Let Your Photos Talk: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks

Yu Liu[†], Jianlong Fu[‡], Tao Mei[‡] and Chang Wen Chen[†]
[†]State University of New York at Buffalo, NY, USA
[‡]Microsoft Research Asia, Beijing, P. R. China
[†]{yliu44,chencw}@buffalo.edu, [‡]{jianf,tmei}@microsoft.com *

Abstract

Automatic generation of natural language description for individual images (a.k.a. image captioning) has attracted extensive research attention. In this paper, we take one step further to investigate the generation of a paragraph to describe a photo stream for the purpose of storytelling. This task is even more challenging than individual image description due to the difficulty in modeling the large visual variance in an ordered photo collection and in preserving the long-term language coherence among multiple sentences. To deal with these challenges, we formulate the task as a sequence-to-sequence learning problem and propose a novel joint learning model by leveraging the semantic coherence in a photo stream. Specifically, to reduce visual variance, we learn a semantic space by jointly embedding each photo with its corresponding contextual sentence, so that the semantically related photos and their correlations are discovered. Then, to preserve language coherence in the paragraph, we learn a novel Bidirectional Attention-based Recurrent Neural Network (BARNN) model, which can attend on the discovered semantic relation to produce a sentence sequence and maintain its consistence with the photo stream. We integrate the two-step learning components into one single optimization formulation and train the network in an end-to-end manner. Experiments on three widely-used datasets (NYC/Disney/SIND) show that the proposed approach outperforms state-of-the-art methods with large margins for both retrieval and paragraph generation tasks. We also show the subjective preference of the machine-generated stories by the proposed approach over the baselines through a user study with 40 human subjects.

Introduction

Generating a human-level narrative from an ordered photo stream, in this research we refer to as "visual storytelling", presents a fundamental challenge to both computer vision and natural language processing areas. This is challenging because it requires not only the full understanding of each photo in a stream as well as the relation among different photos, but also a sophisticated mechanism to generate a natural paragraph from the perspective of language coherence.

Existing researches have focused more on generating natural descriptions for a single photo. In this research we take one step further to investigate the problem of paragraphing a photo stream. We consider this task as a sequence-to-sequence learning problem, where the input is a photo stream (with order) while the output is a sentence sequence, each corresponding to one photo.

Image Captions:
- A group people is play ball.
- A man is playing a ball.
- A group of people is having dinner around a table.

Storytelling:
A group of people gather as a team. The ball game begins. After the *ball game*, *the team winds down with some food*.

Figure 1: The difference between storytelling and image captions in isolation is the coherence among sentences, i.e. the green term "ball game" and the blue term "the team" are introduced from previous photos with corresponding color. [Best viewed in color]

It is a totally different issue from conventional sequence-to-sequence problems. The closely related research to visual storytelling is image and video captioning, where the Recurrent Neural Networks (RNNs) are usually employed for generating a single sentence from a given image or video clip (Venugopalan et al. 2015a)(Donahue et al. 2015), (Venugopalan et al. 2015b), (Pan et al. 2016). However, the task of visual storytelling is more challenging due to the difficulty of modeling the large visual variance in an ordered photo collection and preserving the long-term language coherence among multiple sentences. First, at the input side, photo stream usually has significantly large visual variance, such as the example in Figure 1. Most existing models for sequence learning with deep structure (e.g., RNN) are not designed to deal with such long-term dependency with large visual variance. Second, at the output side, visual storytelling requires much more complicated textual form in coherent and consistent paragraph. As the storytelling in the Figure 1, the overlapping of semantics between sentences often results in complex relational structure of a story.

To deal with the above two challenges, we formulate this task as a sequence-to-sequence learning problem and propose a novel joint learning model by leveraging the semantic coherence in a photo stream.

First, we learn a common semantic space by jointly em-

*This work was performed when Yu Liu was visiting Microsoft Research Asia as a research intern.

bedding each image with its corresponding sentence using semantic embedding in (Kiros, Salakhutdinov, and Zemel 2014) to alleviate visual variance problem. Semantic learning is important in tasks like multimedia representation and search (Mei et al. 2014). As other semantic learning works (Fu et al. 2014; 2015b; 2015a; Wang et al. 2016), we assume that the semantics of visual and text in telling the same story are relevant (Liu, Mei, and Chen 2016). Our idea is to learn a *semantic space* where the related photo-sentence pair can be close enough to reflect the same semantics. As a result, semantically related photos with large visual gap can be bridged by the neighborhood relation in the semantic space. Hence the visual variance is reduced and semantics are learned. For example, as in Figure 1, since it can be observed in the story dataset that "dinner/picnic" almost always follow a "ball game," we learn that these two visually different events should be semantically closed to each other. Moreover, a semantic relation matrix (coherence matrix), can be further identified by distance measure in this space. The coherence matrix is important to describe the semantic structure of the story, and will be used in an attention scheme to guide the training our sequence model for narrative paragraph generation.

