Science in Information Technology Letters
Vol. 1., No. 1, May 2020, pp. 17-23
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

Automated image captioning with deep neural networks

Abdullah Ahmad Zarir [a,1], Saad Bashar [a,2], Amelia Ritahani Ismail [a,3,*]

[a] Department of Computer Science, International Islamic University Malaysia, Malaysia
[3] amelia@iium.edu.my
* Corresponding Author

ARTICLE INFO

Article history
Received 01 June 2019
Revised 11 June 2019
Accepted 10 January 2020

Keywords
Recurrent Neural Networks
Convolutional Neural Networks
Image Captioning

ABSTRACT

Generating natural language descriptions of the content of an image automatically is a complex task. Though it comes naturally to humans, it is not the same when making a machine do the same. But undoubtedly, achieving this feature would remarkably change how machines interact with us. Recent advancement in object recognition from images has led to the model of captioning images based on the relation between the objects in it. In this research project, we are demonstrating the latest technology and algorithms for automated caption generation of images using deep neural networks. This model of generating a caption follows an encoder-decoder strategy inspired by the language-translation model based on Recurrent Neural Networks (RNN). The language translation model uses RNN for both encoding and decoding, whereas this model uses a Convolutional Neural Networks (CNN) for encoding and an RNN for decoding. This combination of neural networks is more suited for generating a caption from an image. The model takes in an image as input and produces an ordered sequence of words, which is the caption.

1. Introduction

The ImageNet Challenge every year starting from 2010 and the MSCOCO Image Captioning Challenge in 2015 has given rise to new state-of-the-art performances in the image classification task. Through these competitions, we have introduced to Krizhevsky et al. [1], their network "AlexNet" is used in the computer vision problems. Following this in 2014, we were introduced to even deeper and wider networks such as VGGNet [2] and GoogLeNet [3]. The GoogLeNet introduced a new architecture called Inception, which started a new era in the field of image classification, which is also a fundamental block in this image captioning model. The most significant impact of GoogLeNet is its cheap use of computational power than its predecessor networks. On top of that, it has made Transfer Learning possible for large scale image classification given a very limited time, computation capacity, and cost-constrained environment.

After the introduction of the first Inception model, a lot of research started based on it, which led to a more robust and efficient version of the Inception model, commonly known as InceptionV2, InceptionV3, etc. This InceptionV3 model does the heavy lifting of recognizing different objects in an image, but for generating caption, a model must also explain how the objectivity between objects and their attributes and activities. In short, the relationship between these objects needs to be explained using natural language (like English), which makes the presence of language models a must-have requirement.

The architecture, followed by the winner of the 2015 MSCOCO Image Captioning Challenge, uses a single joint model that takes image I as input. The input is trained to maximize the order of the corresponding target words, which are each word from a particular dictionary that describes the image. The machine translation becomes the inspiration of this model, which converts source language sentences (S) into maximum target language translations (T). Translation using Recurrent Neural Networks (RNNs) is a simple method [4]–[6] and can achieve sophisticated performance. An 'encoder' RNN reads an S. The next step is to convert it into a fixed-length vector representation. It is used in the initial hidden state of the RNN decoder that produces T. In this study, and the RNN encoder is changed with a deep Convolutional Neural Network (CNN). CNN performed the same task as the previous RNN [7]. That is the reason for using CNN as an 'encoder' of images. The process of using pre-training to classify images and use the last hidden layer as input to the RNN decoder that generates sentences. This model is called the 'Neural Image Caption' (NIC) by the winners of the challenge.

The MNIST [8] was first introduced in the year 1998, which started the whole new era of image classification. Though in the beginning, it took a week-long to train even this dataset, we have come a long way from that in almost two decades. Now MNIST is used to teach different Machine Learning approaches to an image classification problem in a classroom environment. It can be trained in less than an hour. Following the success of data-driven approaches, newer and more detailed datasets were being produced. Introduction of CIFAR-10 [9] in 2009 and ImageNet [3] in 2014 was a big milestone in pushing forward the research on image classification. ImageNet consists of 1.2 million images distributed into 1000 categories. These image datasets are mostly meant for single object recognition in the given image. Soon more detailed dataset of images was available such as the MSCOCO [10] dataset, which has general caption text associated with the images. A significant thing about some of these datasets are that they are active and still growing.

Compared to image datasets, text data were and are more abundant. Text data has an inbuilt grammar in it, which helps to extract and mine specific text data. Which is why natural language processing and other related fields have advanced so far, credit goes to all the available text corpuses. The task of object recognition with a fully-connected network is that it ignores any spatial information about the image. It gives the same importance to all the pixels present in the image regardless of the relation with any neighboring pixels. Because of this, if the object in the image is not centralized, it fails to classify. This issue is resolved with convolutional neural networks [11] as it takes spatial information of the given image into consideration. This is done through three key concepts, which are local receptive fields, shared weights & bias, and pooling.

