Recurrent Topic-Transition GAN for Visual Paragraph Generation

Xiaodan Liang[1] Zhiting Hu[1,2] Hao Zhang[1,2] Chuang Gan[3] Eric P. Xing[2]
[1]Carnegie Mellon University [2]Petuum Inc. [3] Tsinghua University
{xiaodan1,zhitingh,hao,epxing}@cs.cmu.edu, ganchuang1990@gmail.com

Abstract

A natural image usually conveys rich semantic content and can be viewed from different angles. Existing image description methods are largely restricted by small sets of biased visual paragraph annotations, and fail to cover rich underlying semantics. In this paper, we investigate a semi-supervised paragraph generative framework that is able to synthesize diverse and semantically coherent paragraph descriptions by reasoning over local semantic regions and exploiting linguistic knowledge. The proposed Recurrent Topic-Transition Generative Adversarial Network (RTT-GAN) builds an adversarial framework between a structured paragraph generator and multi-level paragraph discriminators. The paragraph generator generates sentences recurrently by incorporating region-based visual and language attention mechanisms at each step. The quality of generated paragraph sentences is assessed by multi-level adversarial discriminators from two aspects, namely, plausibility at sentence level and topic-transition coherence at paragraph level. The joint adversarial training of RTT-GAN drives the model to generate realistic paragraphs with smooth logical transition between sentence topics. Extensive quantitative experiments on image and video paragraph datasets demonstrate the effectiveness of our RTT-GAN in both supervised and semi-supervised settings. Qualitative results on telling diverse stories for an image verify the interpretability of RTT-GAN.

1. Introduction

Describing visual content with natural language is an emerging interdisciplinary problem at the intersection of computer vision, natural language processing, and artificial intelligence. Recently, great advances [18, 3, 4, 31, 24, 33, 22, 21, 36] have been achieved in describing images and videos using a single high-level sentence, owing to the advent of large datasets [23, 17] pairing images with natural language descriptions. Despite the encouraging progress in image captioning [30, 33, 24, 31], most current systems tend to capture the scene-level gist rather than fine-grained entities, which largely undermines their applications in real-world scenarios such as blind navigation, video retrieval, and automatic video subtitling. One of the recent alternatives to sentence-level captioning is visual paragraph generation [11, 16, 34], which aims to provide a coherent and detailed description, like telling stories for images/videos.

a) Generic description:

A group of people are sitting around a living room together. One of the men is wearing black sleeve shirt and blue pants. A man is sitting next to the wooden table. A man and woman are sitting on a couch. There is a brown wooden table in the room.

b) Personalized descriptions:

There is a man sitting on a wooden chair. The man with a white remote with white buttons is wearing a black and white shirt and jean pants. A woman next to him has red shirts and red skirts. There are a man and woman sitting on the floor next to a wooden table.

A smiling woman is sitting on a couch. She has yellow short hair and is wearing a short sleeve shirt. She is holding a white plate. There is a brown couch in the living room. In front of her is a wooden table. There are papers and glasses on the table.

Figure 1. Our RTT-GAN is able to automatically produce generic paragraph descriptions shown in (a), and personalized descriptions by manipulating first sentences (highlighted in red), shown in (b).

Generating a full paragraph description for an image/video is challenging. First, paragraph descriptions tend to be diverse, just like different individuals can tell stories from personalized perspectives. As illustrated in Figure 1, users may describe the image starting from different viewpoints and objects. Existing methods [16, 34, 19] deterministically optimizing over single annotated paragraph thus suffer from losing massive information expressed in the image. It is desirable to enable diverse generation through simple manipulations. Second, annotating images/videos with long paragraphs is labor-expensive, leading to only small scale image-paragraph pairs which limits the model generalization. Finally, different from single-sentence captioning, visual paragraphing requires to capture more detailed and richer semantic content. It is necessary to perform long-term visual and language reasoning to incorporate fine-grained cues while ensuring coherent paragraphs.

To overcome the above challenges, we propose a semi-supervised visual paragraph generative model, Recurrent Topic-Transition GAN (RTT-GAN), which generates diverse and semantically coherent paragraphs by reasoning over both local semantic regions and global paragraph

context. Inspired by Generative Adversarial Networks (GANs) [7], we establish an adversarial training mechanism between a structured paragraph generator and multi-level paragraph discriminators, where the discriminators learn to distinguish between real and synthesized paragraphs while the generator aims to fool the discriminators by generating diverse and realistic paragraphs.