Second, we propose a Bidirectional Attention-based Recurrent Neural Network (BARNN) to use the coherence matrix to enforce the sentence-to-sentence coherence. In our model, as the convention, the memory of each recurrent timestep encodes the deep feature (semantic in our case) of one photo/sentence in the sequence. Existing research (Park and Kim 2015) has proposed to use bidirectional RNN (BRNN) to capture sentence-level transition, and only considers the adjacent memories in the sequence. However, as we have shown in Figure 1, stories usually consist of much more complex structure, where the semantic of arbitrary timesteps in the sequence can be related. In this structure, each semantic memory may be contributed from arbitrary timesteps. For the example shown in storytelling of Figure 1, the term "the team" and "ball game" in sentence 1 and 2 can contributes to the prior knowledge "after the ball game, the team..." in sentence 3. To this end, we propose to design a framework with attention mechanism that integrates the memory semantics from various timesteps, via connections called *skip*. For this purpose, We design a novel recurrent unit by granting classic GRU with skip connections to allow this attention scheme. Thus the unit is called skip-GRU (sGRU). Note that the coherence matrix plays a role here of guiding the attention in how much information to contribute. Therefore, learning this BARNN can enable us to coherently model the sequence of photos. Note that we combine the learning of BARNN and previous semantic embedding into one objective function and train in end-to-end manner.

The contributions of the paper can be summarized as:

- The inherent challenges of visual storytelling are addressed by training an cross-modality embedding model, which can overcome the large visual variance in photo stream and represent the semantic of an underlying story.
- A novel BARNN framework with a new-designed skip gated recurrent unit (sGRU) is proposed to leverage implicit semantic relation in order to enforce the coherence of predicted sentences.
- Extensive experiments on the three storytelling datasets (NYC, Disney (Park and Kim 2015) and SIND (Huang et al. 2016)) have been carried out, and superior performance over the state-of-the-art with large margins in both retrieval and generation tasks has been obtained.

Related Work

Due to rapid growth of research interest recently in visual-to-language translation, there are a good number of related works has been carried out. They can be divided into three categories: single-frame to single-sentence, multi-frame to single-sentence and multi-frame to multi-sentence.

Single-frame to single-sentence modeling These researches focus on image captioning task, which can be classified into two sub-categories: semantic element based methods (Kulkarni et al. 2013; Farhadi et al. 2010; Mitchell et al. 2012; Yang et al. 2011) and Convolutional Neural Network (CNN) based methods(Vendrov et al. 2016; Kiros, Salakhutdinov, and Zemel 2014; Karpathy and Li 2015; Vinyals et al. 2015; Mao et al. 2015). In semantic element based methods, the regions of interest are first detected and represented in intermediate space defined by a group of semantic elements (object, action, scene) to fill in a sentence template. In CNN based model (Krizhevsky, Sutskever, and Hinton 2012), the CNN fully-connected layer output is extracted to represent the input images for classification.

Multi-frames to single-sentence modeling This family of approaches, mainly focus on video captioning to captures the temporal dynamics in variable-length of video frames sequence and to map them to a variable-length of words (Venugopalan et al. 2015a; Donahue et al. 2015; Venugopalan et al. 2015b). The sequence-to-sequence modeling are mainly relied on a RNN framework, such as Long-Short-Term-Memory (LSTM) (Hochreiter and Schmidhuber 1997). Moreover, bidirectional RNN (BRNN) is explored recently to model the sequence in both forward and backward passes (Peris et al. 2016). The approach proposed by (Yao et al. 2015) argues that a video has local-global temporal structure. They employ a 3D CNN to extract local action feature and an attention-based LSTM to exploit the global structure. However, this family of approaches have not yet exploited the visual variance and text coherence problem simultaneously in one single framework.

Multi-frame to multi-sentence modeling The work by (Park and Kim 2015) is the first scheme to explore the task of image streams to sentence sequence. They use a coherence model in textual domain, which is able to resolve the entity transition patterns frequently found between sentences. However, they define the coherence as rigid word reappearance frequency, which is unable to address the semantic gap and therefore cannot fully express the deeply meanings. Moreover, they focuses on textual coherence without acknowledging the problem of large visual variance.

Approach

We formulate the visual storytelling task as a sequence-to-sequence learning problem and propose a novel two-step

