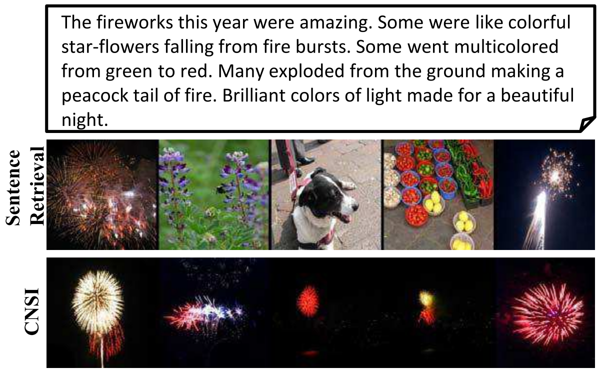
Show Me a Story: Towards Coherent Neural Story Illustration

Hareesh Ravi* , Lezi Wang* , Carlos M.Muniz* , Leonid Sigal† , Dimitris N.Metaxas* , Mubbasir Kapadia*
* Rutgers University
† University of British Columbia
*{hr268,lw462,cmm609,dnm,mk1353}@cs.rutgers.edu, †{lsigal}@cs.ubc.ca

Abstract

We propose an end-to-end network for visual illustration of a sequence of sentences forming a story. At the core of our model is the ability to model the inter-related nature of the sentences within a story, as well as the ability to learn coherence to support reference resolution. The framework takes the form of an encoder-decoder architecture, where sentences are encoded using a hierarchical two-level sentence-story GRU, combined with an encoding of coherence, and sequentially decoded using a predicted feature representation into a consistent illustrative image sequence. We optimize all parameters of our network in an end-to-end fashion with respect to order embedding loss, encoding entailment between images and sentences. Experiments on the VIST storytelling dataset [9] highlight the importance of our algorithmic choices and efficacy of our overall model.



The fireworks this year were amazing. Some were like colorful star-flowers falling from fire bursts. Some went multicolored from green to red. Many exploded from the ground making a peacock tail of fire. Brilliant colors of light made for a beautiful night.

Figure 1. Two image sequences visualize the given story (top). The images predicted by the proposed sequential model with coherence (bottom) demonstrate higher consistency and better alignment with the story than the images retrieved independently sentence-by-sentence (middle).

1. Introduction

There is an unprecedented wealth of multimedia data (image, video and text) on the web which stems from the availability of accessible imaging devices (*e.g*., cell phone and tablets) and the avid use of social media. Availability of this rich data along with recent algorithmic developments in neural architectures has resulted in the wealth of multi-modal image/video-text approaches. Typical problems include image captioning [10], natural language-based image retrieval [22], and joint embedding of text and images to understand the relationship and be able to translate between the two modalities [29]. However, most of these formulations assume atomic image-sentence pairings both at training and test time. This makes it difficult to apply them for storytelling tasks where sentences are implicitly inter-related in a narrative.

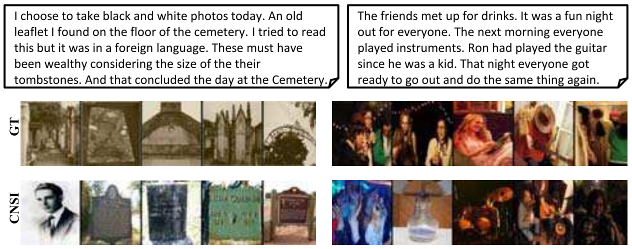
Recent approaches have started to address these challenges by proposing datasets (*e.g*., VIST [9]) and hierarchical language decoders that are able to generate multiple sentences [14], or paragraph descriptions, forming stories [7]. While some limited success has been shown, most of these approaches (with the exception of [12]) attempt to go in a forward direction, producing a multi-sentence description for an image [14], video [33], or image sequence [9]. In this paper, we address the relatively unexplored, inverse problem of generating illustrations for inter-related sentences forming a narrative story. This problem is important for a variety of creative applications, including automatic storyboarding in film, storytelling [9], and story creation. The key difference with story captioning is that images tend to be more expressive (*e.g*., an image is worth a thousand words), making it challenging to produce a coherent sequence of illustrations.

In this paper, we propose a method for retrieving a sequence of illustrative images which correspond to a narrative passage of text. Fig 1 illustrates the problem and our solution (CNSI) for a given input passage (top). The images in the middle row are predicted by the baseline where a sentence-image similarity is learned [27] and images are independently retrieved for each sentence in the passage. The key benefit of encoding context, through hierarchical Gated Recurrent Units (GRU) [2], and coherence is illustrated by images in the bottom row. In fact, only the first sentence in the story mentions the word 'fireworks', but all of the images in the output share that visual, highlighting that the references were correctly resolved. This example clearly illustrates the key benefits of our proposed model.

We develop an end-to-end network for visual illustration of a sequence of sentences forming a story. We refer to an input storytelling sequence of sentences as SIS (story-in-sequence) and the corresponding output visual summary of images as IIS (images-in-sequence). We achieve this by building a neural architecture that takes the form of an encoder-decoder, where sentences are encoded using hierarchical two-level sentence-story GRU, combined with an encoding of coherence, and then sequentially decoded using the predicted feature representation into consistent IIS. We optimize all parameters of our network in an end-to-end fashion with respect to the order embedding loss. The resultant model tries to sequentially *translate* each input sentence vector to a representative output image vector. The image closest to this output vector is then retrieved from a large dataset to illustrate the corresponding sentence.

