Visual Imagination from Texts

Hanock Kwak and Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University
Seoul 151-744, Korea
Email: (hnkwak, btzhang)@bi.snu.ac.kr

***Abstract*—Imagination is a fundamental ability of humans which resides in the cognitive system. We propose a connectionist model that generates images from a given sentence after trained on a dataset of image-sentence pairs. The model is composed of language model and image model that are connected with a latent variable constrained by a prior distribution. The latent variable encodes dual information and it is generalized by Bayesian learning method. We trained on cartoon video series 'Pororo' and 16,066 fine-grained sentences describing short clips. Our model successfully generates plausible images which are highly correlated with a given sentence.**

I. Introduction

Images are composed of several different objects forming a hierarchical structure with various styles and shapes. Deep learning models are used to disentangle those complex underlying patterns [1][2], build distributed feature representations [3], and solve classification [4] and generation [5] problems using large datasets. The objective of model governs way of encoding complex visual informations as well as decoding latent variables.

Natural language offers a general, abstract and flexible descriptions of visual informations. Even language lacks detailed visual properties humans are able to imagine specific objects from abstract information depending on the experiences of what they seen before. Traditionally visual information about an object has been captured in attribute representations[6][7]. The categories of images are most simple one, and descriptions are more general. We connect this abstract representations with detailed visual informations generalized in real vectors.

Recently, deep convolutional and recurrent networks for text and image have successfully learned discriminative and generalizable representations automatically from raw data[8]. These approaches exceed the previous state-of-the-art methods that do not use neural networks. Inspired by these works, our model learns a direct mapping from sentences to image pixels using recent deep learning techniques.

II. Model

The model is combination of language model and image model that are connected with a latent variable constrained by a prior distribution. The language model is mainly made of LSTM [9] which encodes a given sentence and the image model uses transposed convolutional layers[5] that generate images from encoded vector. The model is illustrated in Figure 1.

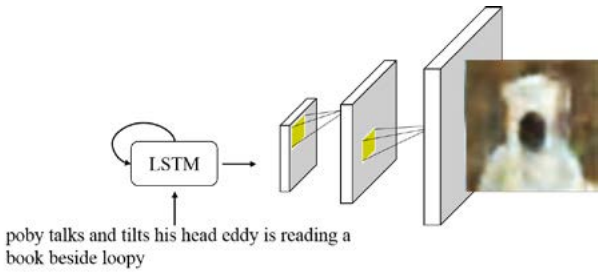

Fig. 1. **Illustration of the model.** Firstly, a sentence is passed in LSTM recurrently word by word. Then the final output of the LSTM is given to transposed convolutional layers which generates an image.

Let a given sentence be a vector $s = [w_1, w_2, ..., w_n]^T$ where w_i is a i-th word. Then a representation z of the sentence s is a final output vector of LSTM:

$$\begin{aligned} h_t &= LSTM(h_{t-1}, Uw_i), \\ z &= h_n. \end{aligned} \tag{1}$$

where h_t is a hidden state of LSTM at t step, and U is an embedding matrix. The final output image y is generated from z through transposed convolutional layers. We used mean square error between y and ground truth image as a loss function.

In addition, several recent deep learning techniques (batch normalization [10], ADAM [11], etc.), critical to the overall performance, were utilized.

III. Experiments

We trained on cartoon video series 'Pororo' and 16,066 fine-grained sentences describing short clips. The 'Pororo' has long play time of 1,232 minutes, various scenes, and some characters. All images are resized to 64×64 with antialiasing. The image model has series of four transposed convolution where channels are halved and size of the filter maps doubled for each time. Initially we set hidden state h_1 as zero vector and U is initialized with pre-trained Glove vectors[12].

The result is shown in Figure 2. The model successfully generated plausible images from sentences and it generalized dual representation of texts and images.

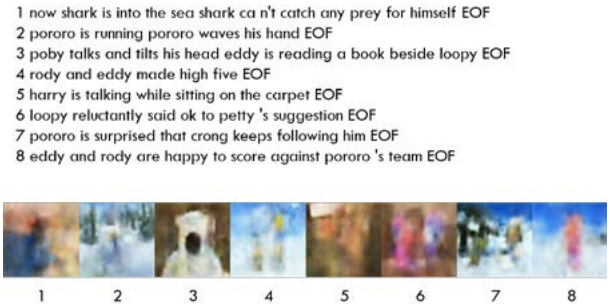

Fig. 2. **Some samples of output images.**

IV. Conclusion

The model generalized dual representation and connected both modality to generate images from sentences. Yet output image is not clear enough and the connection is ambiguous. We need to design models that can perform more accurate and logical inference between texts and images.

References

[1] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1431–1439.

[2] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," *arXiv preprint arXiv:1603.05631*, 2016.

[3] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

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

[7] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 365–372.

[8] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions," in *IEEE Computer Vision and Pattern Recognition*, 2016.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift." in *ICML*, ser. JMLR Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456.

[11] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–43.