Hide-and-Tell: Learning to Bridge Photo Streams for Visual Storytelling

Yunjae Jung,[1] **Dahun Kim,**[1] **Sanghyun Woo,**[1] **Kyungsu Kim,**[2] **Sungjin Kim,**[2] **In So Kweon**[1]
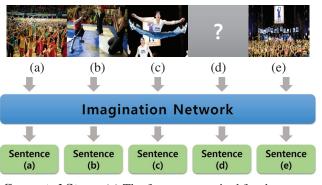
[1]Korea Advanced Institute of Science and Technology (KAIST), Korea
[2]Samsung Electronics Co., Ltd (Samsung Research), Korea

Abstract

Visual storytelling is a task of creating a short story based on photo streams. Unlike existing visual captioning, storytelling aims to contain not only factual descriptions, but also human-like narration and semantics. However, the VIST dataset consists only of a small, fixed number of photos per story. Therefore, the main challenge of visual storytelling is to fill in the visual gap between photos with narrative and imaginative story. In this paper, we propose to explicitly learn to imagine a storyline that bridges the visual gap. During training, one or more photos is randomly omitted from the input stack, and we train the network to produce a full plausible story even with missing photo(s). Furthermore, we propose for visual storytelling a hide-and-tell model, which is designed to learn non-local relations across the photo streams and to refine and improve conventional RNN-based models. In experiments, we show that our scheme of *hide-and-tell*, and the network design are indeed effective at storytelling, and that our model outperforms previous state-of-the-art methods in automatic metrics. Finally, we qualitatively show the learned ability to interpolate storyline over visual gaps.

Introduction

Recent deep learning based approaches have shown promising results for vision-to-language problems (Vinyals et al. 2015; Karpathy and Fei-Fei 2015; Donahue et al. 2015; Yu et al. 2016; Pan et al. 2016; Gao et al. 2017) that require the generation of text descriptions from given images or videos. Most existing methods have focused on giving direct and factual descriptions of visual content. While this is a promising first step, it is still challenging for artificial intelligence to connect vision with more naturalistic and human-like language. One emerging task proposed to take one step closer to human-level description is visual storytelling (Huang et al. 2016). Given a stream (set) of photos, this method aims to create a narrative, evaluative and imaginative story based on semantic visual understanding. While conventional visual descriptions are visually grounded, visual storytelling tries to describe contextual flow and overall situation across the photo stream, and so its output sentences can contain words for objects that do not even appear in the

Generated Story: (a) The fans were excited for the game. (b) There were many people there. (c) The lead singer performed a great performance. **(d) The game was very intense.** (e) This was a great game of the game.

Figure 1: Example of our hide-and-tell prediction. In this example, our hide-and-tell network takes four valid images and one black image. It is designed to learn contextual relations across the photo stream. Despite the hidden photo, our predicted sentence (d) *"The game was very intense"* is semantically natural and plausible with the whole story context.

given image. Therefore, filling in the visual gap between the given photos with a subjective and imaginative story is the main challenge of visual storytelling.

In this paper, we propose to explicitly learn to imagine the storyline that bridges the visual gap. To this end, we present an auxiliary *hide-and-tell* training task to learn such ability. As shown in Fig. 1, one or more photos in the input stack are randomly masked during training. We train our model to produce a full, plausible story even with a missing photo(s). This image dropout in training encourages our model to describe what is happening *in* the given stream of photos, as well as *between* the photos. Since this story imagination task is an ill-posed problem, we follow curriculum learning, in which we start with an original setting in the early steps, and gradually increase the number of image dropout during training.

Furthermore, we propose an imagination network that is

designed to learn non-local relations across photo streams to refine and improve, in a coarse-to-fine manner, the recurrent neural network (RNN) based baseline. We build upon a strong baseline model (XE-SS) (Wang et al. 2018c) that has a CNN-RNN architecture and is trained with cross-entropy loss. Since we focus on learning contextual relations among all given photo slots, even those with missing photos, we propose to add a non-local (NL) layer (Wang et al. 2018b) after the RNN block to refine long-range correlations across the photo streams. Our imagination network is designed with the first CNN block, and a stack of two RNN-NL blocks with a residual connection between; the following gated recurrent unit (GRU) outputs the final storyline.

