딥인코더-디코더 기반의 인공지능 포토 스토리텔러

민경복*, Dang L. Minh*, 이수진**, 문현준*
*세종대학교 컴퓨터 공학과, **중앙대학교 다빈치 SW 교육원
e-mail :hmoon@sejong.ac.kr

AI photo storyteller based on deep encoder-decoder architecture

Kyungbok Min*, L. Minh Dang*, Sujin Lee**, Hyeonjoon Moon*
*Department of Computer Science and Engineering, Sejong University
**Department of Da Vinchi SW Education, Chungang University

요 약

Research using artificial intelligence to generate captions for an image has been studied extensively. However, these systems are unable to create creative stories that include more than one sentence based on image content. A story is a better way that humans use to foster social cooperation and develop social norms. This paper proposes a framework that can generate a relatively short story to describe based on the context of an image. The main contributions of this paper are (1) An unsupervised framework which uses recurrent neural network structure and encoder-decoder model to construct a short story for an image. (2) A huge English novel dataset, including horror and romantic themes that are manually collected and validated. By investigating the short stories, the proposed model proves that it can generate more creative contents compared to existing intelligent systems which can produce only one concise sentence. Therefore, the framework demonstrated in this work will trigger the research of a more robust AI story writer and encourages the application of the proposed model in helping story writer find a new idea.

1. INTRODUCTION

In recent years, beyond understanding and carrying out specific tasks, Artificial Intelligence (AI) has been increasingly applied in the fields which are considered human territory, the idea of using AI to generate artworks such as photo [1, 2] novels, and movie scenarios has become a trending topic in the research community.

The story is one of the earliest and fundamental means for humankind to transfer knowledge to the following generations. A story is highly linked to the visual domain because the narrative and visual context are the primary source of inspiration for stories. The task of creating a short story based on the context of the input image, which this research mentions as "visual story writer", is a challenging topic to that combine computer vision techniques and natural language processing (NLP) technologies. It demands not only the recognition of the context in a photo which involves interaction and relation among objects, but also a complicated system to generate a short story from the prospect of language consistency. Figure. 1 describes a typical image caption, which briefly summarized the content of an image and a short which relies on the caption of the image that was created by our proposed model.

The advancements in computer vision and NLP have allowed the computers to "perceive" and generate accurate descriptions for a given image. The results proved that researchers had made significant improvements in image captioning field [3], that seeks to output a caption for an input image, or video sequence description, which aims to create descriptions for a sequence of frames.

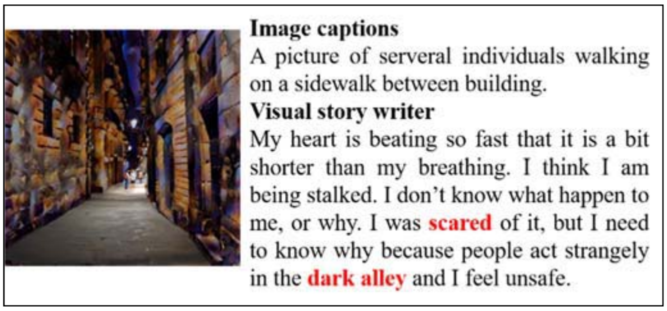


Image captions
A picture of serveral individuals walking on a sidewalk between building.
Visual story writer
My heart is beating so fast that it is a bit shorter than my breathing. I think I am being stalked. I don't know what happen to me, or why. I was **scared** of it, but I need to know why because people act strangely in the **dark alley** and I feel unsafe.

Figure 1. A short caption generated to describe the image content along with a story generated by the proposed model.

The next step is to make a narrative short story about the context of a captioned image. This indicates that the focus is shifted from generating a caption to creating a short story. Existing researches mainly concentrated on creating a simple caption for a single image [4]. We take one step further to try to generate a short story based on the image context. We regard the research topic as an unsupervised sequence-to-sequence problem, with a photo as the input, whereas the output is a sentence sequence. It is an entirely different matter from common sequence-to-sequence issues. Image captioning research is the nearest related research to "visual story writer". In those studies, recurrent neural network (RNN) is frequently implemented to generate a sentence as the output to describe the input image [5]. Compare to the common sequence-to-sequence problem, and the "visual story writer" is more complicated due to the challenge of modelling the huge visual variance and maintaining the long-term consistency among several sentences.