The paragraph generator is built upon dense semantic regions of the image, and selectively attends over the regional content details to construct meaningful and coherent paragraphs. To enable long-term visual and language reasoning spanning multiple sentences, the generator recurrently maintains context states of different granularities, ranging from paragraph to sentences and words. Conditioned on current state, a spatial visual attention mechanism selectively incorporates visual cues of local semantic regions to manifest a topic vector for next sentence, and a language attention mechanism incorporates linguistic information of regional phrases to generate precise text descriptions. We pair the generator with rival discriminators which assess synthesized paragraphs in terms of plausibility at sentence level as well as topic-transition coherence at paragraph level. Our model allows diverse descriptions from a single image by manipulating the first sentence which guides the topic of the whole paragraph. Semi-supervised learning is enabled in the sense that only *single-sentence* caption annotation is required for model training, while the linguistic knowledge for constructing long paragraphs is transfered from standalone text paragraphs without paired images.

We compare RTT-GAN with state-of-the-art methods on both image-paragraph and video-paragraph datasets, and verify the superiority of our method in both supervised and semi-supervised settings. Using only the single-sentence COCO captioning dataset, our model generates highly plausible multi-sentence paragraphs. Given these synthesized paragraphs for COCO image, we can considerably enlarge the existing small paragraph datasets to further improve the paragraph generation capability of our RTT-GAN. Qualitative results on personalized paragraph generation also shows the flexibility and applicability of our model.

2. Related Work

Visual Captioning. Image captioning is posed as a longstanding and holy-grail goal in computer vision, targeting at bridging visual and linguistic domain. Early works that posed the problem as a ranking and template retrieval task [5] performed poorly as it is hard to enumerate all possibilities in one collected dataset due to the compositional nature of language. Therefore, recent works [18, 3, 4, 31, 24, 33, 20] focus on directly generating captions by modeling the semantic mapping from visual cues to language descriptions. These approaches are typically based on deep generative models [10], among which training recurrent network language models conditioned on image features [3, 4, 31, 6] achieves great success by taking advantages of large-scale image captioning datasets. Similar success has been already seen in video captioning fields [4, 32]. Though generating high-level sentences for images is encouraging, massive underlying information, such as relationships between objects, attributes, and entangled geometric structures conveyed in the image, would be missed if only summarizing them with a coarse sentence.

Visual Paragraph Generation. Paragraph generation overcomes shortcomings of standard captioning and dense captioning by producing a coherent and fine-grained natural language description. To reason about long-term linguistic structures with multiple sentences, hierarchical recurrent network [19, 34, 16] has been widely used to directly simulate the hierarchy of language. For example, Krause et al. [16] combine semantics of all regions of interest to produce a generic paragraph for an image. However, all these methods suffer from the overfitting problem due to the lack of sufficient paragraph descriptions. In contrast, we propose a generative model to automatically synthesize a large amount of diverse and reasonable paragraph descriptions by learning the implicit linguistic interplay between sentences. Our RTT-GAN has better interpretability by imposing the sentence plausibility and topic-transition coherence on the generator with two adversarial discriminators. The generator selectively incorporates visual and language cues of semantic regions to produce each sentence.

3. Recurrent Topic-Transition GAN

The proposed Recurrent Topic-Transition GAN (RTT-GAN) aims to describe the rich content of a given image/video by generating a natural language paragraph. Figure 2 provides an overview of the framework. Given an input image, we first detect a set of semantic regions using dense captioning method [13]. Each semantic region is represented with a visual feature vector and a short text phrase (e.g. *person riding a horse*). The paragraph generator then sequentially generates meaningful sentences by incorporating the fine-grained visual and textual cues in a selective way. To ensure high-quality individual sentences and coherent whole paragraph, we apply a sentence discriminator and a topic-transition discriminator on each generated sentence, respectively, to measure the plausibility and smoothness of semantic transition with preceding sentences. The generator and multi-level discriminators are learned jointly within an adversarial framework. RTT-GAN supports not only supervised setting with annotated image-paragraph pairs, but also semi-supervised setting where only a single sentence caption is provided for each image and the knowledge of long paragraph construction is learned from a standalone paragraph corpus.

