

# Vision Mediated Story Generation

Hong Chen Rapael Shu Noriki Nishida Hideki Nakayama

東京大学 情報理工研究科

{chen, raphael, nishida, }@nlab.ci.i.u-tokyo.ac.jp

nakayama@ci.i.u-tokyo.ac.jp

## 1 Introduction

Story generation is the task of automatically generating a complete story. For this task, the dataset is especially important. Looking back to our childhood, after we could understand each word's meaning, it still took a long time for us to achieve the ability to write a good story. According to this observation, a large dataset is undoubtedly necessary. However, unfortunately, the story data is hard to be obtained. So in this work, we propose a unsupervised method to train the story generation model without any story data.

On the other hand, different from the novel, the storybooks for children contain many pictures and we call them the picture-book. It means that the image information would help children to understand the context of the story. The case of picture-books tells us that visual information might play a significant role in a story. For another example, when people start writing a story, they will imagine the pictures of the next event in their minds and then write down what happened in the image. So in this view, we state that the image information is indispensable in the open story generation task.

To summary our contribution:

- It is the first time to solve the open story generation problem in the visual space.
- We advise an unsupervised approach to generate a story.

## 2 Related work

There exists a large number of researches utilising deep learning. The simplest way is to use a sequence to sequence model.

Furthermore, [6] used event as mediation to generate story. They first extracted the event from the sentence, then predicted the next event and transferred the event sentence into the story sentence as the next generated sentence. [3] utilised different attention mechanics learning word, sentence and paragraph hierarchically. They also created their own dataset Writing Prompt Dataset to train their model.

## 3 Method

To overview our model, its input is a topic sentence and it contains three part. The first part is retrieving visual representation with a given topic. The next one is predicting a sequence of the visual representations. The last one is recovering these visual representations into the sentences. Thus, we can get a sequence of narratives. Fig. 1 shows the overview of our model.

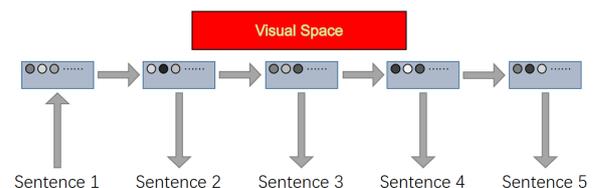


図 1: The overview of our entire method.

### 3.1 Retrieve image

The objective of this stage is to retrieve a desirable image. [2] embedded the text space into the visual space. They passed three types of sentence representations into the network and got the representation in the visual space. Three representations are Bag of Word (BoW) [4], Word2Vec [7] and Gated Recurrent Unit (GRU) [1].

When retrieving the image, they calculated the cosine similarity with each image in the database and found the image with the maximum value as retrieved image. In this work, we borrow their model to retrieve the image. Fig. 2 shows the network. The Flickr30k dataset is used in this training step.

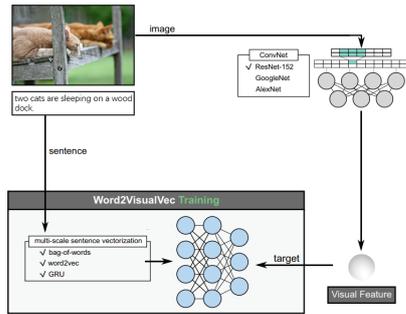


图 2: This network embeds the text space into the visual space.

### 3.2 Predict next image

The predicted next image should contain close semantic meaning to the current one. This regularisation ensures the coherence of our generated story. We experiment three traditional methods in this stage: K nearest neighbour search, Multilayer Perceptron with MSE loss and Multilayer Perceptron with Pairwise Ranking loss.

Method	Top 1	Top 5	Top 10
NNS	9.33%	18.96%	22.47%
MLP(MSE)	0.70%	3.30%	5.84%
MLP(Pairwise)	0.50%	1.30%	3.10%

表 1: The performance of three methods in predicting the next image test on VIST test dataset.

From the Table 2, K nearest neighbour search performs the best. Therefore, we use K nearest neighbour search in this stage. In the test phase, we use the entire images in the VIST dataset as the next image candidates.

### 3.3 Generate sentence

The purpose of this stage is to generate a story-style sentence. To achieve this goal, we separate it into two stages. The first is generating captions with conventional approaches and the second is transferring the captions into the story-style sentences.

To generate the caption for the image, we borrow the NIC model [8]. We also produce a model to transfer the caption to the sentence. This model is the same as the NIC model but with different input. The NIC model's input is the representation of the image and this model's input are the concatenation of the caption and the image representation. The caption representation is the last output of GRU whose input is the word embeddings of the caption. The image representation is the subtraction of the current image representation and the previous image representation. Fig. 3 shows the caption to story-style sentence model. The dataset that we used for generating caption is the MSCOCO dataset. For transferring the caption to sentence, we use the corresponded the caption and story sentence pairs in the VIST dataset.

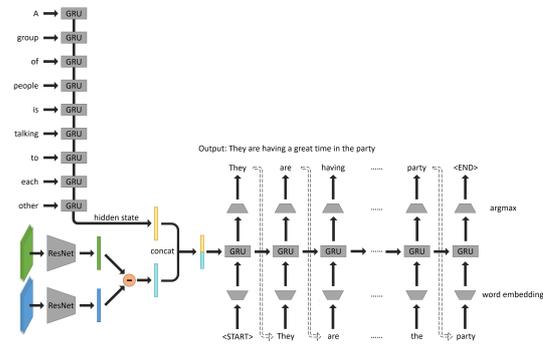


图 3: The caption to story-style sentence model. The input combine the caption feature and the image feature. The GRU model produces the caption feature. The image feature is the subtraction of the feature of the current image and the previous image.