In the experimental section, we evaluate our results with automatic metrics of BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). We conduct a quantitative ablation study to verify the contribution of each of the proposed design components. Also, we compare our imagination network with existing state-of-the-art models for visual storytelling. By conducting a user study, we show that our results are qualitatively better than the baselines. Another user study demonstrates that our hide-and-tell network is able to predict a plausible overall storyline even with missing photos. Finally, we introduce a new task of story interpolation, which involves predicting language descriptions not only for the given images, but also for gaps between the images.

Our contributions are summarized as follows.

- We propose a novel hide-and-tell training scheme that is effective for learning imaginative ability for the task of visual storytelling

- We also propose an imagination network design that improves over the conventional RNN-based baseline.

- Our proposed model achieves state-of-the-art visual storytelling performances in terms of automatic metrics.

- We qualitatively show that our network faithfully completes the storyline even with corrupted input photo stream, and is able to predict inter-photo stories.

Related Work

Visual Storytelling
Visual storytelling is a problem of generating human-like descriptions with images selected from a photo album. Unlike conventional captioning tasks, visual storytelling aims to create a subjective and imaginative story with semantic understanding in the scenes. Early work (Park and Kim 2015) exploits user annotation from blog posts. Newly released VIST (Huang et al. 2016) dataset with a narrative story leads to several follow-up studies. Approaches with hierarchical concept (Yu, Bansal, and Berg 2017; Wang et al. 2019) are proposed. And Wang et al.; Wang et al. (2018a; 2018c) formulate a visual storytelling task using adversarial reinforcement learning methods.

Overcoming Bias

Overfitting is a long-stand problem of deep neural network which causes difficulty in test cases. To alleviate this problem, dropout (Srivastava et al. 2014) is widely adopted. During training, it randomly drops weights in the neural networks to avoid severe co-adapting. For language models, a similar approach (Ji et al. 2015) named blackout is proposed to increase stability and efficiency. While dropout is often used at hidden layers of networks, blackout only targets output layers. Recently, for captioning models, Burns et al. (2018) tries to overcome bias in gender-specific words by occluding gender evidence in training images.

Illustrated hiding methods motivate our input blind learning scheme. It randomly obscures one or two images from the input in the training stage. Since the VIST dataset has a fixed number of input images as five, there can be overfitting in learning relations among images. From this point of view, our hide-and-tell concept gains performance improvement from the perspective of regularization. Also, visual storytelling aims to generate subjective and imaginative descriptions unlike conventional captioning. In that regard, our approach has the advantage that the network learns to imagine the skipped input.

Curriculum Learning
Inspired by the human learning process, Bengio et al. (2009) proposed curriculum learning which starts from relatively easy task and gradually increases the difficulty of training. It benefits both performance improvement and speed of convergence in various deep learning tasks such as optical flow (Ilg et al. 2017), visual question answering (Misra et al. 2018), and image captioning (Ren et al. 2017). We also exploit curriculum learning by scheduling the difficulty of a task. At the early steps of training, there is no obscured input. Then, one of the five input images is omitted in the later step. Lastly, two of the five input images are hidden. If a validation loss is saturated, each step goes into the next step.

Relational Embedding
Recently, a non-local neural network (Wang et al. 2018b) is proposed to capture long-range dependencies with self-attention. in other words, it computes the relations along with spatio-temporal spaces. Also, the non-local layer is a flexible network that can be well suited to both convolution layers and recurrent networks. It is widely used to vision tasks such as scene graph generation (Woo et al. 2018), image generation (Zhang et al. 2018), and NLP tasks such as image and video captioning (Gao et al. 2019), text classification and sequence labeling (Liu et al. 2019). We also exploit the self-attention mechanism of the non-local layer to our networks which try to imagine a story for the hidden images by learning relations between images.