In next sections, we first derive the adversarial frame-

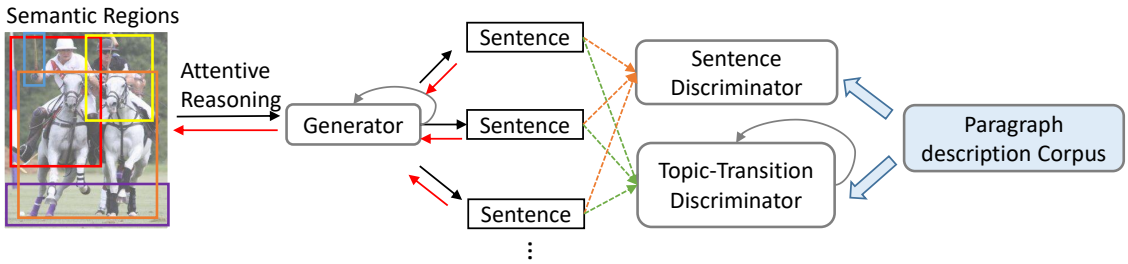

Figure 2. Our RTT-GAN alternatively optimizes a structured paragraph generator and two discriminators following an adversarial training scheme. The generator recurrently produces each sentence by reasoning about local semantic regions and preceding paragraph state. Each synthesized sentence is then fed into a sentence discriminator and a recurrent topic-transition discriminator for assessing sentence plausibility and topic coherence, respectively. A paragraph description corpus is adopted to provide linguistic knowledge about paragraph generation, which depicts the true data distribution of the discriminators .

work of our RTT-GAN, then describe detailed model design of the paragraph generator and the multi-level discriminators, respectively.

3.1. Adversarial Objective

We construct an adversarial game between the generator and discriminators to drive the model learning. Specifically, the sentence and topic-transition discriminators learn a critic between real and generated samples, while the generator attempts to confuse the discriminators by generating realistic paragraphs that satisfy linguistic characteristics (i.e., sentence plausibility and topic-transition coherence). The generative neural architecture ensures the paragraph captures adequate semantic content of the image, which we describe in detail in the next sections. Formally, let G denote the paragraph generator, and let D^s and D^r denote the sentence and topic-transition discriminators, respectively.

At each time step t, conditioned on preceding sentences $\mathbf{s}_{1:t-1}$ and local semantic regions $\mathbf{V}$ of the image, the generator G recurrently produces a single sentence $\mathbf{s}_t$, where each sentence $\mathbf{s}_t = \{\mathbf{w}_{t,i}\}$ consists of a sequence of $\mathbf{N}_t$ words $\mathbf{w}_{t,i}$ $(i = 1, \ldots, \mathbf{N}_t)$:

$$P_G(\mathbf{s}_t|\mathbf{s}_{1:t-1}, \mathbf{V}) = \prod_{i=1}^{\mathbf{N}_t} P_G(\mathbf{w}_{t,i}|\mathbf{w}_{t,1:i-1}, \mathbf{s}_{1:t-1}, \mathbf{V}). \quad (1)$$

The discriminators learn to differentiate real sentences $\hat{\mathbf{s}}$ within a true paragraph $\hat{\mathbf{P}}$ from the synthesized ones $\mathbf{s}_t$. The generator G tries to generate realistic visual paragraph by minimizing against the discriminators' chance of correctly telling apart the sample source. As the original GAN [7] that optimizes over binary probability distance suffers from mode collapse and instable convergence, we follow the new Wasserstein GAN [1] method that theoretically remedies this by minimizing an approximated Wasserstein distance. The objective of the adversarial framework is written as:

$$\begin{aligned} &\min_G \max_{D^s, D^r} \mathcal{L}(G, D^s, D^r) = \\ &\mathbb{E}_{\hat{\mathbf{s}} \sim p_{\text{data}(\hat{\mathbf{s}})}} \left[D^s(\hat{\mathbf{s}})\right] - \mathbb{E}_{\mathbf{s}_{1:t} \sim p_{G(\mathbf{s}_{1:t}|\mathbf{V})}} \left[D^s(\mathbf{s}_t)\right] + \\ &\mathbb{E}_{\hat{\mathbf{P}} \sim p_{\text{data}(\hat{\mathbf{P}})}} \left[D^r(\hat{\mathbf{P}})\right] - \mathbb{E}_{\mathbf{s}_{1:t} \sim p_{G(\mathbf{s}_{1:t}|\mathbf{V})}} \left[D^r(\mathbf{s}_{1:t})\right], \end{aligned} \quad (2)$$