In this paper, we introduce an unsupervised sequence-to-sequence gated recurrent unit (GRU) framework that generates a short story of different genres using the caption content. The model creates stories, sentence by sentence concerning the image context and the previously generated sentence. The proposed framework consists of an RNN decoder that is trained on collected datasets. After that, sentences extracted from the datasets are mapped to a skip-thought vector. Then, the RNN model is conditioned on the skip-thought vector to create the encoded sentence. Moreover, we use visual-semantic embedding model to map conceptual captions dataset images and captions into a common vector space. Our goal with this approach was to force the system to generate stories that contain more narrative and consistent language and that every generated sentence in the story will be affected by previously generated sentences.

2. RELATED WORK

The system proposed by [6] aligned crowdsourced plot synopses with shots from the videos for story-style content retrieval. The authors applied a similarity function between sentences in plot synopses and shots based on keywords and personal identities in subtitles. In [7], the authors implemented a multiple neural network structure based on single and pairwise element-based predictions and exploited both text and image features. The obtained results proved that the proposed model could learn exciting aspects of common temporal sense. In another research proposed by [8], the authors exploited hashtags from an input image and analyzed it to create meaningful anecdotes connecting to the essence of the image. They used the attention-based encoder-decoder framework to create hashtags for an image. After that, a character-level language model was trained using a multi-layer RNN, and this model is applied to generate stories using one of the generated hashtags. Another model demonstrated by Alexander in [9] can produce image captions which were visually descriptive and appropriately. The authors tried to divide semantics and style. NLP techniques and frame semantics were employed to generate concise semantic term representation. Moreover, they implemented a unified language model that decoded sentences with diverse word choices and syntax for different styles.

Our work differs from previous work in several important aspects. We deal with a more challenging area of story/caption alignment. Unlike caption generation, which only describes what the main content of an image is, the story is more verbose and might vary because it is only based on the image context.

3. METHODOLOGY

3.1. Dataset preparation for story telling

This study contains two different datasets: (1) Books download from the Smashwords website, which is a website where authors share their unpublished books to the community. We crawl only books which contained over 20,000 words to filter out perhaps noisier shorter stories. The dataset has books in 2 different genres, romance (500 books) and horror (621 books). (2) Conceptual captions dataset, which is a huge captions dataset introduced by Google in 2018 [14]. It contains over 3.3 million pairs of image and caption, which were created by automatically filtering and extracting caption annotations from the internet. The dataset includes an increased magnitude of captioned images compared to the human-curated MS-COCO dataset. Moreover, the dataset contains a broader diversity of image-caption pairs as the images and annotations were downloaded from billions of web pages, enabling better performance of image captioning models and thus the models will generate better captions.

In order to download the books, we use an automated python crawler. The downloaded books were in PDF format, so we apply a python pdf2txt library to convert them into plain text format. Finally, we used the python NLP library to groom the data by removing blank rows and finally save all contents into one file.

3.2. Proposed AI caption generation model

In this section, a detailed description of the proposed "visual story writer" model is explained thoroughly. The implementation process of the proposed model is described in Figure 2, which includes three main components 1) Train both visual-semantic alignment model and skip-thought encoder on the conceptual captions dataset 2) Extract skip-thought vectors from the collected stories dataset and 3) Perform deep style transfer from caption style to story style using the skip-thought decoder.

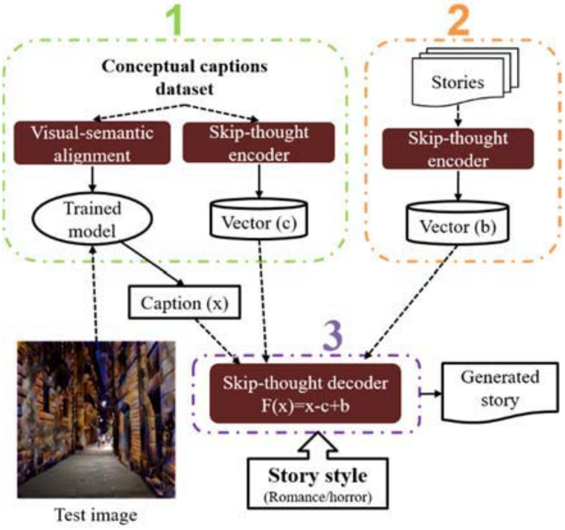


Figure 2. Overall system architecture of photo storyteller

A. Skip-thought vectors

- *Encoder*

Given a sentence tuple $(\boldsymbol{s_{i-1}}, \boldsymbol{s_i}, \boldsymbol{s_{i+1}})$, the model tries to encode the sentence $\boldsymbol{s_i}$ into a fixed vector. After that, based on the generated vector, it attempts to create the previous sentence $\boldsymbol{s_{i-1}}$ and the next sentence $\boldsymbol{s_{i+1}}$. Let $\boldsymbol{w_i^t}$ indicates the $\boldsymbol{t}$-th word for the sentence $\boldsymbol{s_i}$ and let $\boldsymbol{x_i^t}$ be its word embedding. We divide the model into two main sections involving the encoder and decoder.