Proposed Approach
An overview of the proposed imagination network is shown in Fig. 2. Given five input images $I = \{I_1, I_2, I_3, I_4, I_5\}$, the model outputs five corresponding sentences $S = \{s_1, s_2, s_3, s_4, s_5\}$. Each sentence consists of several words
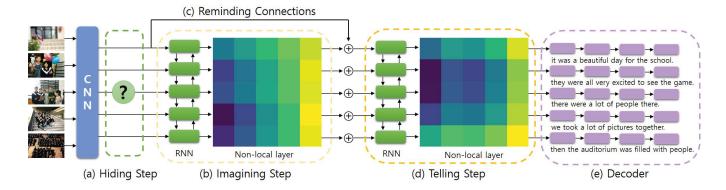
Figure 2: The overall architecture of our imagination network. (a) In hiding step, one (or two) of the five inputs are randomly omitted by zero-masking. (b) In the imagining step, the inter-frame relations are roughly captured by the proposed imagination network which is composed of RNN (*e.g.* GRU) and non-local self-attention module. (c) By utilizing residual connections, the imagination network can focus on recovering the blinded features. (d) The telling step refines the whole features using the same architecture with that of the imagining step, while the parameters are not shared between them. (e) The decoder generates a final story that describes the whole images.

$W = \{w_1, w_2, \cdots, w_T\}$, where T denotes the length of the sentence.

Our model operates in three steps: **Hide**, **Imagine**, and **Tell**. After the first convolutional layer, which extracts visual features from each input photo, the hiding step randomly blinds one or two image features. It is implemented by setting the selected feature values to 0. During training, we employ a curriculum learning scheme, which starts with a normal setting (without hiding) and gradually increases the number of hidden images to two image features. (*i.e.* 0 to 2). In our preliminary experiment, we found that blinding three or more image features does not provide further performance improvement.

Second, the imagining step consists of the aforementioned RNN-NL block. The goal of this step is to make a coarse initial prediction for the omitted features. Together with a residual connection from the CNN feature stack, this step captures contextual relations between the known image features, focusing on recovering the missing features. Finally, the telling step takes the feature stack from the previous imagining step, and refines the relational embedding to capture more concrete semantics throughout the photo stream. The RNN-NL block in this step shares the same architecture as that of imagining step, while the parameters are not shared. The refined feature stack is fed into the decoder to generate the final language output.

Hide-and-Tell Learning

Hiding Step
The input photo stream $I = \{I_1, I_2, I_3, I_4, I_5\}$ is fed into the pre-trained CNN layer, which extracts high-level image features $F = \{f_1, f_2, f_3, f_4, f_5\}$. As shown in Fig. 2-(a), one or two of the features F are randomly dropped in the hiding step. Although the missing information makes the reconstruction task an ill-posed problem, the method of hiding not only has a regularization effect but also helps our model to learn the contextual relations that lead to a performance

gain in testing.

$$F_{blind} = \{b_1 f_1, b_2 f_2, \cdots, b_5 f_5\}, \quad b_n \in \{0, 1\} \quad (1)$$

where n denotes the number of input images, F_{blind} is a feature set including zero-masked features, b_n is a masking weight which is randomly set during training.

Curriculum Learning
It is very challenging even for human intelligence to recover the missing features by using the neighboring photos in the same input stack. To ease the training difficulty in early steps, we adopt a curriculum learning scheme (Bengio et al. 2009). In early training, our imagination network is given fully visible photo stack (*i.e.* $b_{total} = 0$). When the training loss becomes saturated, we start to hide one image feature from the input stack (*i.e.* $b_{total} = 1$). Similarly, we proceed to hide two image features (*i.e.* $b_{total} = 2$) in the later steps.

$$b_{total} = \begin{cases} 0, & \text{if epoch} < \alpha \\ 1, & \text{if } \alpha \leq \text{epoch} < \beta \\ 2, & \text{otherwise} \end{cases} \quad (2)$$

where α, β are hyper parameters which are empirically determined as the saturation point. The effect of curriculum learning is shown in the experiment section (Table. 2).

