Oracle Performance for Visual Captioning

Li Yao[1]
li.yao@umontreal.ca
Nicolas Ballas[1]
nicolas.ballas@umontreal.ca
Kyunghyun Cho[3]
kyunghyun.cho@nyu.edu
John R. Smith[2]
jsmith@us.ibm.com
Yoshua Bengio[1]
yoshua.bengio@umontreal.ca
Frederico A. Limberger[1]
http://www.cs.york.ac.uk/~fal504
Richard C. Wilson[1]
http://www.cs.york.ac.uk/~wilson

[1] Université de Montréal
[2] IBM T.J. Watson Research
[3] New York University
[4] Dept. of Computer Science
University of York
York, UK

With standard datasets publicly available, such as COCO and Flickr in image captioning, and YouTube2Text, MVAD and MPI-MD in video captioning, the field has been progressing in an astonishing speed. For instance, the state-of-the-art results on COCO image captioning has been improved rapidly from 0.17 to 0.31 in BLEU. Similarly, the benchmark on YouTube2Text has been repeatedly pushed from 0.31 to 0.50 in BLEU score.

While obtaining encouraging results, captioning approaches involve large networks, usually leveraging convolution network for the visual part and recurrent network for the language side. It therefore results model with a certain complexity where the contribution of the different component is not clear.

Instead of proposing better models, the main objective of this work is to develop a method that offers a deeper insight of the strength and the weakness of popular visual captioning models. In particular, we propose a trainable oracle that disentangles the contribution of the visual model from the language model. To obtain such oracle, we follow the assumption that the image and video captioning task may be solved with two steps. Consider the model $P(\mathbf{w}|\mathbf{v})$ where $\mathbf{v}$ refers to usually high dimensional visual inputs, such as representations of an image or a video, and $\mathbf{w}$ refers to a caption, usually a sentence of natural language description. In order to work well, $P(\mathbf{w}|\mathbf{v})$ needs to form higher level visual concept, either explicitly or implicitly, based on $\mathbf{v}$ in the first step, denoted as $P(\mathbf{a}|\mathbf{v})$, followed by a language model that transforms visual concept into a legitimate sentence, denoted by $P(\mathbf{w}|\mathbf{a})$. $\mathbf{a}$ referes to *atoms* that are visually perceivable from $\mathbf{v}$. We define the configuration of $\mathbf{a}$ as an orderless collection of unique atoms. That is, $\mathbf{a}^{(k)} = \{a_1, \ldots, a_k\}$ where k is the size of the bag and all items in the bag are different from each other.

The above assumption suggests an alternative way to build an oracle. In particular, we assume the first step is *close to perfect* in the sense that visual concept (or hints) is observed with almost 100% accuracy. And then we train the best language model conditioned on hints to produce captions.

We consider a simple parametrization of $P(\mathbf{w}|\mathbf{a})$ with Long-short term memory networks (LSTMs) in Hochreiter and Schmidhuber [1]

$$\begin{bmatrix} p(\mathbf{w}_t \mid \mathbf{w}_{<t}, \mathbf{a}^{(k)})) \\ \mathbf{h}_t \\ \mathbf{c}_t \end{bmatrix} = \psi(\mathbf{h}_{t-1}, \mathbf{c}_{t-1}, \mathbf{w}_{t-1}, \mathbf{a}^{(k)}), \tag{1}$$

where $\mathbf{h}_t$ and $\mathbf{c}_t$ represent the RNN state and memory of LSTMs at timestep t respectively.

Despite its simplicity, the proposed model serves as a "performance upper bound" in visual captioning tasks. For the comparison of such oracle models with SOTA, please refer to the paper for details.

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.