Commonsense Knowledge Aware Concept Selection For Diverse and Informative Visual Storytelling

Hong Chen[1,3], Yifei Huang[1], Hiroya Takamura[2,3], Hideki Nakayama[1,3]
The University of Tokyo[1], Tokyo Institute of Technology[2]
National Institute of Advanced Industrial Science and Technology, Japan[3]
{chen, nakayama}@nlab.ci.i.u-tokyo.ac.jp
hyf@iis.u-tokyo.ac.jp, takamura.hiroya@aist.go.jp

Abstract

Visual storytelling is a task of generating relevant and interesting stories for given image sequences. We propose to foster the diversity and informativeness of a generated story by using a concept selection module that suggests a set of concept candidates. Then, we utilize a large scale pre-trained model to convert concepts and images into full stories. To enrich the candidate concepts, a commonsense knowledge graph is created for each image sequence from which the concept candidates are proposed. To obtain appropriate concepts from the graph, we propose two novel modules that consider the correlation among candidate concepts and the image-concept correlation. Extensive automatic and human evaluation results demonstrate that our model can produce reasonable concepts.

1 Introduction

Most previous works on Visual Storytelling (VST) constructed end-to-end frameworks [19, 18, 10, 21]. However, their stories tend to be monotonous which contains limited lexical diversity and knowledge [6] (see the example in Figure 1). Recently, two-stage generation methods, also known as plan-write strategy, aroused much research attention in story generation tasks [20, 14, 1]. When adopted to the task of VST, Hsu et.al. [6] shows that this strategy is capable of generating more diverse stories compared with end-to-end methods.

In this work we aim to generate stories that are both diverse and informative for a given input image sequence. Taking the advantage of the previous two-stage models, we detect image concepts and construct concept graphs for proposing a set of concept candidates, and propose two novel methods for better selecting the appropriate concept for the second generation stage. After detecting the concept in each input image, we first extend the concepts into a larger commonsense graph using ConceptNet [13]. This extension step increases the informativeness of generated stories. Since selecting appropriate candidates from the concept graph is critical for generating stories of good quality, a natural way is to use a graph attention network [17] to refine the node features.

[state-of-the-art] *The basketball game was intense. The opposing team was very competitive. The game was intense. The fans cheered. It was a great game .*	**[state-of-the-art]** *It was a great day for the baseball game. It was a lot of fun. The crowd was cheering on. The game was intense. It was a great day to see the game .*
[Ours] *The team was on the court on the last night of the basketball game. The player on the team was on a shot. The ball kept getting blocked during the game. Team fans on the sidelines were cheering the team. The other team on the other shot was also blocked.*	**[Ours]** *This was the baseball game. The players were all excited to be playing against each other in the game. We all watched as the team played together. The players got in on the action and played hard. This was my favorite part of the game. The game was just too great. Our team won the game !*

Figure 1 Stories generated by the existing work [10] and our proposed model using concept selection (red). The existing work tends to generate similar stories (blue) for different inputs. Our model can generate more informative and diverse stories.

For selecting the most adequate concept from the candidates as the input to the second stage of our model, two novel modules are proposed in this work. The first one, named Sequential Selection Module (SSM), operates in a straightforward manner that uses an encoder-decoder for selecting concepts for each image. Differently from SSM, the second module called Maximal Clique Selection Module (MCSM) processes the concept graph as a whole. It learns a probability for each concept in the training phase, and during inference it finds a maximal clique using the Bron Kerbosch algorithm [2]. The concepts within the clique are used for the next story generation step. Our experiments show that improved quality of concept selection can greatly help to increase the diversity of the generated stories while keeping the relevance with the input images.

The second stage of our model generates a story with the image features and the selected concepts. Other than using the same module for fair comparison with existing works, we also propose to modify the large scale pre-trained model BART [11] to input the images and concepts and output the full stories.