Imagination Network
Our imagination network (INet) is designed to learn contextual relations between images in the input stack, and to generate human-like stories even with omitted photo(s). Following a coarse-to-fine pipeline, our network includes a *coarse* imagining step and a *fine* telling step that correspond to Fig. 2-(b) and (d), respectively. We use RNN-NL block in both steps.
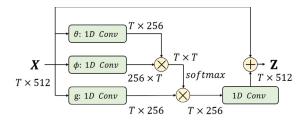
Figure 3: Relational embedding layer. The features are first reshaped as a matrix form. Three parallel 1-D convolutions are used for feature embedding. The non-local operation starts with computing the correlation map ($T \times T$). It is produced by multiplying the output of θ and ϕ with the following $softmax$ normalization. The map is then multiplied with the output of g. The residual connection from X to Z allows the non-local block to be incorporated into existing RNN layers.

Imagining Step

In the imagining step, F_{blind} is fed into the bidirectional gated recurrent unit (GRU). In the forward direction, Bi-GRU takes $F_{blind} = \{b_1 f_1, b_2 f_2, \cdots, b_5 f_5\}$ and embeds according hidden states $(\overrightarrow{h_1}, \cdots, \overrightarrow{h_5})$. Then, in the backward direction, reversed hidden states $(\overleftarrow{h_1}, \cdots, \overleftarrow{h_5})$ are generated. The hidden states are concatenated into $h_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right]$.

To model non-local relations between the images, we employ an embedded Gaussian version of a non-local neural network (Vaswani et al. 2017; Wang et al. 2018b). As illustrated in Fig 3, our relational embedding is different from most existing non-local approaches in that it considers each input image feature as one element and focuses on the relations between the photo streams. Detailed equations for relational embedding are as follows:

$$y = softmax(x^T W_\theta^T W_\phi x) \, g(x), \quad (3)$$
$$Z = W_z \, y + x, \quad (4)$$

where, x is the hidden states from GRU (*i.e.* h_i), and each W denotes 1D convolution layers because our approach does not consider the spatial dimension of each input image, but considers each image feature as one element.

Inspired from residual shortcuts (He et al. 2016), a reminding connection is added to connect initial CNN features to the end of the first RNN-NL block. By adding F_{blind} to Z, which is the output of the relational embedding layer, the first RNN-NL block is encouraged to focus on recovering the missing features.

$$F_{reminded} = F_{blind} + Z. \quad (5)$$

Telling Step

In the telling step in Fig. 2-(d), the features from the previous imagining step, $F_{reminded}$, are fed into the second RNN-NL block, which shares the same architecture as the first block, but does not share the weight parameters. The features that have been hidden during the hiding step are now roughly reconstructed in the feature stack $F_{reminded}$, and the second RNN-NL block refines these features to allow more concrete and associative understanding of all the photos in the input stream. Thus, to make better language predictions, the second block focuses more on refining the features of all photo elements.

The decoder (Fig. 2-(e)) consists of GRU and generates sentences $S = \{s_1, s_2, s_3, s_4, s_5\}$ for each input photos. In order to generate each sentence S, word $W = \{w_1, w_2, \cdots, w_T\}$ are recurrently predicted in one-hot vector v_t as follows:

$$w_t = \text{GRU}(w_{t-1}, [f_{tell}; v_{t-1}]), \quad (6)$$
$$v_t = softmax(W_w \, w_t), \quad (7)$$

where W_w denotes fully connected (FC) layer and non-linearity (*e.g.* hyperbolic tangent function).

