The Summary of ”Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”[1] Paper

Furkan Gül

I. Motivation

Identifying information in an image and automatically generating captions related to that image is a challenging task. Caption generation models must be capable of satisfying two fundamental problems. The first issue related to the computer vision era is to understand which objects are lying in an image The second trouble regarding the natural language processing domain is to explain relationships between inferred objects in a natural language. That's why image captioning has been considered a crucial problem. Although it has a difficult nature as a task, recently there has been occurred an increasing demand for researching the caption generation problem. Before this paper was released, there have been huge improvements in the quality of the image captioning. However, recent approaches to solving the caption generation problem make use of both convolutional neural networks (CNNs) to extract feature vector embeddings of images and recurrent neural networks (RNNs) to decode those feature vector embeddings into a novel sequence of words or equivalently novel natural language sentences.

II. Approach

Previous works compressed whole images into vectorial representations using feature vector embeddings. It is a static representation. However, more dynamic representation generated by attention algorithm is more critical for bringing salient features in an image at the front as needed. Attention mechanism is more beneficial and becomes highly crucial to use when there exists much clutter in an image.

In this paper, two fundamental contributions are proposed clearly. Firstly, two different attention-based models for image caption generating are introduced under the same attention model architecture. These two attention variants are soft deterministic attention mechanism and hard deterministic attention mechanism. Secondly, the attention mechanism focuses on different parts of an image for each word in a sentence. By visualizing those parts of images corresponding to specific words, the performance of the model frameworks can be deduced intuitively. As a result of this paper, both correctly attended objects in an image and their corresponding words can be seen apparently in Figure 1.

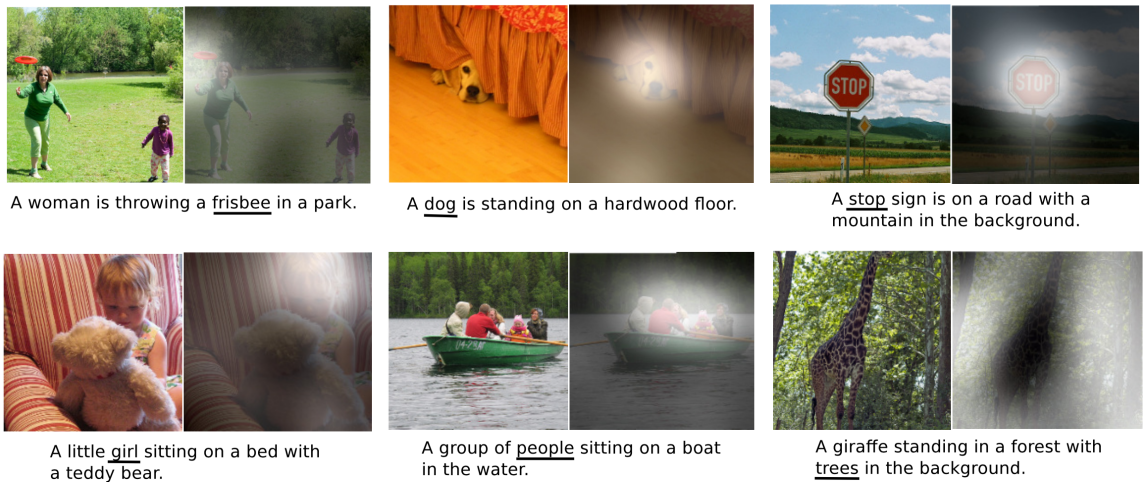

Fig. 1. Examples of attending to the ground truth class of objects (white shows attended regions of images, underlines showed related word in sentences) [1]

The common framework for two variants of attention model includes encoder and decoder parts. The overall model takes a single image and form a caption y which can be seen in Equation 1 below. The caption for each image consists of $\mathbf{y}$ encoded vectors of the size C (vectors are represented with bold font).

$$y = \{\mathbf{y}_1, ..., \mathbf{y}_C\}, \mathbf{y}_i \in \mathbb{R}^K (1)$$

where *K* denotes the vocabulary size and *C* represents the caption length. In encoder section, CNN is used to extract a set of feature vectors at different image locations of a single image. For just a single image, CNN generates *L* vectors. All those vectors are D-dimensional representations of all parts of the image. All *L* vectors concatenated as in Equation 2.

$$a = \{\mathbf{a}_1, ..., \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^D (2)$$