We conduct extensive experiments on the public VIST dataset [9]. Our experiments demonstrate that using our

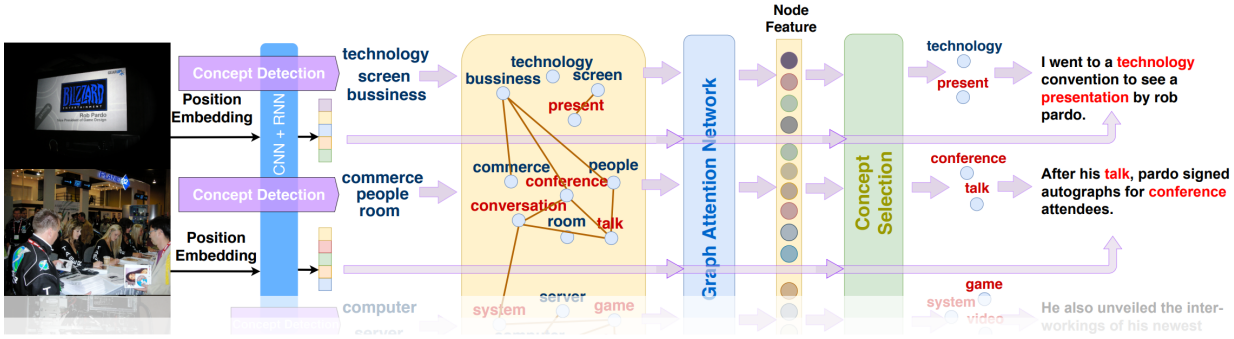

Figure 2 An overview of our visual storytelling model. The image features are obtained by a pretrained CNN combined with a bi-LSTM layer. The concepts are obtained from a concept detection model and enriched by ConceptNet [13]. These concepts from the nodes in a graph and are connected according to the relationship in the knowledge base. Initialized by the word embedding vector, the concept features are then updated by a Graph Attention Network. Our proposed concept selection module is then applied to select concept words using the image and concept features. Finally, both image features and concept words are used to generate a full story.

proposed concept selection modules, our generated stories can achieve better performance on both automatic metric and multiple human evaluation metrics using the same generation module. When equipped with BART, the quality of the stories can be remarkably improved, with the generated story diversity similar to human writing.

In summary, our main contributions are listed as follows:

- We propose two novel modules SSM and MCSM to select concepts from the given candidates concepts under a plan-write two-stage visual storytelling system.
- We exploit modified BART as our story generation module to mitigate the problem caused by limited vocabulary and knowledge in the dataset.
- Large scale experiments using automatic metrics and human evaluation show that our model can outperform previous models by a large margin in both diversity and informativeness, while retaining the relevance and logicality as the previous work.

2 Related Work

Wang et.al. [18] proposed a visual storytelling framework which is widely used as a base model in the coming-up studies. This framework uses an end-to-end structure that first converts the image into features and then transfers its information to the adjacent images by a BiLSTM layer. Finally, a decoder decodes the features separately and merges the sentences into a story. While many succeeding works [8, 10] can achieve high automatic scores, the story may not be interesting and informative [7] for humans as they often contain repetitive texts and limited information. On the other line of the research, to alleviate the low diversity problem, Hsu et.al. [6] proposed to generate several concepts before outputting the full stories. The discrete concept words can guide the decoder to produce more diverse stories. This plan-and-write strategy [20] can substantially increase the diversity of the stories.

3 Method

Figure 2 depicts an overview of our proposed model. Given a sequence of N image features $I = \{I_1, ..., I_N\}$ as input, our model 1) construct a large commonsense graph for the images, 2) update concept feature in the graph, 3) select the concepts from the graph and 4), send concepts and image features into the decoder to output the full story. The details of each step are as follows.

3.1 Commonsense Graph Construction

We use clarifai [16] to obtain the top 10 seed concepts from each image. Each concept is used as a query to select relative commonsense concepts in the ConceptNet [13]. An undirected edge is established between concepts if they are related in ConceptNet. Also, a concept in one image will connect to the related concepts in the adjacent images to allow information flow between images.

3.2 Concept Features Update

Initialized with word embedding vectors, the concept features are updated by a two-layer Graph Attention Network, which passes information between connected concepts and image using attention mechanism.

3.3 Concept Selection Module

We propose two methods to select concepts given the concept features and the image features. To better formalize the procedure in the methods, we denote $c^{i,j}$ as the j-th concept of the i-th ($1 \le i \le N$) image. we let $\mathcal{C}_S = \{c_S^{1,1}, ..., c_S^{N,K}\}$ and $\mathcal{C}_G = \{c_G^{1,1}, ...\}$ denote the concepts set in the source candidate concepts and the full word set in the gold story, respectively. The target concepts are their intersection: $\mathcal{C}_T = \mathcal{C}_S \cap \mathcal{C}_G$.