Experiments

Experimental Setup

Datasets
Our experiments are conducted on the VIST dataset which provides 210,819 unique photos from 10,117 Flickr albums for visual storytelling tasks. Given five input images selected from an album, corresponding five sentences annotated by users are provided as ground truth. For the fair comparison, we follow the conventional experimental settings used in existing methods (Yu, Bansal, and Berg 2017; Wang et al. 2018c). Specifically, three broken images are excluded in our experiments. Also, the same number of training, validation, and test sets are used: 4,098, 4,988, and 5,050.

Evaluation Metrics
In order to quantitatively measure our method for storytelling, automatic metrics such as BLEU, ROUGE-L, CIDEr, and METEOR are adopted. We employ the same evaluation code used in existing methods (Yu, Bansal, and Berg 2017; Wang et al. 2018c). Two sets of human-subjective studies are performed for further comparison.

Implementation Details
We reproduced XE-ss (Wang et al. 2018c) and set as our baseline network. However, our approach is completely different from their adversarial reinforcement learning method except for the baseline (*i.e.* XE-ss). ResNet-152 (He et al. 2016) is used for the pre-trained CNN layer in Fig. 2. We empirically choose hyper parameters for curriculum learning; $\alpha = 50, \beta = 80$. The learning rate starts with $4e-4$, and it decays by half when the training difficulty is changed (*i.e.* epoch $= \alpha$ or β). Adam optimizer is used. For non-linearity in the network, ReLU (Nair and Hinton 2010) is used for pre-trained CNN layers and SELU (Klambauer et al. 2017) is employed for the imagining step and the telling step. In decoding stage, beam search is utilized with beam size $= 3$. For fair experiments, we removed randomness along with

Method	B-1	B-2	B-3	B-4	M	R	C
INet - B	63.7	39.1	23.0	13.9	35.1	29.2	9.9
INet - N	64.4	39.8	23.6	14.3	35.4	29.6	9.4
INet - R	63.5	39.0	22.9	13.9	35.0	29.4	9.2
INet	**64.4**	**40.1**	**23.9**	**14.7**	**35.6**	**29.7**	**10.0**

Table 1: Ablation Study. We block important components of INet to empirically verify its contributions to the final performance. INet-B, INet-N, and INet-R denote INet without blinding, non-local layers and the second RNN-NL block respectively.

Methods	B-1	B-2	B-3	B-4	M	R	C
0	63.7	39.1	23.0	13.9	35.1	29.2	9.9
1	64.2	39.8	23.7	14.6	35.5	29.7	9.9
2	62.7	39.1	23.5	14.4	35.5	29.5	9.2
(0, 1)	63.7	39.6	23.5	14.4	35.4	**29.9**	9.8
(0, 1, 2)	**64.4**	**40.1**	**23.9**	**14.7**	**35.6**	29.7	**10.0**

Table 2: Curriculum learning. To show the effect of the curriculum learning, we experiment by varying the number of image dropouts for each item. The left column denotes the number of hidden input features during training. The (0, 1, 2) means that the number of hiding increases from 0 to 2. And the 0 denotes the number of hiding is fixed to 0.

different experiments by fixing a random seed. In other words, our experimental results do not rely on multiple trials.

Quantitative Results

Ablation Study
We conduct an ablation study to demonstrate the effects of different components of our method in Table. 1.

Our model has three distinctive components; hiding step, non-local attention layer, and imagination network. We investigate the importance of each component. If we provide non-blinded (*i.e.* fully-visible) input features to the model, the model loses the regularization effects. We call this model as *INet-B*. If we omit the non-local attention layers, the network should only rely on the recurrent neural network (RNN) to capture the inter-frame relations, missing the complementary effects of the non-local relations. We named this model as *INet-N*. If we do not use the telling step, the model only has one imagining step which shows insufficient performance to generate more concrete sentences on the photo stream. We named this model as *INet-R*.

