# Design Science Research
# 800 words essay
# GRU Image captioning

Frédéric Sallin

December 17, 2019

Generating captions for images can be done faster with gated recurrent unit.

Using a specific kind of recurrent neural network, gated recurrent unit (GRU) in the task of image captioning makes the training 20% faster than the current state of the art long short term memory (LSTM) based recurrent neural network. On MSCOCO dataset, The BLEU-4 score obtained is 34 (2% more than the LSTM state of the art mentioned above.)[1]

Image captioning is the task of automatically generating a natural language text describing an image. Humans are able to understand images without text descriptions quite easily, but computers need a way to go from an image to its text representation. This task is useful in many cases. Image indexing is a famous example. Being able to look for images based on their description is an important feature for search engines. Social networks use image captioning to analyze and categorize users based on the pictures they upload online.

Image captioning is the meeting of two sub-fields of artificial intelligence : computer vision and natural language processing. Their respective tasks needed to be solved for image captioning are object recognition and natural language text generation. At first, these tasks have been done separately. Creating complex systems capable of detecting objects in image and then combining them in structured language model.[3] However, a model solving both tasks at the same time in a fluent and non-separated way remains to be studied.

Deep learning has been a popular approach for image captioning. The ability of neural networks to automatically learn features has surpassed hand-designed models. Besides, the availability of big datasets of image and captions has increased the use of these networks.

We propose a model of image captioning consisting in a convolutional neural network (CNN) encoder that extracts the features of the image and feeds a gated recurrent unit neural network (GRU) that decodes and builds the caption. Convolutional neural network are widely for image object and features detection and recognition. Recurrent neural networks are popular in text generation since sentences are a sequence of words where each word has an impact on the next one.

The main goal of our model is maximizing the likelihood $p(S|I)$ of generating a sequence $S = S_1, S_2, S_3, ...$ of words describing the image $I$. Precisely, we want to find best probability for the correct caption given the image such that:

$$\begin{aligned}
\theta^* &= \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta) && (1)\\
\theta^* &\text{ are the parameters of the model} && (2)\\
I &\text{ is the image} && (3)\\
S &\text{ is the sentence} && (4)
\end{aligned}$$

The gated recurrent unit (GRU) is a special kind recurrent neural network defined such that:

$$\begin{aligned}
z_t &= \sigma_g \left(W_z x_t + U_z h_{t-1} + b_z\right)\\
r_t &= \sigma_g \left(W_r x_t + U_r h_{t-1} + b_r\right)\\
h_t &= (1 - z_t) \circ h_{t-1} + z_t \circ \sigma_h \left(W_h x_t + U_h \left(r_t \circ h_{t-1}\right) + b_h\right)\\
\text{where } x_t &\text{ is the input vector,}\\
h_t &\text{ is the output vector,}\\
z_t &\text{ is the update gate vector,}\\
r_t &\text{ is the reset gate vector,}\\
W, U, b &\text{ are the weights matrices and the bias vector,}\\
\sigma_g &\text{ is the sigmoid activation function,}\\
\sigma_h &\text{is the hyperbolic tan activation function}
\end{aligned}$$

It has fewer parameters than other recurrent neural network such as LSTM. It is its main valuable point. Having less parameters means less complex architecture and less complex computations.

The set used for our model is the MSCOCO data set. It consist of a training set of 82783 images with 4 captions for each image, a validation set of 40504 images and a test set of 40775 images. The metric used to describe the results is the BLEU score. The BLEU score gives the precision of correct generated words over the reference sentence. BLEU-4 means we use 4 words window to compute the precision. With our model, we have a BLEU-4 score of 34, which is better than the score of the state of the art of 32.[2]. Moreover, the training time is significantly reduced, 20% less than with a LSTM network. It means that it can be trained on less powerful system for the same quality of generated captions

In conclusion, we have presented a new model of image captioning based on a convolutional neural network and gated recurrent unit network which is quite efficient. Even if the results are convincing, this is only the start. Bigger datasets could lead to better image captioning.

# References

[1] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, June 2015, pp. 3156–3164. ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298935. URL: http://ieeexplore.ieee.org/document/7298935/ (visited on 07/08/2019).

[2] Oriol Vinyals et al. “Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (Apr. 1, 2017), pp. 652–663. ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2016.2587640. arXiv: 1609.06647. URL: http://arxiv.org/abs/1609.06647 (visited on 07/10/2019).

[3] Kelvin Xu et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. In: *arXiv:1502.03044 [cs]* (Feb. 10, 2015). arXiv: 1502.03044. URL: http://arxiv.org/abs/1502.03044 (visited on 07/10/2017).