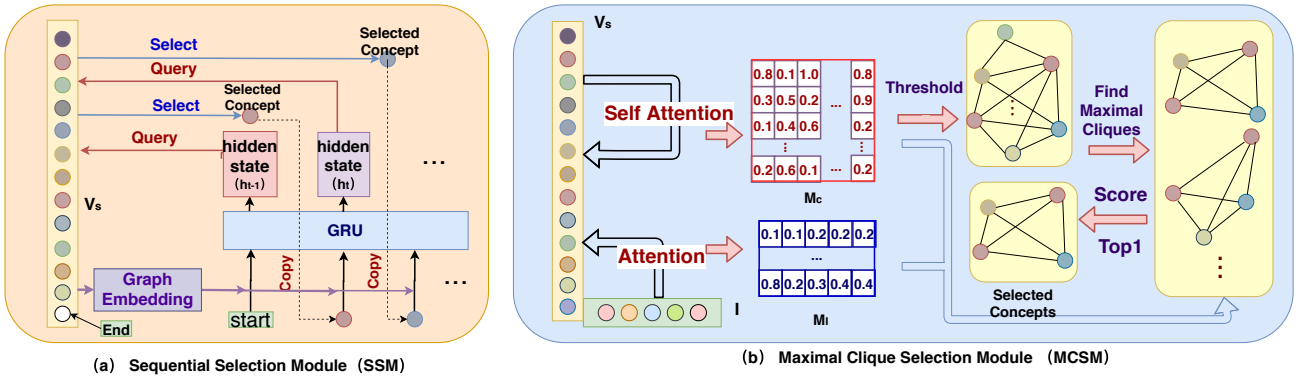

Figure 3 Concept selection modules: (a) Sequential Selection Module (b) Maximal Clique Selection Module

3.3.1 Sequential Selection Module (SSM)

One straightforward way of selecting concepts is to adopt an encoder-decoder model where we can forward the updated concept features into the encoder, and the decoder will output the selected concepts. Inspired by the Copy Mechanism [4], instead of generating a probability distribution with vocabulary size in each step, the SSM outputs are directly chosen from the inputs $\mathcal{C}_S$. As shown in Figure 3(a), we use a GRU [3] to first encode the concept embedding feature v_S^{t-1} and the hidden state into a new hidden state h^t. We then use h^t to query all the concepts in $\mathcal{C}_S$ to get a probability p_S for each concept in the source set. Finally the concept with the highest probability is selected as the output concept, while its feature is directly copied for the generation of the next step:

$$\begin{aligned} h^t &= \mathbf{GRU}\left(h^{t-1}, v_S^{t-1}\right), \\ p_S &= \mathbf{softmax}\left((W_h h^t)^{\mathrm{T}} W_c V_S\right), \\ c_S^t &= \mathbf{argmax}(p_S), \end{aligned} \tag{1}$$

Here W_h and W_c are trainable projection matrices. The objective function is to maximize the probability score of the concepts which locate in C_T.

$$\mathcal{L}_{ssm} = -\Sigma y_{S,T} \log(p_S), \tag{2}$$

where $y_{S,T}$ is an indicator of whether a concept in $\mathcal{C}_S$ is in $\mathcal{C}_T$. The sequence selection step stops when the module generates <end> token. This <end> token is added to the set of candidate concepts with a uniform random initialized feature without any update during the training phase. The same procedure is done to the <start> token except that it is not involved in the candidates.

3.3.2 Maximal Clique Selection Module (MCSM)

Different from SSM, this method aims to calculate the co-occurrence probability of all candidate concepts c_s in the graph. An illustration of MCSM is shown in Figure 3(b). In the beginning, we calculate self-attention to compute a correlation matrix $M_C \in (NK \times NK)$ which contains the correlation between each pair of nodes. We also calculate another correlation matrix for each image $M_I \in (N \times K)$ indicating the correlation between the concept embedding feature (v_S) and image features (I).

$$\begin{aligned} M_C &= \sigma(v_S^{\mathrm{T}} W_a^{\mathrm{T}} W_b v_S). \\ M_I &= \sigma(I^{\mathrm{T}} W_c^{\mathrm{T}} W_d v_S). \end{aligned} \tag{3}$$