In all ablation setups, we observe performance drops. The model *INet-B* shows that simply using all the image features is not enough to get good results as it is prone to overfitting. This shows the effectiveness of the proposed *hide-and-tell* learning scheme. The model *INet-N* suffers from its structural limitation as it purely depends on the recurrent neural networks for modeling the inter-frame relationship, and has difficulty handling complex relations between the frames. The result of model *INet-R* implies that the refinement stage after the first imagination step is crucial.

Method	B-1	B-2	B-3	B-4	M	R	C
Huang et al.	-	-	-	-	31.4	-	-
Yu et al.	-	-	21.0	-	34.1	29.5	7.5
HPSR	61.9	37.9	21.5	12.2	34.4	**31.2**	8.0
GAN	62.8	38.8	23.0	14.0	35.0	29.5	9.0
XE-ss	62.3	38.2	22.5	13.7	34.8	29.7	8.7
AREL (best)	63.8	39.1	23.2	14.1	35.0	29.5	9.4
HSRL	-	-	-	12.3	35.2	30.8	**10.7**
INet	**64.4**	**40.1**	**23.9**	**14.7**	**35.6**	29.7	10.0

Table 3: Comparison to Existing Methods. Following automatic metrics are used: BLUE (B), METEOR (M), ROUGE-L (R), and CIDEr (C). The result shows that our approach achieves new state-of-the-art result.

XE-ss	Hide-and-tell	Tie
24.7 %	55.2 %	20.1 %

Table 4: Baseline vs INet without hiding in the test.

Comparison to Existing Methods
We compare our method with state-of-the-art methods (Huang et al. 2016; Yu, Bansal, and Berg 2017; Wang et al. 2018c; 2019; Huang et al. 2019) in Table. 3. Our approach achieves the best results in BLEU and METEOR metrics. Compared with previous approaches, our approach could better handle complex sentences. However, evaluation metrics are not perfect as there are many reasonable solutions for the narrative story generation. Therefore, we perform a user study and compare our approach with the strongest state-of-the-art baseline (Wang et al. 2018c). For each user study (Table. 4, Table. 5), thirty participants answered twenty five queries. As shown in Table. 4, we see that our approach significantly outperforms the baseline, implying that our method produces much more human-like narrations.

Qualitative Results
Non-hiding Test
We qualitatively compare our model with the baseline (Wang et al. 2018c) in Fig. 4. We can observe that our model produces more diverse and comprehensive expressions. For example, in Fig. 4-(a), the repeated sentences (*e.g.* "The flowers were so beautiful") are generated by the baseline, whereas, the results of ours show a wide variety of sentences (*e.g.* "Some of the flowers were very colorful.", "The flowers were blooming."). Moreover, there exists an apparent gap in depicting the picture. For describing the second photo in Fig. 4-(a), ours "There were many different kinds of shops there" is a better representation than baseline's "There were a lot of people there". We observe a similar phenomenon in the example (b) as well. While the baseline

Full input	Hidden input	Tie
30.9 %	40.5 %	28.6 %

Table 5: INet vs INet$_{hidden}$. In inference stage, we compare the story generated by INet with and without hidden images.

(a)					
Baseline	I went to the market last week.	There were a lot of people there.	There were lots of flowers to choose from.	The flowers were so beautiful.	The flowers were so beautiful.
INet	I went to the market yesterday.	There were many different kinds of shops there.	There were a lot of flowers there.	Some of the flowers were very colorful.	The flowers were blooming.
User Annotation	It is a beautiful morning to visit the original location.	Early morning is a great time to beat the crowds.	There are unusual plants to be seen.	The variety of flowers is simply amazing. There are short flowers.	There are tall flowers. Flowers you never knew existed.
(b)					
Baseline	We had a great time at the party.	There was a lot of food to eat.	There was a lot of food there.	We had a lot of food to eat.	I had a great time .
INet	We went out to dinner at a local restaurant.	The food was amazing.	It was a lot of ingredients.	They had a lot of meat.	I had a great time .
User Annotation	It was time to have a super dinner with family.	We started out with all these taco ingredients.	Then we moved to this squid like substance.	Then we had a bunch of shrimp scampi.	And finished it off with some strong drinks.