Suppose that $w_i^1, \dots, w_i^N$ depicts all N words w in the current sentence s_i. The encoder task is to create a hidden state h_i^t for each time step t that encode the sequence of words from w_i^1 to w_i^t. Therefore, the last hidden state h_i^N of the sentence s_i will encode the entire contents. GRU model of the encoder generates the current hidden state h_i^t by linearly adding the previous hidden state h_i^{t-1} and the newly proposed state $\tilde{h}_i^t$

$$h_i^t = z_i^t \odot h_i^{t-1} + (1 - z_i^t) \odot \tilde{h}_i^t$$

where ⊙ is element-wise multiplication and z_i^t is the update gate. The update gate z_i^t is applied to decides how much information from the past need to be kept for the future. It is calculated by a linear sum between the newly proposed state and the existing state.

$$z_i^t = \sigma(W_z x_i^t + U_z h_i^{t-1} + b_i^z)$$

σ is the logistic nonlinearity, and W_z (n×m matrix), U_z (n×n matrix), b_z (n×1 vector) are fixed sized parameters belong to the update gate (two weights and bias). Thanks to the sigmoid activation function, the update gate output value is in the range of zero to one. If it is one, the update gate will completely forget previous hidden states, which means $h_i^t = \tilde{h}_i^t$. On the other hand, if it is zero, then all hidden states from previous time steps will be copied over, that is $h_i^t = h_i^{t-1}$. The proposed state $\tilde{h}_i^t$ is calculated similarly to the traditional recurrent unit.

$$\tilde{h}_i^t = \tanh(W x_i^t + U(r_i^t \odot h_i^{t-1}) + b_i^t)$$

in which r_i^t is a reset gate. The reset gate r_i^t is calculated similar to the update gate but based on a set of different parameter values: W_r (n×m matrix), U_r (n×n matrix), b_r (n×1 vector).

$$r_i^t = \sigma(W_r x_i^t + U_r h_i^{t-1} + b_i^r)$$

When the status is off (r_i^t equals to 0), the reset gate forces the newly proposed state to behave as it is working on the first word in a sequence of words, enabling it to eliminate the state h_i^{t-1} computed previously.

Although GRU preserves the memory remarkably better than RNN, it just analyzes the forward lingual context. Therefore, GRU is unable to examine the backward context. Thus, the learning process of GRU is considered partially completed because, in a sentence, the context of a word is influenced by its forward context and its backward context. TGRU, which is an improved version of GRU, was introduced to cope with the mentioned problem; it allows the model to encode the lingual context of a word from both sides. It contains two main networks, one network process the forward context, whereas, the other analyzes the backward context. The equations for the hidden state, the proposed hidden state, update gate, and reset gate of the forward and backward GRU are described below:

Forward pass:

$$\overrightarrow{z^t} = \sigma(\overrightarrow{W_z} x^t + \overrightarrow{U_z}\,\overrightarrow{h^{t-1}})$$

$$\overrightarrow{r^t} = \sigma(\overrightarrow{W_r} x^t + \overrightarrow{U_r}\,\overrightarrow{h^t})$$

$$\overrightarrow{\tilde{h}_t} = \tanh(\overrightarrow{W^d} x^t + \overrightarrow{U^d}(\overrightarrow{r^t} \odot \overrightarrow{h^{t-1}}))$$

$$\overrightarrow{h^t} = \overrightarrow{z^t} \odot \overrightarrow{h^{t-1}} + (1 - \overrightarrow{z^t}) \odot \overrightarrow{\tilde{h}^t}$$

The backward pass is the step we added to our model.

$$\overleftarrow{z_t} = \sigma(\overleftarrow{W_z} x^t + \overleftarrow{U_z}\,\overleftarrow{h^{t-1}})$$

$$\overleftarrow{r^t} = \sigma(\overleftarrow{W_r} x^t + \overleftarrow{U_r}\,\overleftarrow{h^{t-1}})$$

$$\overleftarrow{\tilde{h}_t} = \tanh(\overleftarrow{W^d} x^t + \overleftarrow{U^d}(\overleftarrow{r^t} \odot \overleftarrow{h^{t-1}}))$$

$$\overleftarrow{h_i^t} = \overleftarrow{z^t} \odot \overleftarrow{h^{t-1}} + (1 - \overleftarrow{z^t}) \odot \overleftarrow{\tilde{h}^t}$$

- *Decoder*

The decoder is calculated similar to the encoder, except that the calculation is based on the encoder output h_i from sentence s_i and the introduction of C_z, C_r, C that is used as a bias for the hidden state, update gate, and reset gate computation by the encoder output vector h_i.