Due to the ambiguity of the task, existing quantitative metrics, such as mAP [16], produce misleading results, as these metrics are computed based on the exact image IDs. For example, as shown in Fig 2, the precision goes to zero because predictions (in the bottom) consists of different images from the ground truth (middle). However, the predicted images are actually *prefered*, to ground truth, by Amazon Mechanical Turk (AMT) subjects in terms of better story visualization. To address this, we first perform a user study on AMT to evaluate the performance of the proposed architecture, in comparison to ground truth, and existing baselines [27] of visual retrieval for isolated sentences, and ablated model without coherence. Results indicate that the proposed model outperforms the baseline, and users prefer image sequences with coherence. Additionally, we propose a new quantitative metric for this task based on the visual saliency of the retrieved images with respect to the ground truth images. We show this metric serves as a good proxy for measuring whether a predicted image can be considered a good visual illustration.

Contributions: Our core contribution is an end-to-end architecture for retrieving a sequence of illustrative images from a set of sentences, one for each sentence, forming a story. We model context between sentences using hierarchical two-level sentence-story GRU. Further, since it is natural to use references (*e.g*., direct: *he/she/it* and indirect: *they both went there*) within the story once actors and objects are defined, we also introduce a coherence vector to help with such reference resolution. Evaluation of the proposed architecture with a user study shows that the proposed algorithm performs better than the baseline in a comprehensive ablation study. Further, we introduce a new metric for this task that can better deal with ambiguities in the image selection.


I choose to take black and white photos today. An old leaflet I found on the floor of the cemetery. I tried to read this but it was in a foreign language. These must have been wealthy considering the size of the their tombstones. And that concluded the day at the Cemetery.

The friends met up for drinks. It was a fun night out for everyone. The next morning everyone played instruments. Ron had played the guitar since he was a kid. That night everyone got ready to go out and do the same thing again.

Figure 2. Two samples of ground truth (middle) and our prediction (bottom) for the story (top). The bottom sequence of images win more votes from AMT workers for better visual illustration.

2. Related work

Our work is related to the rich literature on multi-modal image-text representation and learning, including caption generation, and natural-language based retrieval. Here we overview only the most relevant related works.

Caption generation: Caption generation for images is a well studied problem in the vision community [5, 18, 24, 30, 31, 32, 34, 35]. Most recent works use some form of Recurrent Neural Network (RNN) to generate the captions word by word given the encoding of the image. Particularly, He *et al*. [8] use Part of Speech (POS) tags from image descriptions, while Jia *et al*. [30] utilized semantic information from images to guide Long Short Term Memory (LSTM) to generate meaningful descriptions. Vinyals *et al*. [28] and Chen *et al*. [1] used CNN-based image encoders and an RNN and bidirectional RNN for decoding, respectively. Attention-based neural networks are also popular and allow the model to focus on specific regions when generating individual words [34]. Johnson *et al*. [10] proposed a convolutional localization network to create dense captions for an image in a single forward pass. Captions, in general, have been defined as a descriptive piece of text that contain information about the objects and scene in the image. We wish to encode the relationship between multiple inter-related sentences that form a narrative story and corresponding images.

Image Retrieval: Image retrieval is considered to be the reverse of caption generation. Most techniques [4, 6, 17, 21, 23, 27] that deal with image retrieval also evaluate their algorithms on caption generation. For example, Wang *et al*. [29] proposed structure preserving and bidirectional ranking constraints, while Zhou *et al*. [23] formulated a Gaussian visual-semantic embedding; Gong *et al*. [6] used a multi-view version of Canonical Correlation Analysis (CCA) for joint representation of phrases and images. Vendrov *et al*. [27] proposed to leverage intuition that a correct caption-image pair can be considered as ordered, with the caption being a more abstract representation of the image. This effectively encoded entailment relationship between images and captions. Evaluations for this method produced state-of-the-art results for image retrieval and caption generation. Reed *et al*. [22] proposed a Generative Adversarial Network to synthesize images from descriptive text. Even though these papers have addressed the problem of visual illustration of text, they lack ability to deal with sequential structure of a textural story, which is the focus of this paper.

Images in sequence to story in sequence (IIS to SIS): The sequence of images is, in general, a continuous stream of consistent images that are part of the same story or event. Park and Kim [20] introduced a coherent recurrent convolutional network to explicitly model coherence within the text and jointly embed a sequence of text and the corresponding sequence of images. Huang *et al*. [9] proposed a novel visual storytelling dataset (VIST) for IIS to SIS generation. They also proposed a simple baseline RNN to enable IIS to SIS task. We make use of this dataset for evaluating our algorithm for the reverse task of SIS to IIS. Yu *et al*. [36] proposed an automatic hierarchically-attentive RNN to automatically summarize an album by selecting the most representative set of photos and then generated a natural language story. Evaluation using the VIST [9] dataset showed state-of-the-art results for this task. All of the papers described above deal with going from sequential images to sequential text; we are looking to do the reverse. Applying the same techniques in reverse is not trivial due to the differences in the information contained in the input and output.

Story in sequence to images in sequence (SIS to IIS): The task of visually illustrating a sequence of inter-related sentences, forming a story, is the problem addressed by this paper. To the best of our knowledge, we could not find any previous work that directly deals with this task. Kim *et al*. [11, 12] are the closest. Both of these papers deal with blogs or large pieces of text that require short visual summaries. In particular, they look at blogs and photo streams corresponding to a common topic. Blogs contain a relatively large volume of text and tend to have information related to 'location', 'time', 'ride name' and other specific details that make the retrieval space more constrained. Our approach deals with sentences that are conceptually more abstract, less descriptive and (by design) more closely interrelated into a story.