Figure 4: Non-hiding test. We qualitatively compare the results of baseline and the results of INet using all input images without hiding in the inference stage. (a) The upper example. (b) The lower example.

(a)			?		
Baseline	The students were ready for the graduation ceremony.	All of the students were excited to be graduating.	Diplomas and family members were there to support the.	The students were very proud of their accomplishments.	My family was so happy to be there.
INet	All the students were ready to go to the graduation.	They were very excited to graduate.	The graduation ceremony was a lot of fun.	After the ceremony , the students posed for a picture.	I took a picture with my family.
(b)					?
Baseline	The students were ready for the graduation ceremony.	The students were very excited to be graduating.	She was so happy to be graduating.	The students were very proud of their diplomas.	Diplomas all day.
INet	We were all excited to finally be graduating.	All of my friends were there to support me.	The dean of the graduation ceremony was a lot of fun.	He was so proud of his diploma.	The ceremony was over, we took a picture of the entire family.

Figure 5: Hiding test. For the obscured input, we qualitatively show the results of the baseline and the results of INet. A story for the hidden image (*i.e.* the third image in (a)) is also generated. Unlike Fig. 4, user annotation is skipped in this experiment because users already know which input image is blinded.

Figure 6: Story Interpolation. Given five input images provided in the VIST dataset, we insert black images in between the five images. Our model is asked to predict the sentence descriptions for both valid and black images. The generated sentences are plausible, and the storyline shows natural contextual flow.

repeats the same expressions such as *"There was a lot of food"*, our network generates a wide variety of descriptions such as *"food", "ingredients", "meat"*. The qualitative results above demonstrate again that our method greatly improves over the strongest baseline (Wang et al. 2018c).

Hiding Test

In this experiment, we explore the strength of INet by hiding the input images in testing. As shown in Fig. 5, one of the five input images is omitted. Specifically, the third and fifth image are masked in Fig. 5-(a) and Fig. 5-(b) respectively. We then show the story generated by ours and the baseline (Wang et al. 2018c). We can clearly see that our method produces a much more natural story and well captures the associative relations between the images. For example, the results of the baseline do not even form a sentence such as (*e.g. "Diplomas and family members were there to support the."* or *Diplomas all day.*). On the other hand, the results of ours not only well maintains the global coherency over the sentences and are more locally consistent with neighboring sentences (*e.g. "The graduation **ceremony** was a lot of fun."* and *"**After the ceremony**, the students posed for a picture."*).

In Table. 5, we show that INet with one hidden image can generate a more human-like story than the INet without any hidden images. Thanks to the proposed hide-and-tell learning scheme, our INet is equipped with a strong imagination ability regardless of the input image masking.

Story Interpolation

The story interpolation is a newly proposed task in this paper. It aims to interpolate the story by predicting sentences in between the given photo stream. Since the photo stream has temporally sparse images, the current task of visual storytelling has limited expressiveness. However, the proposed

story interpolation task can make the whole story more specific and concrete.

As illustrated in Fig. 6, a story for given five input images is generated. Additionally, the inter-story for inserted black images is also created with four sentences. The results of interpolation look thoroughly maintaining both global contexts over the whole situation and local smoothness with adjacent sentences. For instance, the generated sentence *"The Halloween party was over."* maintains both the global context of whole situation (*i.e. halloween party*) and local smoothness (*i.e. party was over*) preceded by *'[male] had a great time."*.

Motivated by the importance of imagination in the visual storytelling task, we extend our blinding test (Fig. 1) to the story interpolation task. While the blinding test recovers a story for the hidden input, story interpolation generates inter-story (*i.e.* five plus four, total nine sentences). Since creating a story by looking only at surrounding images without corresponding input obviously requires imagination, our hide-and-tell approach faithfully performs well due to our new learning scheme and network design.