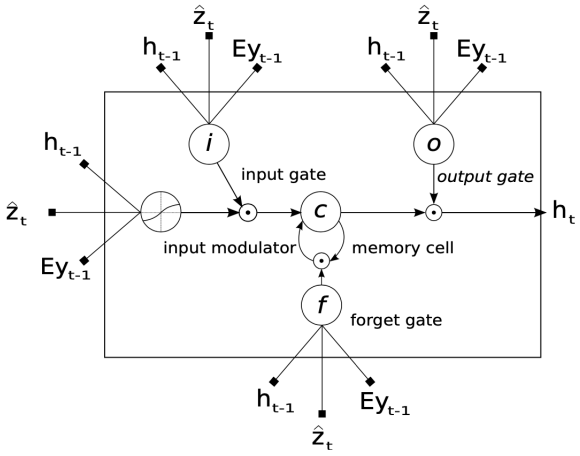

Fig. 2. A LSTM Cell in Paper[1]

In decoder section, very similar LSTM structure mentioned in Zaremba et al.[2] is used (see Figure 2). The equations for this LSTM can be found in Figure 3. $\boldsymbol{i_t}, \boldsymbol{f_t}, \boldsymbol{o_t}, \boldsymbol{g_t}$ represent the input, forget, output, and state update respectively. $T_{s,t} : \mathbb{R}^s \rightarrow \mathbb{R}^t$ represents a basic affine transformation with learned parameters. $\mathbf{E} \in \mathbb{R}^{m \times K}$ is an embedding matrix. Thus, $\boldsymbol{Ey_{t-1}}$ is an embedding vector (I think there is a mistake here. It should be $\boldsymbol{Ey_t}$). $\boldsymbol{h_{t-1}}$ and $\boldsymbol{c_{t-1}}$ are the previous hidden state and previous cell state respectively. $\boldsymbol{h_t}$ and $\boldsymbol{c_t}$ are the current hidden state and current cell state respectively. *n* and *m* are LSTM and embedding dimensionality respectively. $\hat{\boldsymbol{z}}_{\boldsymbol{t}} \in \mathbb{R}^D$ is the context vector. It is a dynamical representation of the corresponding patch of the input image at time *t*. ϕ function calculates $\hat{\boldsymbol{z}}_{\boldsymbol{t}}$ using features vectors in Equation 2. The only difference between soft attention mechanism and hard attention mechanism is the definition of that ϕ function. All remaining parts of the attention model for both variations are the same.

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} \mathbf{E}\mathbf{y}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{z}}_t \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$

Fig. 3. LSTM Equations in Paper[1]

In training phases of both attention model variants, stochastic gradient descent with adaptive learning rate is used. It is figured out that RMSProp[3] works best for the Flickr8k dataset[4] and Adam[5] works best for Flickr30k[6]&COCO[7] datasets. The Oxford VGGnet[8] is used as a CNN architecture for encoding part. It is pre-trained on ImageNet[9] data without finetuning. $14 \times 14 \times 512$ activation/feature map of forth convolution layer just before max pooling layer in the Oxford VGGnet is taken for decoder part. Thus, decoder (LSTM) processes on the flattened 196×512. Here, 196 stands for *L* and 512 represents *D*. For the regularization, besides droupout, early stopping is also used.

III. Dataset & Results

Flickr8K, Flickr30k, and MSCOCO datasets are used in their experiments. All three datasets contain 8000, 31000, 123000 images respectively. There are at least 5 sentence descriptions for every image in these three datasets. The vocabulary size for all their experiments are fixed and 10,000. Since the tokenization in Flickr8k and Flickr30k is just a basic tokenization, to be consistent with MSCOCO, the same basic tokenization is applied to MSCOCO. For reporting the results, BLUE (from 1 to 4) metrics and METEOR metric are used. Both are frequently used standard metrics in the image captioning literature.

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

Fig. 4. BLUE-1,2,3,4/METEOR metrics compared to other methods [1]

By looking at Figure 4, it seems that both soft and hard attention models do not make a noticeable difference in COCO dataset compared to previous models in the literature in terms of BLUE scores and METEOR. However, when the results on Flickr8k and Flickr30k datasets are considered, there is a slight improvement in almost all BLUE scores and METEOR by a large margin around 2. At the time of the publication year, around the end of 2015, these results were the best among all others.

IV. A Couple of Words

For now, details of both soft and hard attention models were not emphasized. After learning attention models in the class, this section will be revisited and clarified. The next paper to be summarized will be "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering"[10] from 2018.

References

[1] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, R. Salakhutdinov, R. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *ArXiv*, abs/1502.03044, 2015.

[2] W. Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *ArXiv*, abs/1409.2329, 2014.

[3] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5 - rmsprop. *Technical report*, 2012.

[4] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). *J. Artif. Intell. Res.*, 47:853–899, 2013.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[6] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[7] Tsung-Yi Lin, M. Maire, Serge J. Belongie, J. Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[8] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[9] Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, 2009.

[10] Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.