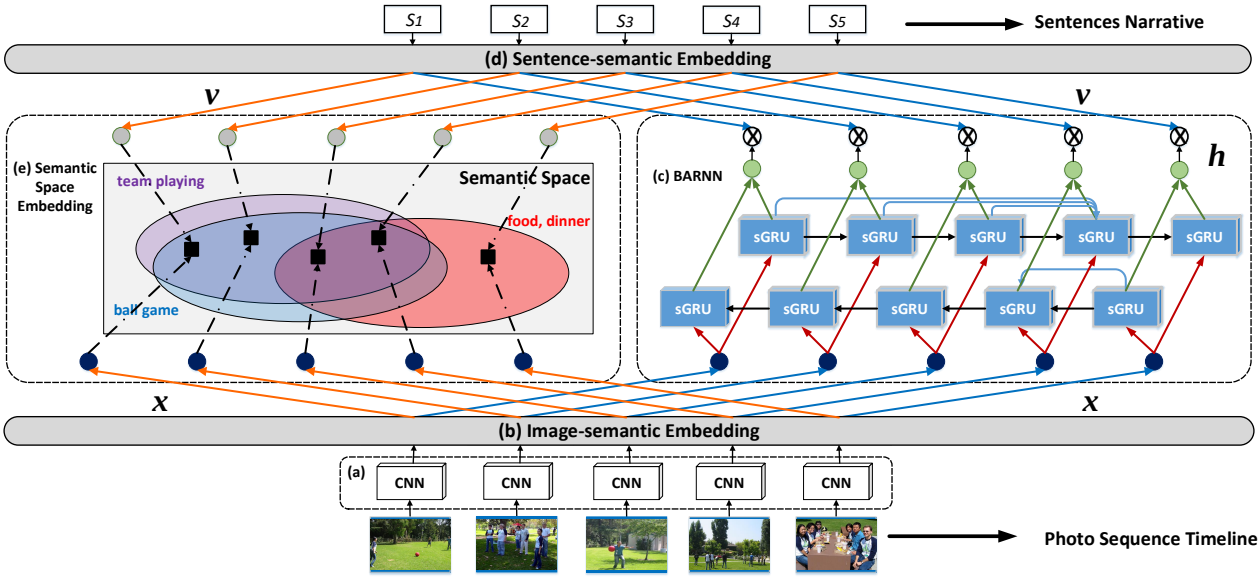

Figure 2: The framework of our approach. The images sequence is input to (a) to obtain 4096-dim VGG features, which is then mapped to 300-dim semantic embedding vectors $\boldsymbol{x}$ via (b). Then on the right side, $\boldsymbol{x}$ is fed to the (c) BARNN as a whole sequence to predict their corresponding sentence embedding vectors $\boldsymbol{h}$. Meanwhile, on the left side, $\boldsymbol{x}$ is matched in semantic space (e) with the groundtruth sentence embedding $\boldsymbol{v}$, which is the output of (d). To learn the model, the embedding loss (e) and BARNN loss (c) are minimized in the same objective.

approach. As shown in Figure 2, the proposed framework includes two main parts: (1) image and sentence embedding for joint semantic space learning, corresponding to part (a,b,d,e), and (2) BARNN for paragraph generation using semantic relation of coherence in part (c). The images embedding feature $\boldsymbol{x}$ from (a,b) are further translated to predict sentence embedding features $\boldsymbol{h}$ in (c). Meanwhile, we obtain the groundtruth sentence embedding vectors $\boldsymbol{v}$ from (d). Then we calculate the embedding loss by matching $\boldsymbol{x}$ with $\boldsymbol{v}$ in (e), as well as the BMRNN prediction loss of $\boldsymbol{h}$ given $\boldsymbol{v}$ in (c). Finally, the framework can be trained by iteratively minimizing the loss in both (e,c) as one objective. In test process, we feed $\boldsymbol{h}$ from (c) to a pre-trained language model to generate narrative paragraph.

Joint Embedding for Semantic Space

The goal of embedding model is to learn a common semantic space for photos and sentences. We assume the paired image and sentence in story share same semantics, as in image captioning problem. Thus, common semantic space is learnable using image captioning model, where we employ (Kiros, Salakhutdinov, and Zemel 2014) in our paper.

As in the approach by Kiros, the embedding model consist of two pipelines, the bottom-up for image and top-down for sentence. In Figure 2, the image pipeline consists of CNN in (a) for the 4096 dimension VGG features and a Feed-forward Network in (b). The output of (b) is a K-dim (K=300) vector $\boldsymbol{x}$ representing the image embedding. The sentence embedding (d) represents one sentence with the last hidden $\boldsymbol{v}$ of an LSTM which takes the Word2Vecs (Mikolov et al. 2013) of word sequence as input. Finally in (e), the image and sentence are embedded together by minimizing a contrastive loss:

$$C^{emb}(\boldsymbol{x},\boldsymbol{v}) = \sum_{\boldsymbol{x}\in\boldsymbol{X},\boldsymbol{v}\in\boldsymbol{V},\boldsymbol{v'}\in\boldsymbol{V'}} \max(0,\alpha-\boldsymbol{x}\boldsymbol{v}+\boldsymbol{x}\boldsymbol{v'}) + \sum_{\boldsymbol{x}\in\boldsymbol{X},\boldsymbol{x'}\in\boldsymbol{X'},\boldsymbol{v}\in\boldsymbol{V}} \max(0,\alpha-\boldsymbol{x}\boldsymbol{v}+\boldsymbol{x'}\boldsymbol{v}), \quad (1)$$

where $\boldsymbol{X}$ (or $\boldsymbol{V}$) are the image (or sentence) embedding vectors, $\boldsymbol{V'}$(or $\boldsymbol{X'}$) are the negative paired sentence (or image) samples. In our research, 127 negative pairs are randomly chosen from training set for each positive sample. α denote the contrastive margin (0.1 in our experiment).