Two different decoders are created, one is applied to create the previous sentence s_{i-1} and one is used to generate the next sentence s_{i+1}. The two decoders use different parameters to calculate their hidden states. However, they are based on one vocabulary dictionary V that uses a hidden state to calculate a distribution over words. As a result, two decoders are similar to an RNN language model but reply on the encoder output h_i.

Assume that $\overrightarrow{h^t}$ indicate the hidden state of the decoder of the next sentence s_{i+1} at time step t. The update gate, reset gate, and hidden state of the s_{i+1} decoder is given as follows

$$z^t = \sigma(W_z x^{t-1} + U_z h^{t-1} + C_z h_i)$$

$$r^t = \sigma(W_r x^{t-1} + U_r h^{t-1} + C_z h_i)$$

$$\tilde{h}_t = \tanh(W^d x^{t-1} + U^d(r^t \odot h^{t-1}) + C h_i)$$

$$h^t = z^t \odot h^{t-1} + (1 - z^t) \odot \tilde{h}^t$$

Where x^{t-1} is the word embedding of the previous word as the decoder try to guess the current word based on the context of the previous word.

3.3. Visual-semantic-embedding processing

This topic [10] is also one of the key area because storyteller will use caption not directly from the image. Input image apply visual-semantic embedding between COCO images and captions. In this model, captions and images are mapped into a common vector space. After training, new images can be embedded and retrieve captions. Encoder is used by a deep convolution network (CNN) and Long short-term memory recurrent network (LSTM) for learning a joint image-sentence embedding. Decoder is used by a neural language model that combines structure and content vectors for generating words one at a time in sequence.

3.4. Conditional neural language models

Given these models, a method to link the gap between retrieved image captions and passages in novels is required. That is, if we had a function F that maps a collection of image caption vectors X to a book passage vector $F(\mathrm{x})$, then we could feed $F(\mathrm{x})$ to the decoder to get our story. There is no such parallel data, so we need to construct F another way.

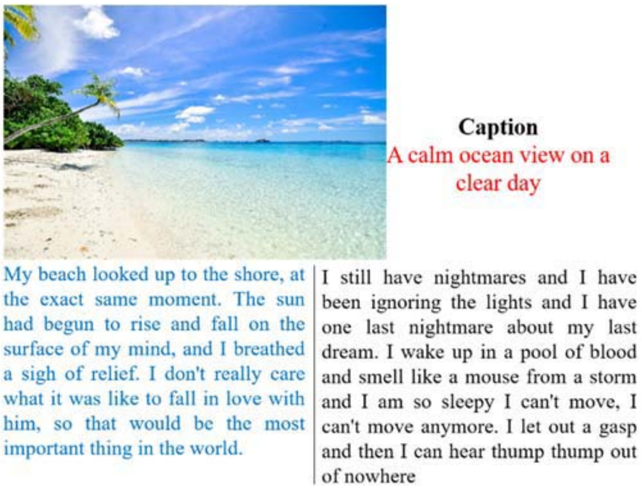
Suppose that we have 3 vectors: a caption x, a "caption style" vector c and a "story style" vector b. Then we define F as:

$$F(x) = x - c + b$$

which intuitively means: keep the "thought" of the caption, but replace the image caption style with that of a story. Then, we feed $F(x)$ to the decoder.

4. EXPERIMENT RESULTS

Figure 3 shows two examples of the generated stories, which include both romance stories and horror stories. The results prove that our model can create vivid stories for images. In particular, firstly, the model used deep visual-semantic alignment to create a short caption for an image. After that, this caption is used as a query to transfer the horror or romance story style to the caption using the proposed sentence encoder-decoder model. The retrieved stories are semantically understandable, and they also based on the general content of the image.



Caption
A calm ocean view on a clear day

My beach looked up to the shore, at the exact same moment. The sun had begun to rise and fall on the surface of my mind, and I breathed a sigh of relief. I don't really care what it was like to fall in love with him, so that would be the most important thing in the world.