where $p_{\text{data}(\hat{\mathbf{s}})}$ and $p_{\text{data}(\hat{\mathbf{P}})}$ denote the true data distributions of sentences and paragraphs, respectively, which are empirically constructed from a paragraph description corpus. The second line of the equation is the objective of the sentence discriminator D^s that optimizes a critic between real/fake sentences, while the third line is the objective of the topic-transition discriminator D^r. Here $p_{G(\mathbf{s}_{1:t}|\mathbf{V})}$ indicates the distribution of generated sentences by the generator G.

To leverage existing image-paragraph pair dataset in the supervised setting, or image captioning dataset in the semi-supervised setting, we also incorporate the traditional word reconstruction loss for generator optimization, which is defined as:

$$\mathcal{L}^c(G) = -\sum_{t=1}^{T} \sum_{i=1}^{\mathbf{N}_t} \log P_G(\mathbf{w}_{t,i}|\mathbf{w}_{t,1:i-1}, \mathbf{s}_{1:t-1}, \mathbf{V}). \quad (3)$$

Note that the reconstruction loss is only used for supervised examples with paragraph annotations, and semi-supervised examples with single-sentence caption (where we set $T = 1$). Combining Eqs.(2)-(3), the joint objective for the generator G is thus:

$$G^* = \arg\min_G \max_{D^s, D^r} \lambda \mathcal{L}(G, D^s, D^r) + \mathcal{L}^c(G), \quad (4)$$

where λ is the balancing parameter fixed to 0.001 in our implementation. The optimization of the generator and discriminators (i.e., Eq.(4) and Eq.(2), respectively) is performed in an alternating min-max manner. We describe the training details in section 3.4.

The discrete nature of text samples hinders gradient back-propagation from the discriminators to the generator [9]. We address this issue following SeqGAN [35]. The state is the current produced words and sentences, and the action is the next word to select. we apply Monte Carlo search with a roll-out policy to sample the remaining words until it sees an END token for each sentence and maximal number of sentences. The roll-out policy is the same with the generator, elaborated in Section 3.2. The discriminator is trained by providing true paragraphs from the text corpus