In this learned space, an image becomes closer to others in one story if they share same semantics. This is because, as we assumed, it is trained by image-sentence pair shares semantics and the sentences are also coherently close. Thus it defines a semantic space where the distance between photos describes the semantic relation. Additionally, the distance between images in a story is reduced compared their original setting. As a result, the embedded input stream that has shorter dependency among states will be much easier for the model in next phase to learn.

Bidirectional Attention RNN (BARNN) for Textual Story Generation

The role of BARNN is to model the semantic structure of stories and generate coherent textual narratives. As in Figure 2 (c), the BARNN take sequence of image embedding vectors $\boldsymbol{x}$ as input, and product corresponding sentence embedding sequence $\boldsymbol{h}$ as output.

Attention with Semantic Structure From the embedding space that models the semantic closeness between images, one can easily infer the semantic relation by inner product of any two semantic embedding vectors:

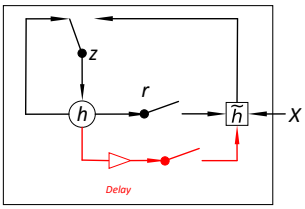

Figure 3: Our skip-GRU model. The black circuit stands for the classic GRU, and red part denotes the preservation scheme we added for skip with a specific delay

$$\boldsymbol{R}_{pt} = \boldsymbol{x}_p \boldsymbol{x}_t, \tag{2}$$

where $\boldsymbol{x}_p$ and $\boldsymbol{x}_t$ are image embeddings of image p and t. Note that $\boldsymbol{x}_p$ and $\boldsymbol{x}_t$ are normalized to keep $\boldsymbol{R}_{pt} <= 1$.

Taking advantage of the neighborhood preserving property (Hadsell, Chopra, and LeCun 2006) of contrastive loss in the learning of the embedding space, the relation $\boldsymbol{R}_{pt}$ reflects the coherence of sentence $\boldsymbol{v}_p$ and $\boldsymbol{v}_t$ to be predicted, since we assume that paired image-sentence share semantic. Thus, to leverage this implicit relation at the output end, we will attend on the relation $\boldsymbol{R}_{pt}$ as weight to allow the semantic of $\boldsymbol{x}_p$ to affect the semantic of $\boldsymbol{x}_t$ in a sequence model.

Skip-GRU (sGRU) In the BARNN model, we define sGRU and use it as basic unit. As in Figure 2 (c), the blue connections across arbitrary timesteps are skips. We name the new GRU as *skip Gated Recurrent Unit* (sGRU). In this section, we will first introduce the classic GRU, followed by the newly-designed sGRU. The complete BARNN framework will be described in details at the end.

The classic GRU, as proposed in (Cho et al. 2014), is a hidden unit used in RNN model for capturing long-range dependencies in sequence modeling. In Figure 3, the circuit in black shows the graphical depiction of the GRU design, specified by following operations:

$$\begin{aligned}
\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_{zx}\boldsymbol{x}_t + \boldsymbol{W}_{zh}\boldsymbol{h}_{t-1}) \\
\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_{rx}\boldsymbol{x}_t + \boldsymbol{W}_{rh}\boldsymbol{h}_{t-1}) \\
\widetilde{\boldsymbol{h}} &= tanh(\boldsymbol{W}_{hx}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) \\
\boldsymbol{h}_t &= \boldsymbol{z}_t\widetilde{\boldsymbol{h}} + (1-\boldsymbol{z}_t)\boldsymbol{h}_{t-1}
\end{aligned} \tag{3}$$

where t is the current time, $\boldsymbol{x}_t$ is the input, $\widetilde{\boldsymbol{h}}$ is the current hidden state and $\boldsymbol{h}_t$ is the output. $\boldsymbol{z}_t$ and $\boldsymbol{r}_t$ are *update gate* and *reset gate*, respectively.

To further take advantage of the implicit semantic relation and model the coherence structure in RNN, we propose a skip scheme to allow the communication between arbitrary states in the RNN, with attention on other semantics. As shown in Figure 3, we add a preservation scheme (red part) to the original design of GRU. Given the current time t and the previous time p, the memory of $\boldsymbol{h}_p$ has been saved, and reused after a delay $|t-p|$ with weight $\boldsymbol{R}_{pt} <= 1$ to help $\boldsymbol{h}_{t-1}$ predict the current hidden state of $\boldsymbol{h}_t$:

$$\begin{aligned}
\boldsymbol{s}_t &= \sigma(\boldsymbol{W}_{sx}\boldsymbol{x}_t + \boldsymbol{W}_{sh}\boldsymbol{h}_p) \\
\widetilde{\boldsymbol{h}} &= tanh(\boldsymbol{W}_{hx}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1} \\
&\quad + \sum\nolimits_{p<t} \boldsymbol{R}_{pt} \cdot \boldsymbol{W}_{hp}\boldsymbol{s}_t \odot \boldsymbol{h}_p) \\
\boldsymbol{h}_t &= \boldsymbol{z}_t\widetilde{\boldsymbol{h}} + (1-\boldsymbol{z}_t)\boldsymbol{h}_{t-1}
\end{aligned} \tag{4}$$

The formulas of gates $\boldsymbol{r}_t$ and $\boldsymbol{z}_t$ are the same with that in equation 3. $\boldsymbol{s}_t$ is skip gate to control how much information is used from $\boldsymbol{h}_p$. Similar to other gates, the design of skip gate $\boldsymbol{s}_t$ is controlled by current input and outputs of skip ancestor.