Among the latest architecture in object recognition, GoogLeNet, with its inception design, outperforms all the previous architectures. There have many optimizations on the Inception Model [12] that know its relative version number. In this paper, Inception V3 has been used to perform the task of classifying several objects, where natural language descriptions can enhance the generation of images from visual data in currents times. The Recognition, object detection, and advances in language generation systems are the triggers for all these successes. The detection of Farhadi et al. [13] able to convert a triplet of scene elements into text. The process uses powerful language models based on language parsing. These can be used for pictures 'in the wild'. However, it is designed by hand and is not too general for text creation.

Many works that discuss the completion of the image given from the problem of ranking descriptions [14]–[16]. Embedding images and text in the same vector space becomes the approach used. Later for the description of an image, it fetches the text that is closely positioned in the given vector space. These methods fail to properly describe unseen images. The NIC model creates a single network to describe images. This network is used to classify images using deep convolutional networks [17]. Besides, it is modeled in sequence using a combination of recurrent neural networks [18]. The training process uses RBB with a single "end-to-end" network. This mode is adopted by machine translators [4]–[6]. The difference is that convolutional networks process images from the starting of a sentence.

2. Method

The proposed NIC is to describe images with a neural and probabilistic framework. The development of statistical machine translation provided a robust sequence model. It can maximize the probability of correct translation by providing input sentences in "end to end" fashion, both the training process and its predictions. That is the reason the NIC model as the model chosen, given the picture (not the input sentence in the source language), the same principle of "translating" into the applied description.

Calculations process the probability of correct descriptions as in (1).

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta) \quad (1)$$

Where θ is the model parameter, I is an image, and S is correct caption. S represents any sentence has no length restriction. It is why the chain rule is used to model a joint probability of more than $S_0, S_1, \ldots, S_N$ where N is the length of this particle example as (2).

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t, I, S_0, \ldots, S_{t-1}) \quad (2)$$

The formula has components, such as (S,I) is a training sample pair and optimization of the sum of log probabilities. Recurrent Neural Net (RNN) is the best choice for modeling $p(S_t, I, S_0, \ldots, S_{t-1})$, for which the number of variable words conditioned upon $t-1$ is the fixed-length hidden state h_t. This memory is updated after the new input x_t uses a non-linear function f (3).

$$h_{t+1} = f(h_t, x_t) \quad (3)$$

For f, a Long-Short Term Memory (LSTM) net was used, For f a Long-Short Term Memory (LSTM) net was used, which has sophisticated work in translating sequentially.

Representation of the input image is done with a CNN. They are widely used and currently are state-of-the-art for object detection and recognition. The version of CNN used in this demonstration is the one that performed best on the ILSVRC 2014 classification competition [18]. It is based on the Inception V3 architecture. This model can generalize other tasks by means of transfer learning [19]. The associated words in this demonstration are represented in an embedding model [20].

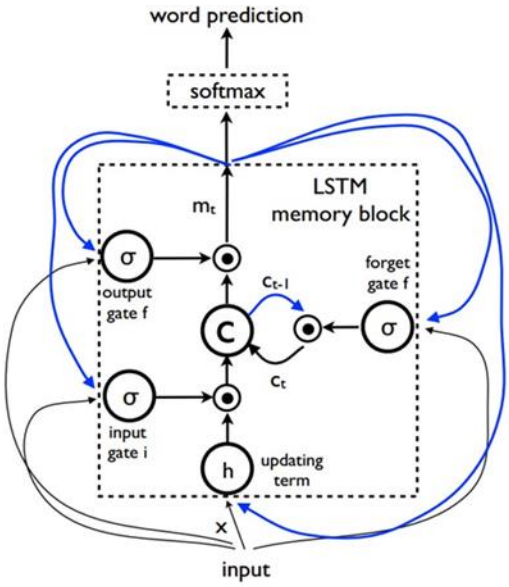

Fig. 1. LSTM: the memory block contains a cell c

Fig. 1. Indicates that LSTM: the memory block has a cell c, which three gates as control. The blue color is a recurring connections – the output m at time t-1 is fed back to the memory at time t through the three gates; the cell value are fed back through the forget gate; the word prediction at time t – 1 is fed back to memory output m at time t into the Softmax.

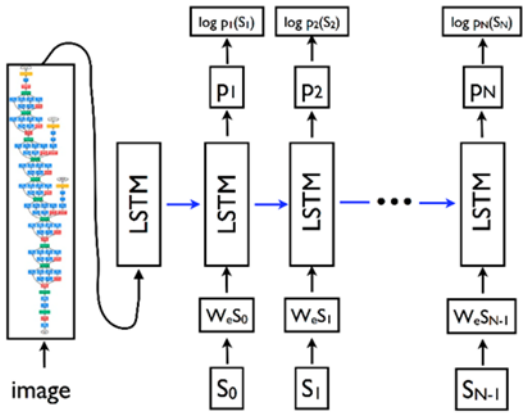

Fig. 2. LSTM model combined

Fig. 2. is a combination of CNN image embedder and word embeddings in the LSTM Model. The blue color in LSTM memory indicates connections unrolled, and they correspond to the recurrent connections in Fig. 1. All LSTMs share the same parameters.