## 4 Experiment

### 4.1 Topic Coherence Score

[5] produced a network to evaluate the score of the coherence. Inspired by their idea, we propose a model to

produce a topic coherence score for our task. Fig. 4 shows the Topic Coherence model.

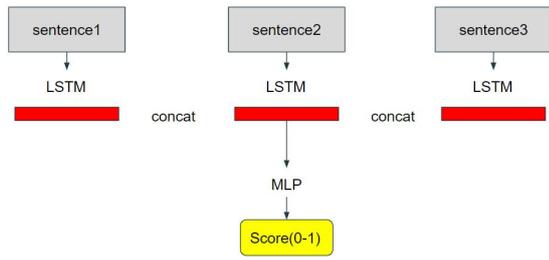


图 4: Topic Coherence model

For training this model, we use the story in VIST dataset as training data. The stories in this dataset has high topic coherence, so we give one as their label. On the other side, we counterfeit two kinds of stories with no topic coherence. 1) We randomly choose five sentences from the sentence corpus, compose them into a story. As for the objective function, We use the margin ranking loss. 2) We randomly duplicate one sentence in the story. Both of the faked data are labeled zero for training.

$$loss = margin + score(fake) - score(real) \quad (1)$$

We test the trained topic coherence model with 1000 story in test data and 1000 fake data. These fake data use the first sentence in a real story as the start sentence and extend four randomly chosen sentences from sentence corpus. We get 86.45% in the test data.

$$Acc = \frac{1}{n} \sum_{i=0}^n \mathbb{I}(score(real_i) > score(fake_i)) \quad (2)$$

This result shows this topic coherence model has ability to distinguish the bad story and the good story.

## 4.2 Consistency and Diversity Score

The second evaluation method that we produce is a counting-based method. We assume that a good story should both satisfy the topic consistency and the word diversity. Based on this assumption, we introduce the function to evaluate the story. Equation 3 shows the function

of this evaluation.

$$Score(D) = \frac{1}{|D|} \sum_{d \in D} \frac{NumberOfUniques(d)}{TopicInconsistency(d)}$$

$$NumberOfUniques(d) = \text{len}(\text{set}(\text{flatten}(d)))$$

$$TopicInconsistency(d) = \mathbb{E}_{w_i, w_j \in d} [\text{distance}(w_i, w_j)] \quad (3)$$

NumberOfUniques shows how many different words are in the story and TopicInconsistency is the mean value of the distance of all possible word pairs in the story. In short, we collect all the words appeared in the generated story. The number of the word collections is the NumberOfUniques while TopicInconsistency calculates the distances for all the possible word pairs in the collections. The higher score means the better performance.

## 4.3 Result

We use the Topic Coherence Score and the Consistency and Diversity Score to evaluate the model. We randomly generate 100 stories using the first sentence of the story in VIST test data as input sentence in both our method and the baseline method. From the results, we can see that we get higher score than the baseline in Table 2. Fig. 5 present some examples using our method and the baseline model.

Method	Ours	Baseline	GroudTruth
Model based	0.4371	0.3855	0.7234
Counting based	1.882	0.945	2.99
Human	1.852	1.480	2.108

表 2: The Model based score is the Topic Coherence Score and the Counting based score stands for the Consistency and Diversity Score. We also propose the human preference which is a three-grade marking system.

## 5 Future Work

As for the future work, we present two possible ways of improving the quality of the story generation. First, as we said, it is possible to add visual reasoning in the image

<b>Input:</b> the catering was well done .
<b>Ground Truth:</b> they were many different dishes for all of our tastes . we were able to try many types of flavors . the day went really well at the office party . good music and food brought every one closer together .
<b>Baseline:</b> everyone had a great time at the ceremony . there were a lot of people waiting for the ceremony to start . they were all very proud of being there . we had a great time .
<b>Ours:</b> we all had a great time eating and drinking and celebrating with everyone . we all had a great time at the party and eating some delicious food . we all had a great time and we learned a lot about it and the food . we had a great time at the party .
<b>Predicted image:</b>


図 5: Exmaples compared with our method and the baseline method and human annotated story.

prediction part. With the KNN method we used in this paper, the reasoning part is not concerned, which makes the event flow becomes not clear. For the second one, The dataset we used is not good enough for generating a story with good quality. Our goal is to generate a story with any topic. However, in the training dataset, if it does not contain the all kinds of story, for testing, we also can not generate the story with such style. Thus, we need a better dataset.

## 6 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16H05872. The members in AIST also produced useful advices for this work.

## 参考文献

- [1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [2] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 2018.
- [3] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.
- [4] Zellig S Harris. Distributional structure. *Word*, Vol. 10, No. 2-3, pp. 146–162, 1954.

- [5] Alice Lai and Joel Tetreault. Discourse coherence in the wild: A dataset, evaluation and methods. *arXiv preprint arXiv:1805.04993*, 2018.
- [6] Lara J Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O Riedl. Event representations for automated story generation with deep neural nets. *arXiv preprint arXiv:1706.01331*, 2017.
- [7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.
- [8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.