The advantages of the skip gate design are two folds. First, the attention scheme explores the semantic structure and guarantees the coherence between arbitrary states. The semantic invisible in current photo can be retrieved from other semantically close photos. Second, the skip gate ensures the non-linear mapping through skips. Such non-linear function is more powerful and more flexible in expressing complicated mappings. In contrast, linear combination proposed in (Ghosh et al. 2016), used as one of our baseline is expected to yield worse performance in the experiments.

Bidirectional Framework In this research, the bidirectional framework enables us to consider timesteps of both past and future. We rewrite the sGRU in the equation 4 into a compact form: $(\boldsymbol{z}_t, \boldsymbol{r}_t, \boldsymbol{s}_t, \widetilde{\boldsymbol{h}}, \boldsymbol{h}_t) = sGRU(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{R}, \boldsymbol{h}_p; \boldsymbol{W})$, to define the operations of the proposed BARNN components as:

$$\begin{aligned}
(\boldsymbol{z}_t^f, \boldsymbol{r}_t^f, \boldsymbol{s}_t^f, \widetilde{\boldsymbol{h}}^f, \boldsymbol{h}_t^f) &= sGRU(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^f, \boldsymbol{R}, \boldsymbol{h}_p^f; \boldsymbol{W}^f) \\
(\boldsymbol{z}_t^b, \boldsymbol{r}_t^b, \boldsymbol{s}_t^b, \widetilde{\boldsymbol{h}}^b, \boldsymbol{h}_t^b) &= sGRU(\boldsymbol{x}_t, \boldsymbol{h}_{t+1}^b, \boldsymbol{R}^T, \boldsymbol{h}_p^b; \boldsymbol{W}^b) \\
\boldsymbol{h}_t &= \boldsymbol{W}_h^f \boldsymbol{h}_t^f + \boldsymbol{W}_h^b \boldsymbol{h}_t^b
\end{aligned} \tag{5}$$

where f indicates forward pass and b denotes backward pass. The two passes neither have inter-communication nor share parameters, except for the input $\boldsymbol{x}_t$. Each pass is learned independently. In training, we learn the parameters $\boldsymbol{W} = \{\boldsymbol{W}^f, \boldsymbol{W}^b, \boldsymbol{W}_h^f, \boldsymbol{W}_h^b\}$ in equation 5. Note that the skip relation matrix of backward pass $\boldsymbol{R}^b$ can be obtained by transposing the forward pass $\boldsymbol{R}^f$, i.e. $\boldsymbol{R}^b = (\boldsymbol{R}^f)^T$.

For the compatibility measure in part (c), we employ again the contrastive loss with margin, which calculates:

$$\begin{aligned}
C_{(\boldsymbol{h},\boldsymbol{v})}^{cpt} &= \sum\nolimits_{\boldsymbol{v'} \in \boldsymbol{V'}} \max\{0, \gamma - \boldsymbol{h}\boldsymbol{v} + \boldsymbol{h}\boldsymbol{v'})\} \\
&\quad + \sum\nolimits_{\boldsymbol{h} \in \boldsymbol{H'}} \max\{0, \gamma - \boldsymbol{h}\boldsymbol{v} + \boldsymbol{h'}\boldsymbol{v}\}.
\end{aligned} \tag{6}$$

Similar to formula 1, $\boldsymbol{h}$ and $\boldsymbol{v}$ are the positive paired vectors, $\boldsymbol{v'}$(or $\boldsymbol{x'}$) are the negative paired (or image) sample to $\boldsymbol{h}$ (or $\boldsymbol{v}$). γ denote the contrastive margin (0.2 in our experiment).

Combination of Two Models

We jointly measure the embedding between visual to language and semantic compatibility of coherence in one objective function by summing all terms of C^{emb} and C^{cpt}:

$$C = \sum\nolimits_{(\boldsymbol{X},\boldsymbol{V})} C^{emb}(\boldsymbol{x}, \boldsymbol{v}) + \sum\nolimits_{(\boldsymbol{H},\boldsymbol{V})} C^{cpt}(\boldsymbol{h}, \boldsymbol{v}), \tag{7}$$

where $\boldsymbol{H} = \{\boldsymbol{h_1}, \boldsymbol{h_2}, ..., \boldsymbol{h_N}\}$ is the output of our BARNN model with a story photo stream input and $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_N}\}$ is the sentence sequence to be matched. And $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ...\boldsymbol{x_N}\}$ are the image embedding vectors. For the training, we take an iterative scheme which alternatively minimizes the two terms of the object. We iterate between the training of the two parts when validation error stops decreasing in m=5 epochs.