3. Proposed Method

We develop the SIS to IIS retrieval algorithm as follows. Let $\mathbf{S} = \{s_1, s_2, ..., s_n\}$ be a set of n sentences (though in principal these can also be paragraphs or other natural text elements) that tell a narrative story and let $\mathbf{I} = \{i_1, i_2, ..., i_n\}$ be the set of n corresponding images from the dataset that best illustrate the input SIS. We consider n to be equal to five in our paper to adhere with the VIST [9] dataset, however, the algorithm is general and is able to deal with arbitrary length sequences. There is, however, an implicit assumption that illustration is one-to-one, meaning that for every sentence (element of the story) we retrieve one illustration.

We build an end-to-end network, as shown in Fig 3, which encodes the input sentences and coherence and predicts encoding of the feature representation for corresponding illustrative images. We now explain each component of the network in detail.

3.1. Sentence Encoding

The process of sentence encoding is illustrated in Fig 3(b). Let each input sentence s_j have n_j words $\{w_1, w_2, ..., w_{n_j}\}$. The GRU RNN network in the first stage sequentially encodes $\{w_1, w_2, ..., w_{n_j}\}$, to generate feature vector $f_1(s_j)$ representing the *j*-th sentence. The corresponding image feature vector $g_1(i_j)$ is obtained by using a pre-trained VGG-16 model [26]. This entire network is trained on the MS-COCO dataset [15] so that $f_1(s_j)$ will be aligned with $g_1(i_j)$. Conceptually, this is to form initial sentence representations that are closer to image vector representations for each sentence in the story. The Order embedding loss (OE-Loss) function as defined in [27] is used to train this network. These encoded representations are then used to initialize the next part of the network that performs sequential story encoding in Fig 3(c).

3.2. Story encoding

The output of the sentence encoder is a sequence of individual sentence encodings: $\{f_1(s_1), f_1(s_2), ..., f_1(s_n)\}$. To encode the story structure among them, we pass this to a higher level network that encodes the sequential (from left to right) nature of the input story to produce vectors: $\{f_2(s_1), f_2(s_2), ..., f_2(s_n)\}$. The target vectors remain the same from above $g_1(i_j)$. A sequential, order embedding loss function is used to train this network to constrain $f_2(s_j)$ to be as close as possible to $g_1(i_j)$. This process is depicted as the sequential model in Fig 3 (c). For sequential story encoding, a three layered RNN with GRU cells [2] is applied to model the shared context between sentences in a story.

3.3. Coherence Model

Even though the GRU sequential model encodes the relationship between sentences, we also explicitly model coreferences between sentences to further improve the ability of our model to capture story structure. We do this by making use of the coherence model proposed in [20]. The authors represent the coherence between sentences, within a piece of text, by a 64 dimensional coherence vector obtained from the parse tree associated with each sentence in the story. They use this vector as input to the final Fully Connected (FC) layer in their network, after zero padding

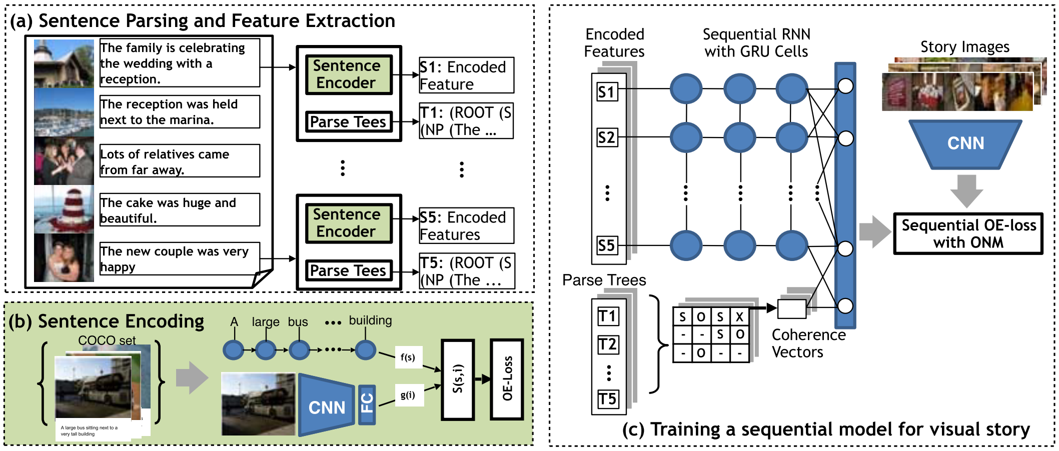

(a) Sentence Parsing and Feature Extraction
The family is celebrating the wedding with a reception.
The reception was held next to the marina.
Lots of relatives came from far away.
The cake was huge and beautiful.
The new couple was very happy
Sentence Encoder
Parse Tees
S1: Encoded Feature
T1: (ROOT (S (NP (The ...
Sentence Encoder
Parse Tees
S5: Encoded Features
T5: (ROOT (S (NP (The ...
(b) Sentence Encoding
COCO set
A large bus ... building
f(s)
g(i)
CNN
FC
S(s,i)
OE-Loss
Encoded Features
Sequential RNN with GRU Cells
Story Images
S1
S2
S5
CNN
Sequential OE-loss with ONM
Parse Trees
T1
T2
T5
S O S X
- - S O
- O - -
Coherence Vectors
(c) Training a sequential model for visual story

Figure 3. The proposed approach for story visualization: (a) Shows modeling of isolated sentences for sentence encoding and parse tree extraction for coherence vector computation. (b) Uses encoded sentence vectors to train sentence-RNN using Order Embedding (OE) loss function [27]. (c) Uses encoded sentence and image vectors along with coherence vector to sequentially encode the input story. Story-RNN is trained using modified OE-loss function with On-line Negative Mining(ONM) for generating negative samples within a batch.

to match dimensions of the input vectors. We, however, directly concatenate the vector with each sentence before the final FC layer. This is visualized in Fig 3(c).