2.1. Datasets

The training dataset consisted of images and sentences in English describing those images. A total amount of 82783 samples from the MSCOCO dataset were used. Another 40775 samples were separated for the test.

The validation dataset consists of 40504 samples from the COCO-4k dataset, and the result is based on this.

3. Results and Discussion

3.1. Evaluation Metrics

Not all translation processes match the correct sentence from the source language to the target language, as there can be many correct answers to that. In this scenario, evaluation is asking the raters to judge the usefulness of each description of the image subjectively, but obviously, this is time-consuming. For evaluation of the NIC model, both subjective and automatic metrics were reinforced and was shown that there is indeed some correlation between these two scorings. The process is following the guidelines in [21]. The graders assess each sentence produced with a range of 1 to 4 (shown in Table 1).

Table 1. Ranking scale.

Score	Meaning
1	Unrelated description.
2	Somewhat related description.
3	Minor errors.
4	No errors.

The matrix used in the image description is the BLUE score [22]. A BLUE score is a precise form of the word n-grams between the sentence produced and the reference. Besides, the confusion model calculates the geometric mean of the inverse probability of each word prediction. More recently, CIDER [23] was introduced in the use of the MSCOCO image writing challenge organization There are Metrics the validation was run on (Table 2).

Table 2. Scores in different Metrics

Model	Scores of		
	BLEU	METEOR	CIDER
NIC	27.7	23.7	85.5
NIC v2	32.1	25.7	99.8

4. Conclusion

There were many limitations in carrying out this work due to a lack of resources. Because of the advancement in transfer learning, it was possible only to train the last layer of the NIC model and do the demonstration. The scale in which machine learning research is going on right now it is fundamental to have a strong resource background to carry out tests after tuning different hyperparameters of any model. The results current models are producing are very promising, and it knows no indication of slowing down. This is just the beginning of the coming automation. NIC models so far can generate general descriptions of images. For future work, it can be made more capable in targeted descriptions.

Acknowledgment

This research is supported by the International Islamic University Malaysia under the Research Initiative Grants Scheme (RIGS): RIGS16-346-0510

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105, available at: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv Prepr. arXiv1409.1556*, Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556.

[3] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[4] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv Prepr. arXiv1406.1078*, Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.1078.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv Prepr. arXiv1409.0473*, Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473.

[6] I. Sutskever, O. Vinyals, and Q. V Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112, available at: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

[7] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," *arXiv Prepr. arXiv1312.6229*, Dec. 2013, [Online]. Available: http://arxiv.org/abs/1312.6229.

[8] Y. LeCun, C. Cortes, and C. J. C. Burges, "The MNIST database of handwritten digits," 1998. http://yann.lecun.com/exdb/mnist/.

[9] A. Krizhevsky, V. Nair, and G. Hinton, "The CIFAR-10 dataset," *2014*. https://www.cs.toronto.edu/~kriz/cifar.html.

[10] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, Springer International Publishing, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[11] Y. Lecun and Y. Bengio, "Convolutional networks for images, speech, and time-series," *Handb. brain theory neural networks*, vol. 3361, no. 10, p. 1995, 1995, available at: https://dl.acm.org/doi/10.5555/303568.303704.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1778–1785, doi: 10.1109/CVPR.2009.5206772.

[14] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2011, pp. 1143–1151, available at: https://papers.nips.cc/paper/4470-im2text-describing-images-using-1-million-captioned-photographs.

[15] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring Nearest Neighbor Approaches for Image Captioning," *arXiv Prepr. arXiv1505.04467*, May 2015, [Online]. Available: http://arxiv.org/abs/1505.04467.

[16] M. Kolář, M. Hradiš, and P. Zemčík, "Technical Report: Image Captioning with Semantically Similar Images," *arXiv Prepr. arXiv1506.03995*, Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.03995.

[17] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv Prepr. arXiv1502.03167*, Feb. 2015, [Online]. Available: http://arxiv.org/abs/1502.03167.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[19] J. Donahue *et al.*, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, 2014, pp. 647–655, available at: https://dl.acm.org/doi/10.5555/3044805.3044879.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector

Space," *arXiv Prepr. arXiv1301.3781*, Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781.

[21] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, Aug. 2013, doi: 10.1613/jair.3994.

[22] S. An, T. Bleu, O. G. Hallmark, and E. J. Goetzl, "Characterization of a Novel Subtype of Human G Protein-coupled Receptor for Lysophosphatidic Acid," *J. Biol. Chem.*, vol. 273, no. 14, pp. 7906–7910, Apr. 1998, doi: 10.1074/jbc.273.14.7906.

[23] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 4566–4575, doi: 10.1109/CVPR.2015.7299087.