	NYC			
	R@1	**R@5**	**R@10**	**Medr**
Random	0.17	0.25	0.59	763
1NN	5.95	13.57	20.71	63.5
BARNN-sGRU	16.23	28.7	39.53	19
BARNN-EMB	17.27	29.42	38.97	19
BCLSTM	15.10	29.91	41.07	18
CRCN	11.67	31.19	43.57	14
BARNN	**29.37**	**45.43**	**52.10**	**8**

Table 1: Sentence retrieval evaluation on NYC

	Disney			
	R@1	**R@5**	**R@10**	**Medr**
Random	0.26	1.17	1.95	332
1NN	9.18	19.05	27.21	45
BARNN-sGRU	19.97	37.48	46.04	14
BARNN-EMB	21.57	39.24	46.50	12
BCLSTM	19.77	38.92	45.20	14
CRCN	14.29	31.29	43.2	16
BARNN	**35.01**	**49.07**	**57.83**	**6**

Table 2: Sentence retrieval evaluation on Disney

	SIND			
	R@1	**R@5**	**R@10**	**Medr**
Random	0.0	0.04	0.10	2753
1NN	4.8	13.00	21.07	74
BARNN-sGRU	21.39	38.72	46.96	14
BARNN-EMB	21.63	38.54	47.01	14
BCLSTM	21.47	37.30	47.39	18
CRCN	9.87	28.74	39.51	21
BARNN	**24.07**	**44.29**	**53.06**	**9**

Table 3: Sentence retrieval evaluation on SIND

Test Process with Language Model

In the testing, we feed the image test data to the framework and obtain the predicted sentence embedding features $\boldsymbol{h}$ from (c) in Figure 2. Note that we do not have the ground truth sentences in testing. The predicted features $\boldsymbol{h}$ are then stacked in order and input to a pre-trained language model to obtain paragraph output.

We build the language model using a LSTM as (Venugopalan et al. 2015a). The LSTM read each embedding feature of $\boldsymbol{h}$ as input, and then generate one word at each timestep. Finally, multiple sentences are stacked in order to produce the narrative paragraph.

Experiment

Both retrieval and generation tasks are evaluated for our approach. For retrieval task, we compare the performance of the approach in three datasets against a group of the-state-of-art methods consisting of both existing models and variations of the proposed model. Two type of measures are used: quantitative measures and user study. For generation task, we perform in test set to produce novel paragraph as a whole. We then evaluate the language by METEOR/BLUE/CIDEr and compared to the-state-of-art baselines.

Experiment Setting

Dataset We make use of three recently proposed datasets, the SIND (Huang et al. 2016), NYC and Disneyland dataset (Park and Kim 2015). All three datasets consist of sequential image-stream-to-sentence-sequence pairs.

Specifically, the SIND is the first dataset particularly created for sequential vision-to-language and other story related tasks (Agrawal et al. 2016). It contains 48,043 stories with 210,819 unique photos. The image streams are extracted from Flickr and the text stories are written by AMT. Each story consists of 5 images and 5 corresponding sentences for a story. The dataset has been split into 38,386 (80%) stories as training set, 4,837 (10%) as test set and 4,820 (10%) as validation set. The NYC and Disney datasets are automatically generated from blog posts searched with travel topics NYC and Disneyland, in total 11,861 and 7,717 stories, respectively. We follow the splitting of dataset in that 80% as training set, 10% as validation set and the others as test set.

Retrieval Task The framework retrieves the best stories from the training set and compares with groundtruth (GT). For evaluation, both quantitative measures and user study are employed. For quantitative measure, the Recall@K metric and median rank are used. Recall@K indicates the recall rate of the GT retrieval given top K candidates while the median rank is the median rank value of the retrieved GT. The higher Recall@K and lower median rank value, the better the performance. For user study, 40 users are invited to give rating on 200 randomly chosen results of the proposed approach, another the-state-of-art baseline method and the GT.

Baselines for Retrieval Task In the experiments, we consider both state-of-art methods from the existing models and variations of the proposed model. Since the visual storytelling is a relatively new research direction, there are only few existing research works to compare with. To the best of our knowledge, the most closely related work is (CRCN) (Park and Kim 2015). Besides, we also adopt the state-of-art models in vision-to-language tasks as baseline, such as video description using CNN and BRNN of (Peris et al. 2016). Particularly, we keep the semantic embedding part in this framework, which makes a variation to our model without attention scheme and sGRU, so its called (BARNN-sGRU). Comparing against (BARNN-sGRU) we evaluate the effectiveness of the proposed sGRU architecture. Likewise, without the part of semantic embedding, (BRNN-EMB) evaluate the embedding part. To validate our claim on the non-linear mapping of sGRU, we compare to CLSTM unit which is in linear fashion (Ghosh et al. 2016). It's a bidirectional framework, hence we call this baseline as (BCLSTM). We also test the K-NN search (1NN) without sequential modeling part, which equals to a search based single image captioning baseline. This comparison demonstrates the value of modeling the entire photo stream rather than single one. We add random retrieval scheme (Random) as a simple baseline.