I still have nightmares and I have been ignoring the lights and I have one last nightmare about my last dream. I wake up in a pool of blood and smell like a mouse from a storm and I am so sleepy I can't move, I can't move anymore. I let out a gasp and then I can hear thump thump out of nowhere

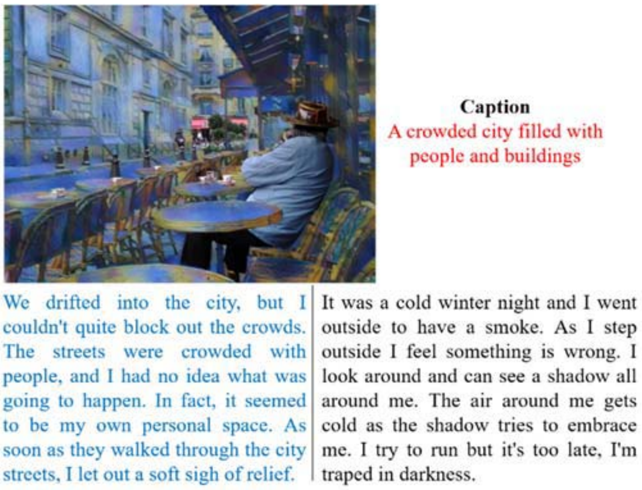
Figure 3. Generated results, including a general caption (red), romance story (blue), and horror story (black).

A slightly more exciting idea is shown in Figure 4; we apply our model on the images generated by computer from the research proposed by [11]. The created captions prove that the deep visual-semantic alignment works well in describing the general contents from the images even if the images are generated by the computer.

5. CONCLUSION

This study introduced a storytelling system from the contents of a photo. Firstly, we collected a large dataset of novels for each romance and horror genre. After that, a model, which applied NLP and encoder-decoder structure to generate the story were proposed. Finally, we generated stories which were related to the picture in different ways.

There are several ways to extend this study work in the future. In our study, we are not using grammar, detail context module in the architecture. Those two functions can make a more natural story and longer story to keep with the narrative. Another weakness of the system is deep learning model. We used only basic CNN model with LSTM method, but it is originally developed for image processing. In order to make a more advanced story creator, we need to find a suitable deep learning model with NLP.



Caption
A crowded city filled with people and buildings

We drifted into the city, but I couldn't quite block out the crowds. The streets were crowded with people, and I had no idea what was going to happen. In fact, it seemed to be my own personal space. As soon as they walked through the city streets, I let out a soft sigh of relief.

It was a cold winter night and I went outside to have a smoke. As I step outside I feel something is wrong. I look around and can see a shadow all around me. The air around me gets cold as the shadow tries to embrace me. I try to run but it's too late, I'm traped in darkness.

Figure 4. Generated results from the proposed model for a photo generated by computer.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2019-0-00136, Development of AI-Convergence Technologies for Smart City Industry Productivity Innovation), and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Agri-Bio Industry Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (316033-04-2-338 SB030).

REFERENCES

[1] Dang, L. Minh, Syed Ibrahim Hassan, Suhyeon Im, and Hyeonjoon Moon. "Face image manipulation detection based on a convolutional neural network." Expert Systems with Applications 129 (2019): 156-168.

[2] Dang, L., Syed Hassan, Suhyeon Im, Jaecheol Lee, Sujin Lee, and Hyeonjoon Moon. "Deep learning based computer generated face identification using convolutional neural network." Applied Sciences 8, no. 12 (2018): 2610.

[3] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[4] Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2422-2431).

[5] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).

[6] Tapaswi, M., Bäuml, M., & Stiefelhagen, R. (2015). Aligning plot synopses to videos for story-based retrieval. International Journal of Multimedia Information Retrieval, 4(1), 3-16.

[7] Agrawal, H., Chandrasekaran, A., Batra, D., Parikh, D., & Bansal, M. (2016). Sort story: Sorting jumbled images and captions into stories. arXiv preprint arXiv:1606.07493.

[8] Gaur, S. (2019, April). Generation of a Short Narrative Caption for an Image Using the Suggested Hashtag. In 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW) (pp. 331-337). IEEE.

[9] Mathews, A., Xie, L., & He, X. (2018). Semstyle: Learning to generate stylised image captions using unaligned text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8591-8600).

[10] Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models." arXiv preprint arXiv:1411.2539 (2014).

[11] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414-2423. 2016.