Here, W_a, W_b, W_c, W_d is trainable weights, σ denotes sigmoid activation function. Intuitively, the concepts that appear in a gold story should own high correlations with each other, and the image should be highly related to the gold concepts to describe it. Thus, our target correlation maps can be written as follow:

$$\begin{aligned} \hat{M}_C[i,j] &= \begin{cases} 1, & c_i \in \mathcal{C}_T \wedge c_j \in \mathcal{C}_T. \\ 0, & \text{otherwise.} \end{cases} \\ \hat{M}_I[i,j] &= \begin{cases} 1, & c_j \in \mathcal{C}_T^i. \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

Then, the objective is to minimize the difference between predicted and target correlation maps:

$$\mathcal{L}_{mcsm} = \lambda_1 \left\| M_C - \hat{M}_C \right\|_2^2 + \lambda_2 \left\| M_I - \hat{M}_I \right\|_2^2 \tag{5}$$

In the testing phase, M_C can be viewed as a fully connected graph in which the edge weights are the values in the matrix. Therefore, a low edge weight means less co-occurrence probability between two concepts. Based on this assumption, we set a threshold τ to remove the edges whose weight is less than τ. Then we apply Bron Kerbosch algorithm [2] to find all maximal cliques from the remaining sub-graph. Finally, we score each of them with equation 6 and select a clique with maximum score s. The output concepts are the nodes within the selected cliques.

$$\begin{aligned} s &= s_C + s_I. \\ s_C &= \frac{1}{(\|\mathcal{C}_P\| - 1)\,\|\mathcal{C}_P\|} \sum_i \sum_{j \neq i} \log(M_C[i,j]). \\ s_I &= \frac{1}{\|\mathcal{C}_P\|} \sum_{i=1}^{N} \sum_{c_j \in \mathcal{C}_P^i} log(M_I[i,j]). \end{aligned} \tag{6}$$

Choices(%)	MCSM vs INet		MCSM vs KS		MCSM vs SSM		MCSM+BART† vs KS		MCSM+BART† vs MCSM	
	MCSM	INet	MCSM	KS	MCSM	SSM	MCSM+BART	KS	MCSM+BART	MCSM
Revelence	**47.4**	35.6	26.3	**31.6**	**50.5**	40.0	28.8	**33.6**	35.2	35.2
Informativeness	**51.0***	31.6	**46.3***	28.9	**44.7**	41.2	**62.5****	18.8	**58.8****	23.5
Logicality	**35.5**	34.3	**34.2**	29.0	32.9	**42.3**	**35.3**	33.3	**40.2**	37.5
Overall	**55.0****	30.0	**44.7**	34.2	**48.3**	37.1	**43.5****	23.0	**47.0***	31.6

Table 1 Human evaluation. Numbers indicate the percentage of annotators who believe that a model outperforms its opponent. Methods without (+BART) means using RNN as the story generation module. Cohen's Kappa coefficients (κ) for all evaluations are in Moderate or Fair agreement, which ensures inter-annotator agreement. We also conduct a sign test to check the significance of the differences. The scores marked with * denotes $p < 0.05$ and ** indicates $p < 0.01$ in sign test.

Method	Dist-2	Dist-3	Dist-4
INet☆	8.36	18.92	31.02
KS☆	10.84	22.90	36.51
KG-Story†	18.73	38.65	57.22
Ours (MCSM)	13.98	34.01	54.11
Image+BART†	21.63	46.23	67.57
Ours (MCSM)+BART†	**34.95**	**69.88**	**88.74**
Gold	47.76	82.27	95.05

Table 2 Diversity of generated stories by different methods. † denotes the story generation module is pre-trained with other dataset. ☆ denotes the model is end-to-end trained.

where $\mathcal{C}_P$ denotes the concepts in a clique, and $\mathcal{C}_P^i$ denotes the concept of the i-th image in the clique.

3.4 Concept to Story Module

The selected concepts are assigned to its corresponding image to generate the sentences. We tried two kinds of encoder-decoder to decode the story: 1) **RNN**: a simple RNN based encoder-decoder module that uses multi-head pooling to encode the concept embedding, and decode the sentences with a RNN decoder. 2) **BART**: a large scale pre-trained encoder-decoder which both can encode the input and output the sentences.

4 Experiment

We conduct experiments on the widely used VIST dataset [9]. For fair comparison, we follow the same experiment setting as [10] except that we set the vocabulary size to 28,000. All models use the same fixed random seed. We use the following baselines:

INet [10] uses a "hide-and-tell" strategy to train an end-to-end model. In this method no concept is used.

KS [19] uses sigmoid attention to incorporate concept features into the model. We change the structure of the visual encoder and decoder the same as **INet** for fair comparison.