Generation Task We pre-train a LSTM language model by using an additional Book Corpus Dataset (Zhu et al. 2006), and then tune it on the storytelling dataset SIND. If we view the sentence embedding part (d) as an encoder from text sentence to embedding features, the language model will act as a decoder in the opposite way. Therefore, we can use (d) to create a synthetic dataset from books to pre-train the language model which can approximate the inverse mapping of (d). Then it will be fine-tuned by the storytelling

Ground truth: It was an exciting day, very dramatic! Someone there had extra flags for us. There was already a big crowd when we got there. We got a few nice pictures near the fountain. We continued to demonstrate, making sure to be heard.
CRCN: The people kept marching as the crowd got even bigger. The people worked together to have their opinion heard. The group kept getting bigger and bigger. The crowd marched in peaceful protest. Some of the people held signs in protest.
Proposed: Everyone was excited to see what type of event would unfold. Some people came through with a banner, and we were hoping it was the start of a parade. More and more people kept coming, and it started getting hard to see what was happening. There was a celebration in town, and everyone came down to see what was going on. Some people started to head home, and we thought that might a good idea for us too.

Ground truth: The flag corps led the parade down the streets. The mayor closed up the parade waving at the crowd as he passed. The flag twirlers came out, but had some problems staying in sync. Onlookers gathered on the sides of the street for the parade. Then the local marching band entertained us.
CRCN: They're much trickier than a normal kite. This group finally got theirs in the air long enough to place second. It's time for the annual kite flying contest in location. One of the contests is managing to fly a box kite. Some of the kites go pretty high in strong wind.
Proposed: Quite a lot of participants and spectators. Participants gather the night before the wildflower triathlon. Getting ready to start the first leg. Participants in good spirits for the event. Another wave gets ready to swing.

Figure 4: Examples of visual storytelling result on SIND. Three stories are generated for each photo stream: story by GT, story by baseline CRCN and story by the proposed BARNN. The colored words indicate the semantic matches between the generation results with the GT. The proposed scheme shows better semantic alignment with the GT than the baseline. [Best viewed in color]

dataset. The generation performance of our approach is compared with the baselines in (Huang et al. 2016) with a machine translation metric METEOR.

Results and Discussion

The quantitative results of story sentences retrieval are shown in Table 1, 2 and 3. We observe that we perform better in a large margins than other baselines on all datasets, that confirms our analysis on the visual variance and semantic relation, and the proposed BARNN model can effective capture and leverage this semantic relations to improve the performance in visual storytelling. Specifically, comparison with (CRCN) shows that visual modeling on photo stream can better accomplish the visual storytelling task, rather than just capturing rigid coherence in textual domain. We also found that the proposed scheme outperforms (BARNN-sGRU) and (BARNN-EMB) variations, which verifies the two importance phase of our proposed approach, the sGRU model and semantic embedding model, respectively. Moreover, it is because of the non-linear design that empowers the sGRU to capture the skipping information, since it achieves higher results than (BCLSTM) where a linear scheme is used. Note that we take use the same VGGNet fc7 feature (Simonyan and Zisserman 2015) in all baselines for fair comparison.

We also discover that the visual models (BARNN-sGRU) and (BARNN-EMB) all yield much better results than (CRCN) under the retrieval metrics. This verifies that modeling from visual domain rather than textual domain can better accomplish the storytelling task. The (1NN) baseline shows unsatisfactory results, indicating that the visual story is not a simple concatenation of individual image captions.

In Table 4, the narrative generation results are measured by METEOR on the SIND dataset and compared with the baselines methods in (Huang et al. 2016). This validates the capability of the proposed model in creating novel sentence that describes the semantic of story. The four baseline models are all built on a regular sequence-to-sequence RNN, with beam search $beam = 10$ (**Beam=10**), greedy search (**Greedy**), rule-based de-duplication (**-Dup**) and visually grounded words from captioning (**Grounded**). The proposed scheme yields better performance because (1) the coherence problem is infeasible in the plain RNN in (**Beam=10**) and (**Greedy**), and (2) the semantic is properly modeled and the deeply meaning in sentences is suitably expressed to avoid the rigid heuristic rules employed by (**-Dup**) and (**Grounded**).

In Table 5, we compare our narrative generation results against CRCN with BLUE (Unigram) and CIDEr, on both NYC and Disney datasets, denoted as **BLUE(N)**, **CIDEr(N)**, **BLUE(D)** and **CIDEr(D)** in the table. The scores of CRCN are reported in paper (Park and Kim 2015).

Proposed	Beam=10	Greedy	-Dup	+Grounded
33.32	23.13	27.76	30.11	31.42

Table 4: METEOR score of our generation approach and baseline.

Method	BLUE(N)	CIDEr(N)	BLUE(D)	CIDEr(D)
CRCN	26.83	30.9	28.15	51.3
Proposed	39.3	41.6	37.7	54.1

Table 5: METEOR score of our generation approach and baseline.