3.4. Loss function

We use an order embedding loss function based on the one proposed in [27] for our network. The assumption is that a short textual description of an image is more abstract than the image itself. The description-image pair can therefore be considered as an ordered pair. Since SIS text is in general more abstract, we use a similar order embedding constraint based model. For ours, the cost of a sentence-image ordered pair violating the order is defined as:

$$E(x, y) = max||0, (y - x)||^2 \tag{1}$$

where $E(x, y) = 0 \iff x \preceq y$ according to the reversed product order. If the order is not satisfied, then $E(x, y)$ is positive. If we treat the sentence-image pair (s_j and i_j) as a two level partial ordering, then we can define $S(s_j, i_j)$ as follows:

$$\begin{aligned} S(s_j, i_j) &= -E(g_1(i_j), f_2(s_j)) \\ &= -max||0, (f_2(i_j) - g_1(s_j))||^2 \end{aligned} \tag{2}$$

where $S(s_j, i_j)$ is the negative order violation penalty for a ground-truth sentence-image pair. The objective is then to maximize this for a ground truth pair relative to other pairs by a margin. Here, $f_2()$ and $g_1()$ are the SIS and IIS encoders as described in Sec 3.2. The loss function to be minimized is then:

$$\begin{aligned} c = \sum_{k=1}^{l} \sum_{j=1}^{5} \Bigg(& \sum_{s'_{k,j}} max\{0, \alpha - S(s_{k,j}, i_{k,j}) + S(s'_{k,j}, i_{k,j})\} \\ & + \sum_{i'_{k,j}} max\{0, \alpha - S(s_{k,j}, i_{k,j}) + S(s_{k,j}, i'_{k,j})\} \Bigg) \end{aligned} \tag{3}$$

where c is the cost for a batch with in size of l. Index k iterates over each story within a batch while index j iterates over each positive ground truth sentence-image pair within each story. Given a batch of story sentence-image pairs, we apply Online Negative Mining (ONM) to generate negative samples [27]. The negative samples for each ground-truth pair are taken from all other stories except the one in consideration. In other words, for a sample $(s_{1,1})$, the corresponding negative samples are $(s'_{k,j}, k \neq 1, j = \{1, 2, 3, 4, 5\})$. Also, j is chosen uniformly at random between the five indices for each negative story. Before each epoch, all the samples are arranged and shuffled carefully to avoid identical images occurring in different stories.

4. Training Details

The proposed hierarchical GRU network with order embedding loss function can be completely seen in Fig 3(c). The sentence encoder is trained on MS-COCO dataset [15] using a joint image-sentence embedding formulation. We believe this ensures a good initial aligned representation for both modalities. The resultant vector for each sentence is given as input to the sequential model to encode the relationship between the sentence vectors. The loss function is explained in Sec 3.4; it calculates loss between the five encoded vectors and the corresponding five image vectors

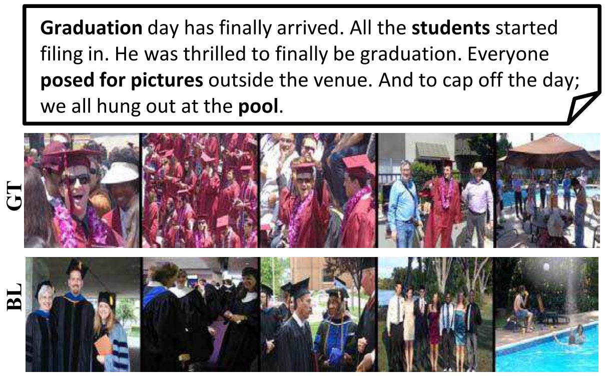

Graduation day has finally arrived. All the **students** started filing in. He was thrilled to finally be graduation. Everyone **posed for pictures** outside the venue. And to cap off the day; we all hung out at the **pool**.
GT
BL

Figure 4. Two image sequences visualize a "graduation" story. AMT workers prefer the `GT` over `BL`, though both look similar.

obtained from a pre-trained VGG16 CNN network. For learning we use Adam optimizer [13] with a learning rate of 0.001. The batch size is 32 stories, a relatively low number to prevent repetition of stories or images in each batch.

Note that out of the $40,149$ stories present in the training dataset for VIST [9], there are only $16,041$ unique story image sequences. This means that multiple SIS can correspond to a single sequence of images. There is a high possibility that in a naïve implementation the order embedding loss function would get the same sequence of images as both positive and negative during the training. This is obviously undesirable. In addition, multiple stories can share different permutations of the same sequence of images causing same image to be seen as both positive and negative illustration for a sentence. Also, the algorithm performs prediction over the entire dataset for each time instant (sentence) during retrieval; therefore during training, we apply Online Negativing Mining to obtain the negative samples from dissimilar stories ($k \neq 1$ in Eq 3) from a disjoint set of instants (j chosen uniformly at random in Eq 3).

The learning is set to run for 150 epochs. From observations, we find that after 130 epochs the loss value starts to saturate at around 3.0. During testing, each sequence of input sentences goes through the network and produces a sequence of image vectors. All images in the dataset go through the CNN part of the pre-trained baseline. Then, for each output vector from network, the image with the closest CNN feture vector is chosen as the predicted image.