KG-Story† [6] is a strong baseline that use two stage plan-write strategy and pre-train the decoder on RocStories Corpora [15]. †indicates the model uses a pre-trained model.

Image+BART†is an end-to-end baseline that uses BART on top of image features to directly generate the story. This baseline is one-stage that does not generate concepts.

We also change the concept selection module and story generation module in our model to validate the effectiveness of each component. Specifically, we compare: Rand+RNN, C_Attn+RNN, SSM+RNN, MCSM+RNN, and MCSM+BART† .

Human Evaluation To better evaluate the quality of generated stories, we conduct human evaluation to compare pairwise outputs with several models via the Amazon Mechanical Turk (AMT). We sample 200 image sequences from the test set and generate stories using each model. For each sample pair, two annotators participate in the judgement and decide their preference on either story (or tie) in terms of **Relevance**, **Informativeness**, **Logicality** and **Overall**. Table 1 shows the human evaluation result. From the comparison between MCSM and INet and the comparison between MCSM and KS, we can see that our two-stage planning method greatly outperforms the end-to-end models, especially in the informativeness score. The MCSM module also outperforms the SSM module, which indicates positive correlation between the quality of concept selection and the overall quality of generated stories. Finally, using BART with MCSM can help to achieve further informativeness and generate even better stories.

Comparison on diversity We report Distinct-n scores [12] in Table 2 that calculate the percentage of unique n-gram in all generated stories in the test data. Higher score means less inter-story repetition. From the table, two stage methods (KG-Story and ours) achieve significantly higher diversity scores. Our MCSM can generate the most diverse stories among all the methods without using external pre-trained models. When equipped with BART, we can achieve diversity close to human writing.

5 Conclusion

In this work we exploit concept selection for improving the diversity and informativeness of stories generated from image sequences. By constructing a commonsense graph and two novel modules for concept selection, our proposed model outperforms all previous works in diversity by a large margin while still preserving the relevance and logical consistency on the VIST dataset.

Acknowledgements This paper is based on the conference paper presented at AAAI2021 [5] and was based on the results from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and also supported by JSPS KAKENHI Grant Number JP19H04166.

References

[1]P. Ammanabrolu, E. Tien, W. Cheung, Z. Luo, W. Ma, L. J. Martin, and M. O. Riedl. Story realization: Expanding plot events into sentences. In *AAAI*, 2020.

[2]C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. 1973.

[3]K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv*, 2014.

[4]J. Gu, Z. Lu, H. Li, and V. O. Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv*, 2016.

[5]H. T. H. N. Hong Chen, Yifei Huang. Commonsense knowledge aware concept selection fordiverse and informative visual storytelling. *AAAI*, 2021.

[6]C.-C. Hsu, Z.-Y. Chen, C.-Y. Hsu, C.-C. Li, T.-Y. Lin, T.-H. Huang, and L.-W. Ku. Knowledge-enriched visual storytelling. *arXiv*, 2019.

[7]J. Hu, Y. Cheng, Z. Gan, J. Liu, J. Gao, and G. Neubig. What makes a good story? designing composite rewards for visual storytelling. In *AAAI*, pages 7969–7976, 2020.

[8]Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *AAAI*, 2019.

[9]T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *ACL2016*, 2016.

[10]Y. Jung, D. Kim, S. Woo, K. Kim, S. Kim, and I. S. Kweon. Hide-and-tell: Learning to bridge photo streams for visual storytelling. *arXiv*, 2020.

[11]M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv*, 2019.

[12]J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv*, 2015.

[13]H. Liu and P. Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 2004.

[14]L. J. Martin, P. Ammanabrolu, X. Wang, W. Hancock, S. Singh, B. Harrison, and M. O. Riedl. Event representations for automated story generation with deep neural nets. *arXiv*, 2017.

[15]N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, and J. Allen. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, 2017.

[16]G. Sood. *clarifai: R Client for the Clarifai API*, 2015.

[17]P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv*, 2017.

[18]X. Wang, W. Chen, Y.-F. Wang, and W. Y. Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv*, 2018.

[19]P. Yang, F. Luo, P. Chen, L. Li, Z. Yin, X. He, and X. Sun. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *IJCAI*, 2019.

[20]L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan. Plan-and-write: Towards better automatic storytelling. In *AAAI*, 2019.

