Show and Tell: A Neural Image Caption Generator

Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan
{vinyals,toshev,bengio,dumitru}@google.com Google, Mountain View, CA, USA.

Automatically describing the content of an image using properly formed English sentences is a fundamental problem in artificial intelligence, but it could have great impact, for instance by helping visually impaired people better understand the content of images on the web. This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community [3]. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

In this paper, we present a generative model based on a deep recurrent neural network that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is composed of an image encoder, which is implemented as a deep convolutional neural network (DCNN) inspired by the latest winning entry of the ILSVRC 2014 classification competition [4], followed by a sentence decoder, which is implemented as a special type of recurrent neural network called Long-Short Term Memory network (LSTM) [1], that generates the sequence of words corresponding to the best sentence describing the image. An example is shown in Figure 1.

The model is jointly trained to maximize the likelihood of the target description sentence given the training image, as follows:

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

where θ are the parameters of our model, I is an image, and S its correct transcription. Since S represents any sentence, its length is unbounded. Thus, it is common to apply the chain rule to model the joint probability over $S_0,\ldots,S_N$, where N is the length of this particular example as

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

where we dropped the dependency on θ for convenience.

It is convenient to think of the LSTM in an unrolled form, where a copy of the model is created for the image and each sentence word as can be seen in Figure 2. The first copy of the LSTM reads the representation of the image as obtained by the CNN, and each subsequent LSTM copy receives as input an embedding representation of the previous word ($W_e S_t$), and produces as output a posterior probability of the next word (P_{t+1}).

Experiments are performed on five different benchmark datasets of various sizes, ranging from a few thousands to a million images. Our main results (see Tables 1 and 2) are reported in terms of BLEU score, which is typically used in automatic machine translation to estimate the quality of the proposed translations, as well as METEOR and CIDER. We used the MSCOCO dataset [2] with the tools that they kindly provided.

Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Table 1: Scores on the MSCOCO development set.

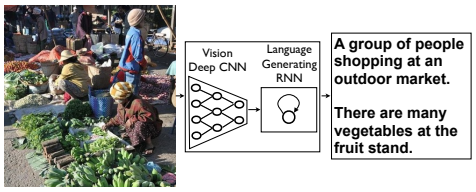

Figure 1: NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating LSTM.

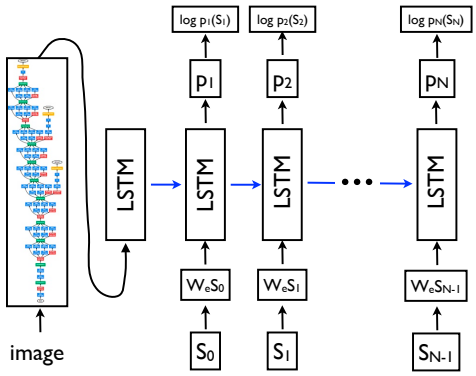

Figure 2: LSTM model combined with a CNN image embedder and word embeddings. The unrolled connections between the LSTM memories are in blue. All LSTM steps share the same parameters.

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text				11
TreeTalk				19
BabyTalk	25			
Tri5Sem			48	
m-RNN		55	58	
MNLM		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Table 2: BLEU-1 scores. We report previous work results when available (see full text for references). SOTA stands for the current state-of-the-art.

References

[1] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8), 1997.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *arXiv:1405.0312*, 2014.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *arXiv:1409.4842*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.