We implement our networks and loss function in python using Tensorflowand Keras. The network is trained on the $40,154$ training set stories in the VIST [9] dataset. We remove stories that have broken URLs or images. We perform qualitative and quantitative evaluation on a subset of test set stories that have captions for all of the images. This reduces the number of test stories from $5,054$ to $3,384$. Retrieval is performed over this entire set with $5,055$ candidate images, *i.e.*, $3,384$ stories with 5 images each in the test set have a total of $5,055$ *unique* images. Each epoch takes approximately 250 seconds in a desktop with an NVIDIA Quadro K2200 GPU and CPU With 32 GB RAM and 1 TB Hard disk space. Prediction takes about 0.5 sec. per image.

5. Experiments

Evaluating the performance of the algorithm for SIS to IIS is non-trivial as there may be multiple correct sequences of images that each can visually describe a given story. For example, Fig 4 shows a story that is visually represented by two sequences of images. It is hard to tell which one of the two visual depictions is the correct one as both sequences describe the story adequately. The VIST [9] dataset has many stories where visual coherence is not explicit (in terms of common objects or scenes throughout the story). Also, since stories are short, the possibility of a visually coherent object or scene being present is generally low. Hence, we resort to evaluation with a user study for a reference of what is correct. We then try to replicate the results using a visual saliency based quantitative metric. We first describe the dataset and our baselines and then present the results.

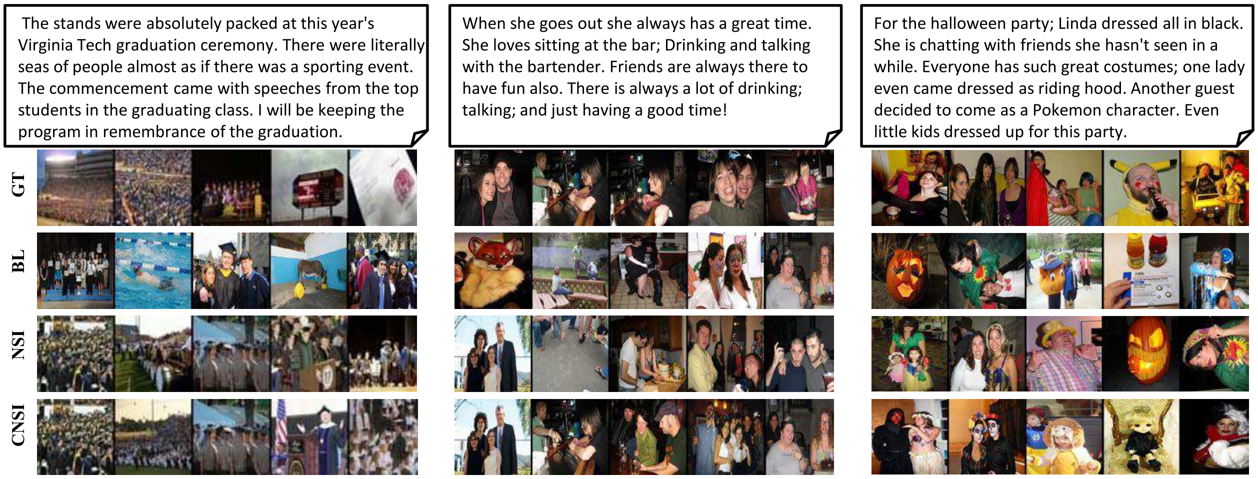
5.1. Dataset

We use VIST dataset [9] for all of our experiments. To our knowledge this is the only dataset that consists of sequences of images with sequences of text descriptions that form narrative stories. The dataset consists of approximately 40,154 stories for training, with each story made of 5 sentences and corresponding set of 5 images. The sentences are unique for each story, but the sets of images are not unique. In fact, out of the 200,770 images (40,154 x 5) only 65,145 are unique. In addition, there is a test set (5,054 stories) and validation set. Note that SIS is present for all of the 40,154 stories while DII are present only for a subset of 28,000 stories. This is the case with the test set as well. Even though the dataset contains story image sequence repeats, this shortcoming actually represents a more realistic scenario and so we use the dataset as is.

5.2. Comparative Evaluation

We compare our approach with two baseline networks to analyze the different aspects of the proposed network.

`BL`: Baseline Network. For the baseline network, we use the one-to-one description to image retrieval algorithm proposed in [27]. The network is pre-trained on MS-COCO [15] and then trained on VIST [9] dataset. Each sentence in the SIS and image in the corresponding IIS are separated from their story to get 200,770 separate sentence-image pairs. We train on this and rearrange the test set similarly for evaluation. This experiment is to study the importance of sequential modeling of the input story.

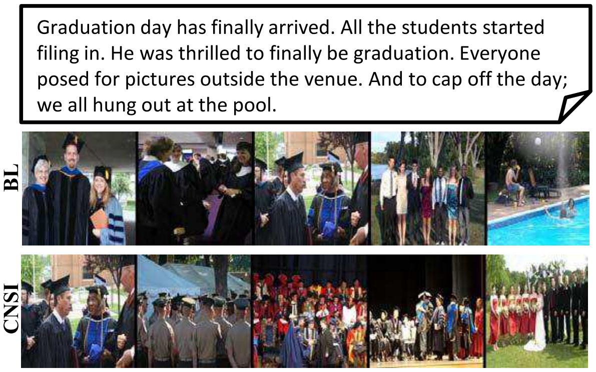


The stands were absolutely packed at this year's Virginia Tech graduation ceremony. There were literally seas of people almost as if there was a sporting event. The commencement came with speeches from the top students in the graduating class. I will be keeping the program in remembrance of the graduation.