[21]L. Yu, M. Bansal, and T. L. Berg. Hierarchically-attentive rnn for album summarization and storytelling. *arXiv*, 2017.

6 Appendix

6.1 Experiment on Concept Selection

Similar as Keyphrase generation tasks, we apply precision, recall and f measure to evaluate the efficiency of concept selection methods. We compare among several methods:

Rand: A simple baseline where we randomly pick 3 concepts from the candidates for each image. On average, each image contains 2.65 gold concepts.

C_Attn: We extract the attended concepts where the attention score is larger than a threshold from the model of Yang et.al. [19]. This is an end-to-end model with sigmoid attention on concept words. We choose 0.8 as the threshold since this contributes the best f-score.

Image to concept(I2C): This is a straightforward version of concept selection where the concepts are directly generated from the images. We simply add a projection layer on each hidden state to predict the concept words from the vocabulary size of the concepts, which is very similar to the model of Hsu et.al. [6].

SSM: Our proposal which uses a copy mechanism in each step of selection.

MCSM: Our proposal which calculates the correlation score for concept-concept and image-concept and uses maximal clique selection.

Qualitative results are shown in Table 3. We can see that our proposed SSM and MCSM can achieve significantly higher f-score than other methods. This helps our model to keep the story relevance to the input images while generating diverse stories.

Method	Precision	Recall	F measure
Rand	2.68	2.45	2.56
C_Attn	30.38	**43.37**	35.86
I2C	31.32	20.75	24.96
SSM	40.43	40.30	40.36
MCSM	**45.30**	40.90	**42.99**

Table 3 Concept selection performance of different methods. The results show that our MCSM achieved the best f-score among all methods.

6.2 Case Study

We show a qualitative result of a random test sample in Figure 4. This is a hard example because the last three images are very similar and the objects in all images are hard to recognize. We can see that INet generates monotonous and even irrelevant sentences. KS can generate better sentences but still low in lexical diversity. For the stories generated by two-stage strategy with RNN (SSM+RNN, MCSM+RNN), we can see that the story follows the selected concepts and the stories seem more reasonable than that of end-to-end training methods. When using BART, we compare three methods that represent no concept selection (Image+BART), bad concept selection (Rand+BART) and ours concept selection (MCSM+BART). We can see that without using concepts or using randomly selected concepts, the generated stories are of low quality and to a certain extent irrelevant to the images. However, when guided by the selected concept, the story becomes vivid, relevant and logical.

INet	This is a picture of a bird. There was a lot of birds on the ground. There was a bird in the water. The birds were in the water. It was a great day for a swim.
KS	It was a nice day for a walk on the water. There were many birds in the sky. There were some pretty white ones. the water was very calm. There were also a lot of birds in the water.
SSM +RNN	Concepts: (1) day (2) ∅ (3) bird (4) water time lot (5) bird frozen back This is a picture of a day. The water was covered in snow. This is a picture of a bird. The bird spent was a lot of time in the water. The bird is frozen and he could not go back.
MCSM +RNN	Concepts: (1) day winter cold (2) spot (3) bird (4) water fish (5) lake It was a cold day for winter. We spotted some of the trees we saw. I saw a bird.The bird was fishing in some water. We had a great time at the lake.
Image +BART	We went to the beach. There was a lot of interesting things to see. There was a bird in the sky. We had a great time. It was very cold at night.
Rand +BART	Concepts: (1) home debating (2) dock alarm (3) bucket cap (4) oyster photo (5) slush americana There was a debate at home before heading out to the dock. Our alarm went off and we kept watch over the dock. Then they noticed the bird with its cap in the air and watched the bird fly away from the safety of the bucket.He jumped up to take a photo of this magnificent oyster. We ended our day by the water feeding with american birds.
MCSM +BART	Concepts: (1) day winter cold (2) spot (3) bird (4) water fish (5) lake A cold winter day at the lake. [male] spotted an animal on the lake and thought that it was a strange bird he could spot. The bird seemed peaceful , and seemed to be enjoying the cold. It then tried to fish in the water. [male] kept a close eye on the bird, but had to keep his distance from the lake.
Gold	Deer looking for food. And disappeared after hearing noises. Bird waiting patiently. Perching in the cold. About to take a flight.

Figure 4 The examples of generated stories by different methods. Our MCSM and SSM can generate better stories compared with other baselines that do not use BART. When using the pretrained BART, the concept selection with MCSM can produce vivid and informative story.