Conclusion

In this paper, we propose the hide-and-tell learning scheme with imagination network for visual storytelling task which addresses subjective and imaginative descriptions. First, input hiding block omits an image from an input photo stream. Then, in imagining block, features of the hidden image are predicted by associating inter-photo relations with RNN and 1D convolution-based non-local layer. At the last, concrete relations between images are refined to generate sentences in the decoder. In experiments, our approach achieves state-of-the-art performance both in automatic metrics and human-subjective user studies. Finally, we propose a novel story interpolation task and show that our model well imagines the inter-story between given photo streams.

References

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. of Association for Computational Linguistics Workshop*, 65–72.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. 41–48. ACM.

Burns, K.; Hendricks, L. A.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. of Computer Vision and Pattern Recognition*, 2625–2634.

Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T. 2017. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia* 19(9):2045–2055.

Gao, L.; Li, X.; Song, J.; and Shen, H. T. 2019. Hierarchical lstms with adaptive attention for visual captioning. *IEEE Trans. Pattern Anal. Mach. Intell.*

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition*, 770–778.

Huang, T.-H. K.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual storytelling. In *Proc. of North American Chapter of the Association for Computational Linguistics*, 1233–1239.

Huang, Q.; Gan, Z.; Celikyilmaz, A.; Wu, D.; Wang, J.; and He, X. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proc. of Association for the Advancement of Artificial Intelligence*, volume 33, 8465–8472.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. of Computer Vision and Pattern Recognition*, 2462–2470.

Ji, S.; Vishwanathan, S.; Satish, N.; Anderson, M. J.; and Dubey, P. 2015. Blackout: Speeding up recurrent neural network language models with very large vocabularies. *arXiv preprint arXiv:1511.06909*.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. of Computer Vision and Pattern Recognition*, 3128–3137.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-normalizing neural networks. In *Advances in neural information processing systems*, 971–980.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Liu, P.; Chang, S.; Huang, X.; Tang, J.; and Cheung, J. C. K. 2019. Contextualized non-local neural networks for sequence learning. In *Proc. of Association for the Advancement of Artificial Intelligence*.

Misra, I.; Girshick, R.; Fergus, R.; Hebert, M.; Gupta, A.; and van der Maaten, L. 2018. Learning by asking questions. In *Proc. of Computer Vision and Pattern Recognition*, 11–20.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. 807–814.

Pan, P.; Xu, Z.; Yang, Y.; Wu, F.; and Zhuang, Y. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proc. of Computer Vision and Pattern Recognition*, 1029–1038.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of Association for Computational Linguistics*, 311–318.

Park, C. C., and Kim, G. 2015. Expressing an image stream with a sequence of natural sentences. In *Proc. of Neural Information Processing Systems*, 73–81.

Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; and Li, L.-J. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proc. of Computer Vision and Pattern Recognition*, 290–298.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proc. of Computer Vision and Pattern Recognition*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proc. of Computer Vision and Pattern Recognition*, 3156–3164.

Wang, J.; Fu, J.; Tang, J.; Li, Z.; and Mei, T. 2018a. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Proc. of Association for the Advancement of Artificial Intelligence*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *Proc. of Computer Vision and Pattern Recognition*, 7794–7803.

Wang, X.; Chen, W.; Wang, Y.-F.; and Wang, W. Y. 2018c. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*.

Wang, B.; Ma, L.; Zhang, W.; Jiang, W.; and Zhang, F. 2019. Hierarchical photo-scene encoder for album storytelling. In *Proc. of Association for the Advancement of Artificial Intelligence*.

Woo, S.; Kim, D.; Cho, D.; and Kweon, I. S. 2018. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems*, 560–570.

Yu, L.; Bansal, M.; and Berg, T. L. 2017. Hierarchically-attentive rnn for album summarization and storytelling. *arXiv preprint arXiv:1708.02977*.

Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; and Xu, W. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proc. of Computer Vision and Pattern Recognition*, 4584–4593.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.