When she goes out she always has a great time. She loves sitting at the bar; Drinking and talking with the bartender. Friends are always there to have fun also. There is always a lot of drinking; talking; and just having a good time!

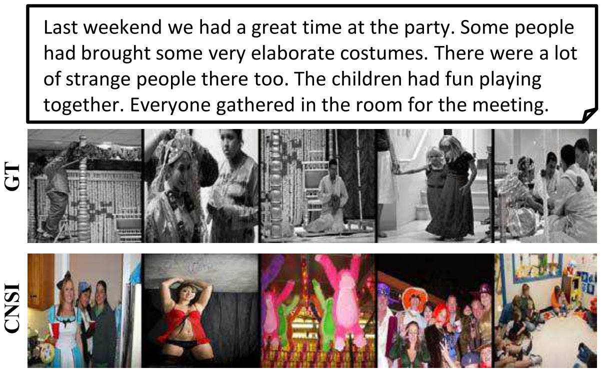
For the halloween party; Linda dressed all in black. She is chatting with friends she hasn't seen in a while. Everyone has such great costumes; one lady even came dressed as riding hood. Another guest decided to come as a Pokemon character. Even little kids dressed up for this party.

Figure 5. Samples of predicted images for three stories where the image sequences in the last row predicted by our `CNSI` model wins the most votes from AMT workers.



Graduation day has finally arrived. All the students started filing in. He was thrilled to finally be graduation. Everyone posed for pictures outside the venue. And to cap off the day; we all hung out at the pool.

Figure 6. Example story where `CNSI` failed to better illustrate the story than `BL` according to the user study.



Last weekend we had a great time at the party. Some people had brought some very elaborate costumes. There were a lot of strange people there too. The children had fun playing together. Everyone gathered in the room for the meeting.

Figure 7. The Amazon workers prefer the predicted image sequence in the bottom which shows less visual consistency than the middle one.

`NSI`: Network without Coherence. In order to evaluate the effect of coherence on the performance of the network, we also consider the proposed network without the coherence vector. Since each sequence of images in the dataset has a different amount of coherence, the role of coherence in our algorithm needs to be assesed. The training procedure is identical to our full model.

5.3. User Study

We perform evaluation with the help of AMT workers. We obtain prediction results from the network without coherence (`NSI`), network with coherence (`CNSI`), baseline (`BL`), and ground truth sequences (`GT`). Five experiments are performed: 1)`BL` vs. `GT`; 2) `NSI` vs. `BL`; 3) `CNSI` vs. `NSI`; and 4) `CNSI` vs. `GT` and 5) `BL` vs `CNSI`. We can draw a conclusion that `CNSI` is the best model among the three approaches if `CNSI` is preferred in (3), (4) and (5). Additionally, results of experiments (1) and (2) shows the performance gain of `NSI` over `BL`.

For the AMT experiment, we take 200 random stories from the test set. The test set consists of 3,335 stories with 4,980 unique images. Two image sequences corresponding to the same text obtained from `GT`, `BL`, `NSI`, or `CNSI` are presented to the user, who is asked to make a binary selection of which visual story best characterizes the text. The order of occurrence of the two representations are randomly shuffled. Each experiment has 200 stories rated by 5 workers, for a sum of 1,000 total evaluations. The number of votes where algorithm A is preferred over algorithm B for an A vs B experiment is shown in Table 1. In addition, Table 1 also lists the details of how many story visualizations generated by algorithm A are preferred by N workers, where N varies from 5 to 0. Also, we perform another experiment that compares all four visual stories simultaneously, considering the possibility that transitivity might not hold in the pairwise experiments. The results for this experiment are shown in Table 2.

BL vs. GT	#Votes	5	4	3	2	1	0
10.6% (106/894)	#Samples	1	0	5	14	58	122
CNSI vs. GT	#Votes	5	4	3	2	1	0
22.2% (222/778)	#Samples	3	3	20	37	61	76
NSI vs. BL	#Votes	5	4	3	2	1	0
67.7% (677/323)	#Samples	44	57	54	26	15	4
CNSI vs. NSI	#Votes	5	4	3	2	1	0
53.3% (533/467)	#Samples	15	41	56	47	32	9
BL vs. CNSI	#Votes	5	4	3	2	1	0
38.5% (385/615)	#Samples	3	17	40	62	59	20

Table 1. The results of pairwise preference test on story visualization of workers reviews via AMT. Comparisons are conducted in the manner of A vs. B. The numbers indicates the percentage of responses that A is a better visualization than B for a given story.

Results and Discussion: Table 1 outlines the pairwise preference results from the AMT user study. For the comparison between BL and GT, BL was preferred 106 times, in comparison to GT which was selected 894 times. Thus, the users' preference of BL over GT was 10.6%. Preference of BL over GT is 3% (6 out of 200), if 3 of the 5 users (majority) preferred the visual story obtained using BL. The user preference for NSI over BL is 67.7%. Similarly, the user preference for CNSI over NSI is 53.3%. The results indicate that the proposed model outperforms the baseline and users prefer image sequences that are coherent and consistent. Fig 5 shows example visual stories obtained from GT, BL, NSI, and CNSI, where users preferred the results generated using the proposed model (CNSI). Fig 6 shows an example scenario where users preferred BL over CNSI.

From Table 2, we can also see that when shown all the four results, users tend to prefer the result by the proposed algorithm more than even GT. Both NSI and CNSI perform reasonably well, but we think CNSI could perform significantly better if dataset was more coherent and/or if contribution of coherence is more dynamically modulated by the network (*e.g.*, through some form of attention).

