To validate the effectiveness, generality and robustness of our models, we conduct experiments on three benchmark datasets: Flickr8K [31], Flickr30K [43] and MSCOCO [21].

Flickr8K. It consists of 8,000 images and each of them has 5 sentence-level captions. We follow the standard dataset divisions provided by authors, 6,000/1,000/1,000 images for training/validation/testing respectively.

Flickr30K. An extension version of Flickr8K. It has 31,783 images and each of them has 5 captions. We follow the public accessible1 dataset divisions by Karpathy et al. [11]. In this dataset splits, 29,000/1,000/1,000 images are used for training/validation/testing respectively.

MSCOCO. This is a recent released dataset that covers 82,783 images for training and 40,504 images for validation. Each of images has 5 sentence annotations. Since there is lack of standard splits, we also follow the splits provided by Karpathy et al. [11]. Namely, 80,000 training images and 5,000 images for both validation and testing

4.2 Implementation Details

Visual feature. We use two visual models for encoding image: Caffe [9] reference model which

is pre-trained with AlexNet [14] and 16-layer VggNet model [33]. We extract features from last fully connected layer and feed to train visual-language model with LSTM. Previous work [39, 23] have demonstrated that with more powerful image models such as GoogleNet [37] and VggNet [33] can achieve promising improvements. To make a fair comparison with recent works, we select the widely used two models for experiments.

Textual feature. We first represent each word w within sentence as one-hot vector, w ∈ RK, K is vocabulary size built on training sentences and different for different datasets. By performing basic tokenization and removing the words that occurs less than 5 times in the training set, we have 2028, 7400 and 8801 words for Flickr8K, Flickr30K and MSCOCO dataset vocabularies respectively. Our work uses the LSTM implementation of [4] on Caffe framework. All of our experiments were conducted on Ubuntu 14.04, 16G RAM and single Titan X GPU with 12G memory. Our LSTMs use 1000 hidden units and weights initialized uniformly from [-0.08, 0.08]. The batch sizes are 150, 100, 100, 32 for Bi-LSTM, Bi-S-LSTM, Bi-F-LSTM and Bi-LSTM (VGG) models respectively. Models were trained with learning rate η = 0.01 (except η = 0.005 for Bi-LSTM (VGG)), weight decay λ is 0.0005 and we used momentum 0.9. Each model is trained for 18∼35 epochs with early stopping. The code for this work can be found at https:// github.com/ deepsemantic/ image captioning.

4.3 Evaluation Metrics

We evaluate our models on two tasks: caption generation and image-sentence retrieval. In caption generation, we follow previous work to use BLEU-N (N=1,2,3,4) scores [28]: BN = min(1, e 1− r c) · e 1 N PN n=1 log pn (16) where r, c are the length of reference sentence and generated sentence, pn is the modified n-gram precisions. We also report METETOR [18] and CIDEr [38] scores for further comparison. In image-sentence retrieval (image query sentence and vice versa), we adopt R@K (K=1,5,10) and Med r as evaluation metrics. R@K is the recall rate R at top K candidates and Med r is the median rank of the first retrieved ground-truth image and sentence. All mentioned metric scores are computed by MSCOCO caption evaluation server2 , which is commonly used for image captioning challenge3 .

4.4 Visualization and Qualitative Analysis

The aim of this set experiment is to visualize the properties of proposed bidirectional LSTM model and explain how it works in generating sentence word by word over time. First, we examine the temporal evolution of internal gate states and understand how bidirectional LSTM units retain valuable context information and attenuate unimportant information. Figure 6 shows input and output data, the pattern of three sigmoid gates (input, forget and output) as well as cell states. We can clearly see that dynamic states are periodically distilled to units from time step t = 0 to t = 11. At t = 0, the input data are sigmoid modulated to input gate i(t) where values lie within in [0,1]. At this step, the values of forget gates f(t) of different LSTM units are zeros. Along with the increasing of time step, forget gate starts to decide which unimportant information should be forgotten, meanwhile, to retain those useful information. Then the memory cell states c(t) and output gate o(t) gradually absorb the valuable context information over time and make a rich representation h(t) of the output data. Next, we examine how visual and textual features are embedded to common semantic space and used to predict word over time. Figure 7 shows the evolution of hidden units at different layers. For T-LSTM layer, units are conditioned by textual context from the past and future. It performs as the encoder of forward and backward sentences. At MLSTM layer, LSTM units are conditioned by both visual and textual context. It learns the correlations between input word sequence and visual information that encoded by CNN. At given time step, by removing unimportant information that make less contribution to correlate input word

References:

1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate.arXiv:1409.0473, 2014.

[2] K. Cho, B. van Merrienboer, C. Gulcehre , F. Bougares,H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder - decoder for statistical machine translation.In EMNLP, 2014.

[3] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP,

2013.
[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010.
[5] R. Gerber and H. - H. Nagel. Knowledge representation for the generation o f quantified natural language descriptions of vehicle traffic in image sequences. In ICIP. IEEE, 1996.
[6] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image - sentence embeddings using large weakly annotated photo collections. In ECCV, 2014.
[7] S. Hochreiter and J. Schmidhuber. Long short - term memory.Neural Computation, 9(8), 1997.
[8] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 47, 2013.
[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In arXiv:1502.03167, 2015.
[10] A. Karpathy, A. Joulin, and L. Fei - Fei. Deep fragment embeddings for bidirectional image sentence mappi ng. NIPS,2014.
[11] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual - semantic embeddings with multimodal neural language models. In arXiv:1411.2539, 2014.
[12] R. Kiros and R. Z. R. Salakhutdinov. Multimodal neural language models. In NIPS De ep Learning Workshop, 2013.
[13] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011.
[14] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y . Choi. Collective generation of natural image descriptions. In ACL, 2012.
[15] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi. Treetalk:Composition and compression of trees for image descriptions.ACL, 2(10), 2014.
[16] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web - scale n - grams. In Conference on Computational Natural Language Learning,2011
[17] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Explain images with multimodal recurrent neural networ ks. In arXiv:1410.1090, 2014.
[18] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. C.Berg, K. Yamaguchi, T. L. Berg, K. Stratos, and H. D. III.Midge: Generating image descriptions from computer vision detections. In EACL, 2012.
[19] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In NIPS,2011.
[20] R. Socher, A. Karpathy, Q. V. Le, C. Manning, and A. Y. Ng.Grounded compositional semantics for finding and describing images with sentences. I n ACL, 2014.
[21] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In NIPS, 2014.
[22] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. - C. Zhu. I2t: Image parsing to text description. Proceedings of the IEEE, 98(8) , 2010.
[23] Marc Tanti, Albert Gatt, Kenneth P. Camilleri :Where to put the Image in an Image Caption Generator .arXiv:1703.09137