User Study

We perform user studies to test the preference on the stories by groundtruth, the proposed model and the baselines. Since only (CRCN) is originally proposed for the sequential vision-to-language task, we choose this method as baseline in the user study. We randomly choose 200 stories from test set of the SIND, each associated with three stories: story from groundtruth (GT),story generated by (CRCN) and story by our proposed method (BARNN). Please see Figure 4 for examples. 40 users are invited to score on the stories with a subjective score of 1-10 (Best story = 10). All these three stories, including the groundtruth stories, are read by the users and scored.

Table 5 shows the user study results, where the last row is the mean score over all samples. We infer the user preference between two stories by comparing their scores from the same person. Equal scores indicate no preference to any method and thus are not considered in preference inference. Rows 1-3 in Table 5 show pairwise preference of each method against others.

We observe that the stories by the proposed scheme is much preferred over (CRCN) in user study. Over all users, the mean score are higher than (CRCN). On average, 68.3%

of users prefer the proposed scheme over (CRCN), while only 13.5% of users prefer (CRCN). From another prospective, 7.0% of the users even prefer the proposed scheme over GT while there is only 2.0% of the user prefer the (CRCN) over GT. These results all confirm the results obtained by the proposed scheme are preferred over the baseline.

	GT	CRCN	BARNN
GT	-	96.3%	86.8%
CRCN	2.0%	-	13.5%
BARNN	7.0%	68.3%	-
Mean Score	8.49	3.67	5.16

Table 6: The evaluation results of user study. The row 1-3 are pairwise preference and the last row is the mean evaluation scores.

Conclusion

In this paper, we presented a framework for visual storytelling, to generate human-level narrative from photo stream. We addressed the inherent challenges of visual variance and textual coherence. In this research, we designed a novel BARNN with a new-designed sGRU model, with attention on sematic relation extracted from space space to enhance the textual coherence in narrative output. Extensive experiments confirm the effectiveness of the proposed model in both retrieval and narrative generation tasks. The proposed BARNN outperforms the-state-of-art models with large margins.

References

Agrawal, H.; Chandrasekaran, A.; Batra, D.; Parikh, D.; and Bansal, M. 2016. Sort story: Sorting jumbled images and captions into stories. *arXiv:1606.07493*.

Cho, K.; Merrienboer, B. V.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder decoder for statistical machine translation. *EMNLP*.

Donahue, J.; Hendricks, L. A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. *CVPR*.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. *ECCV*.

Fu, J.; Wang, J.; Rui, Y.; Wang, X.-J.; Mei, T.; and Lu, H. 2014. Image tag refinment with view-dependent concept representations. *CSVT*.

Fu, J.; Mei, T.; Yang, K.; Lu, H.; and Rui, Y. 2015a. Tagging personal photos with transfer deep learning. *WWW*.

Fu, J.; Wu, Y.; Mei, T.; Wang, J.; Lu, H.; and Rui, Y. 2015b. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. *ICCV*.

Ghosh, S.; Vinyals, O.; Strope, B.; Roy, S.; Dean, T.; and Heck, L. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv:1602.06291*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. *CVPR*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*.

Huang, T. H.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; Zitnick, C. L.; Parikh, D.; Vanderwende, L.; Galley, M.; and Mitchell, M. 2016. Visual storytelling. *NAACL*.

Karpathy, A., and Li, F.-F. 2015. Deep visual-semantic alignments for generating image descriptions. *CVPR*.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual semantic embeddings with multimodal neural language models. *NIPS deep learning workshop*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *NIPS*.

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. 2013. Babytalk: Understanding and generating simple image descriptions. *TPAMI*.

Liu, Y.; Mei, T.; and Chen, C. W. 2016. Automatic suggestion of presentation image for storytelling. *ICME*.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*.

Mei, T.; Rui, Y.; Li, S.; and Tian, Q. 2014. Multimedia search reranking: A literature survey. *ACM Computing Survey*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *NIPS*.

Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; and III, H. D. 2012. Midge: Generating image descriptions from computer vision detections. *EACL*.

Pan, Y.; Mei, T.; Yao, T.; Li, H.; and Rui, Y. 2016. Jointly modeling embedding and translation to bridge video and language. *CVPR*.

Park, C. C., and Kim, G. 2015. Expressing an image stream with a sequence of natural sentences. *NIPS*.

Peris, Á.; Bolaños, M.; Radeva, P.; and Casacuberta, F. 2016. Video description using bidirectional recurrent neural networks. *arXiv:1604.03390*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2016. Order-embeddings of images and language. *ICLR*.

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015a. Sequence to sequence - video to text. *ICCV*.

Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; and Saenko, K. 2015b. Translating videos to natural language using deep recurrent neural networks. *NAACL*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. *CVPR*.

Wang, J.; Fu, J.; Xu, Y.; and Mei, T. 2016. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. *IJCAI*.

Yang, Y.; Teo, C. L.; III, H. D.; and Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. *EMNLP*.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Describing videos by exploiting temporal structure. *ICCV*.

Zhu, Y.; Kiros, R.; Zemel, R. S.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2006. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CVPR*.