Algorithm	GT	BL	NSI	CNSI
#Samples	52.5	44.5	47.5	**55.5**

Table 2. Preference of algorithm based on maximum voting of 5 workers for 200 samples. To avoid ties, if GT and CNSI get 2 votes each and BL gets one, then both GT and CNSI gets half a point.

Importance of Consistency and Coherence: In terms of the predicted images, consistency indicates visual similarity between images of a sequence while coherence is interpreted as images having common entities, such as a person or an object. In Fig 7, it is clear that images of GT (middle) show higher visual consistency and coherence than our prediction (bottom). However, our predictions were preferred by majority of workers for the visual description of the input story. The fact is that users' preference is highly related to the alignment between the images and the corresponding sentences apart from visual consistency and coherence. Ideally, consistency and coherence in the output sequence is preferable as shown by the results in Table 1 but not always. For example, a set of 5 images that lack visual coherence can still be *perceived* by a user as forming a story. This is the case for many samples in the VIST dataset. Also, in [7], authors show failure cases that result from giving order to unordered sequence of images and sentences within the same story in the VIST dataset. They observe that failure cases were due to lack of coherence in the dataset itself. This motivates explicit encoding of coherence in the input story, but not constraining the predicted images to have common objects and scenes.

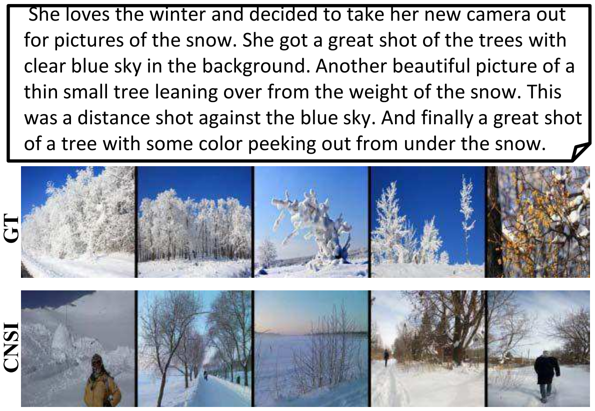

She loves the winter and decided to take her new camera out for pictures of the snow. She got a great shot of the trees with clear blue sky in the background. Another beautiful picture of a thin small tree leaning over from the weight of the snow. This was a distance shot against the blue sky. And finally a great shot of a tree with some color peeking out from under the snow.

Figure 8. A sample that our proposed method (bottom) gets less votes from AMT users than the ground truth (middle). However the two image sequences look visually similar.

5.4. Visual Saliency based Metric

We observe that defining one visually correct sequence of images is not trivial in storytelling. As shown in Fig 4 (GT and BL), multiple visual summaries can describe a story without ambiguity. Majority of workers preferred BL predictions over GT in this example. Then, evaluation boils down to checking if the predicted images describe the input story as adequately as ground truth does. The correspondence between ground truth and predicted images may be caused by the presence of common salient objects/scenes in the images. For example, both GT and BL image sequences in Fig 4 visually describe a graduation day scenario. Given that the SIS are not specific with respect to entities, the BL can be considered a correct representation with respect to GT. We propose a visual saliency based similarity metric to evaluate this kind of correctness of a predicted story.

Text Processing: We consider a test subset with DII data available. Each image has three captions associated with it. The captions are processed using Stanford core NLP [25] parser to extract noun entities. Some abnormal entities are extracted due to spelling, grammar and typograph-

	Recall @1(%)	Recall @ 2(%)	Recall @ 5(%)
BL	5	31	**54**
NSI	16	30	43.5
CNSI	**20.5**	**33.5**	50

Table 3. Visual Saliency based $Recall$@1, 2 and 5.

ical errors (*e.g.*, the word 'advertisement' had five different spellings). Even though 'autocorrect' [1] corrected most of them, the corrections were not always acceptable (*e.g.* "abike" was changed to "alike" instead of "a bike"). Others like "PyEnchant" [2] required manual verification. To automate the process of correcting thousands of words, we use autocorrect and consider only modifications for noun entities. The extracted entities generally represent objects and scenes present in the images. There is a total 13,000 unique entities over the entire set.

Visual Processing: We train a VGG-19 [26] model on ImageNet for the 20,754 categories [3] and classify the images of the story test set using this network. The top-10 most probable categories are chosen as they are mostly interchangeable. These categories were too specific compared to entities extracted from the VIST dataset (*e.g.*, image of a 'daisy flower' had 'flower' as an entity in the descriptions while the exact type of daisy was the predicted result from ImageNet). Hence we take immediate two *hypernyms* of the predicted labels using WordNet [19] for each of the 10 categories to make the visual label list. The union of the visual label list and textual entity list, made up the salient entity set that has both visually and textually grounded entities.

Evaluation Metric: We provide $Recall@k$ ($k = \{1, 2, 5\}$) for a story in the top-k predictions to have the same salient entities as the ground truth. For each sentence in the story, we retrieve the top 'k' images to get 'k' visual stories. If at least one of the 'k' stories have for each of the images, more than 'n' salient entities common with the GT, then it is positive. 'n' is experimentally chosen as 10% of the entities of GT as lower values had erroneous results and higher values had poor Recall.

Results and Discussion: Table 3 show the Recall at 1, 2 and 5 for a predicted sequence of images to be visually similar to the images in ground truth. Visual similarity can be explicitly verified in Fig 8, where an example story (bottom) predicted by CNSI was in top 1 with respect to GT. However, GT was preferred by majority of workers. Hence, we believe that user study alone or the metric alone would not suffice to measure the performance of the proposed algorithm. Even though there exists some mismatch between the results, we can see a clear pattern with respect to which models perform the best on the dataset. We can also see that, as k in Recall at k increases, performance of BL starts to increase more than that of the proposed model. This might be because, as the number of considered images increase, finding an image with same visual entities as GT become easier while finding images that adhere to the story and are also visually similar to GT becomes more difficult.

Recall	@10(%)	@ 50(%)	@ 100(%)	@ 500(%)
BL	0	0.5	0.5	3
NSI	0	**2**	4	22
CNSI	0	1.5	**4.5**	**24.5**

Table 4. Visual Saliency based Recall of GT images@10, 50, 100 and 500.

The values for Recall of retrieving GT images are shown in Table 4. we can see that the proposed algorithm performs better though GT images are not retrieved in ranking top 10 or 20. Note that GT images are not the unique visual representation of input story ([7]) and sometimes stories retrieved by our algorithm are preferred as shown in Table 2.

6. Conclusion, Limitations and Future Work

We propose a two stage network as a solution for the problem of visual illustration of natural language stories. Two networks, along with a baseline were evaluated on a comprehensive dataset using both qualitative and quantitative metrics. We observe that the proposed model performs better than the baseline and in a few cases, better than the ground truth itself, as verified by the user study. We observe that evaluation metrics for the storytelling task is ill-defined and hence propose a visual saliency based recall metric as the new measure. It is observed from evaluation that this task is non-trivial and more research is necessary to study the relationship between SIS and corresponding images.

Limitations and Future Work: Even though the proposed networks perform well, we note that there is a big space for improvement. Particularly, coming up with stronger evaluation metrics, studying the importance of coherence for visualization, and developing a more comprehensive and proper dataset. We are looking to explore these paths to define a more robust and well defined solution to the problem of visual storytelling.

Acknowledgement

Hareesh Ravi, Carlos M. Muniz, and Mubbasir Kapadia were funded, in part, by DARPA SocialSim-W911NF-17-C-0098. Mubbasir Kapadia was funded in part by NSF IIS-1703883 and NSF S & AS-1723869. D. Metaxas has been partially funded by NSF, AFOSR DDDAS program and an ARO-MURI.

[1] https://github.com/phatpiglet/autocorrect/

[2] https://pythonhosted.org/pyenchant/

References

[1] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR 2015*, pages 2422–2431, June 2015. 2

[2] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. *Empirical evaluation of gated recurrent neural networks on sequence modeling*. 2014. 1, 3

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 8

[4] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In *ICCV*, 2017. 2

[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollr, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015. 2

[6] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV 2014*. 2

[7] D. B. D. P. H. Agrawal, A. Chandrasekaran and M. Bansal. Sort story: Sorting jumbled images and captions into stories. In *EMNLP, 2016*. 1, 7, 8

[8] X. He, B. Shi, X. Bai, G.-S. Xia, Z. Zhang, and W. Dong. Image caption generation with part of speech guidance. *Pattern Recognition Letters*, 2017. 2

[9] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, J. Devlin, A. Agrawal, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *NAACL*, 2016. 1, 3, 5

[10] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1, 2

[11] G. Kim, S. Moon, and L. Sigal. Joint photo stream and blog post summarization and exploration. In *CVPR 2015*, pages 3081–3089, June 2015. 3

[12] G. Kim, S. Moon, and L. Sigal. Ranking and retrieval of image sequences from multiple paragraph queries. In *CVPR 2015*, pages 1993–2001, June 2015. 1, 3

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 2015. 5

[14] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 2017. 1

[15] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *ECCV 2014: 740-755*. 3, 4, 5

[16] L. LIU and M. T. ÖZSU, editors. *Mean Average Precision*, pages 1703–1703. Springer US, Boston, MA, 2009. 2

[17] E. Mansimov, E. Parisotto, J. Ba, and R. Salakhutdinov. Generating images from captions with attention. In *ICLR*, 2016. 2

[18] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR 2016*, pages 11–20, June 2016. 2

[19] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995. 8

[20] C. C. Park and G. Kim. Expressing an image stream with a sequence of natural sentences. In *NIPS 2015*, pages 73–81, 2015. 3

[21] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV 2015*, pages 2641–2649, Dec 2015. 2

[22] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *ICML*, 2016. 1, 3

[23] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. Joint image-text representation by gaussian visual semantic embedding. In *MM*, 2016. 2

[24] Z. Ren, X. Wang, N. Zhang, X. Lv, and L. J. Li. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR 2017*, pages 1151–1159, July 2017. 2

[25] S. Schuster and C. D. Manning. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *LREC*, 2016. 7

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3, 8

[27] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. *ICLR 2016*. 1, 2, 3, 4, 5

[28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR 2015*, pages 3156–3164, June 2015. 2

[29] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR 2016*, pages 5005–5013, June 2016. 1, 2

[30] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng. Structured matching for phrase localization. In *ECCV 2016*, pages 696–711, 2016. 2

[31] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. In *CVPR 2017*, pages 7378–7387, July 2017. 2

[32] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR 2016*, pages 203–212, June 2016. 2

[33] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *ICML 2015, pp. 2048-2057)*. 2

[35] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR 2016*, pages 4651–4659, June 2016. 2

[36] L. Yu, M. Bansal, and T. L. Berg. Hierarchically-attentive RNN for album summarization and storytelling. In *EMNLP 2017*, pages 977–982, 2017. 3