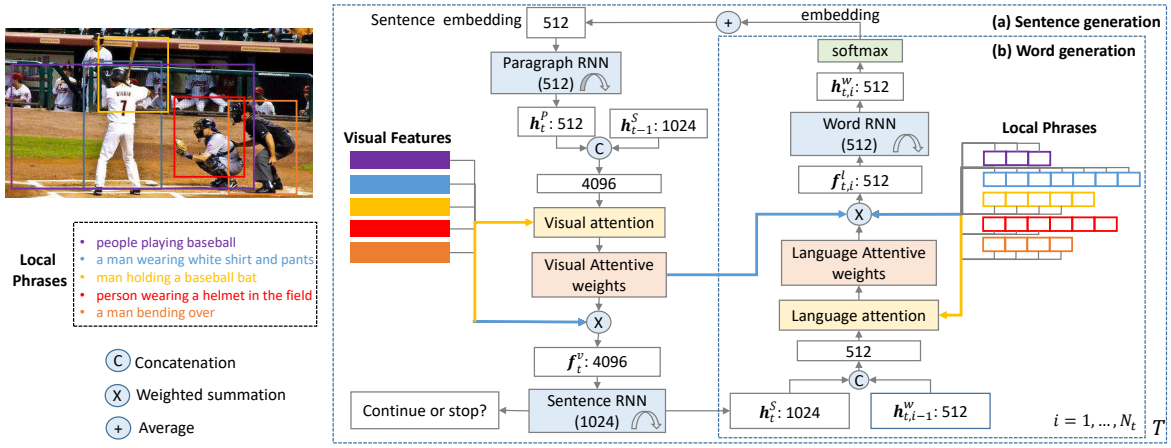

Figure 3. Illustration of our paragraph generator. Given visual features and local phrases of semantic regions, the paragraph generator is performed for most T steps to sequentially generate each sentence. At t-th step, the paragraph states $\mathbf{h}_t^P$ is first updated with the embedding of preceding sentences by *paragraph RNN*. Then, the visual attention takes features of semantic regions, current paragraph states $\mathbf{h}_t^P$ and previous hidden states $\mathbf{h}_{t-1}^S$ as input to manifest a visual context vector $\mathbf{f}_t^v$. $\mathbf{f}_t^v$ is then fed into *sentence RNN* to obtain the encoded topic vector $\mathbf{h}_t^S$ and determine whether to generate next sentence. The *word RNN* with language attention then generates each word.

and synthetic ones from the generator. The generator is updated by employing a policy gradient based on the expected reward received from the discriminator and the reconstruction loss for fully-supervised and semi-supervised settings, defined in Eq. 4. To reduce the variance of the action values, we run the roll-out policy starting from current state till the end of the paragraph for five times to get a batch of output samples. The signals that come from the word prediction for labeled sentences (defined in Eq. 3)) can be regarded as the intermediate reward. The gradients are passed back to the intermediate action value via Monte Carlo search [35].

3.2. Paragraph Generator

Figure 3 shows the architecture of the generator G, which recurrently retains different levels of context states with a hierarchy constructed by a paragraph RNN, a sentence RNN, and a word RNN, and two attention modules. First, the paragraph RNN encodes the current paragraph state based on all preceding sentences. Second, the spatial visual attention module selectively focuses on semantic regions with the guidance of current paragraph state to produce the visual representation of the sentence. The sentence RNN is thus able to encode a topic vector for the new sentence. Third, the language attention module learns to incorporate linguistic knowledge embedded in local phrases of focused semantic regions to facilitate word generation by the word RNN.

Region Representation. Given an input image, we adopt the dense captioning model [13, 16] to detect semantic regions of the image and generate their local phrases. Each region $\mathbf{R}_j$ $(j \in 1, \ldots, M)$ has a visual feature vector $\mathbf{v}_j$ and a local text phrase (i.e., region captioning) $\mathbf{s}_j^r = \{\mathbf{w}_{j,i}^r\}$ consisting of $\mathbf{N}_j$ words. In practice, we use the top $M = 50$ regions.

Paragraph RNN. The paragraph RNN keeps track of the paragraph state by summarizing preceding sentences. At each t-th step ($t = 1, \ldots, T$), the paragraph RNN takes the embedding of generated sentence in previous step as input, and in turn produces the paragraph hidden state $\mathbf{h}_t^P$. The sentence embedding is obtained by simply averaging over the embedding vectors of the words in the sentence. This strategy enables our model to support the manipulation of the first sentence to initialize the paragraph RNN and generate personalized follow-up descriptions.

Sentence RNN with Spatial Visual Attention. The visual attentive sentence RNN controls the topic of the next sentence $\mathbf{s}_t$ by selectively focusing on relevant regions of the image. Specifically, given the paragraph states $\mathbf{h}_t^P$ from the paragraph RNN and previous hidden states $\mathbf{h}_{t-1}^S$ of the sentence RNN, we apply an attention mechanism on the visual features $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$ of all semantic regions, and construct a visual context vector $\mathbf{f}_t^v$ that represents the next sentence at t-th step:

$$\begin{aligned}\mathbf{f}_t^v &= \text{att}^v(\mathbf{V}, \mathbf{h}_t^P, \mathbf{h}_{t-1}^S) \\ &= \sum_{j=1}^{M} \frac{\alpha(\mathbf{v}_j, \beta(\mathbf{h}_t^P, \mathbf{h}_{t-1}^S))}{\sum_{j'=1}^{M} \alpha(\mathbf{v}_{j'}, \beta(\mathbf{h}_t^P, \mathbf{h}_{t-1}^S))} \mathbf{v}_j \\ &:= \sum_{j=1}^{M} a_j \mathbf{v}_j,\end{aligned} \tag{5}$$

where $\beta(\mathbf{h}_t^P, \mathbf{h}_{t-1}^S)$ is a linear layer that transforms the concatenation of $\mathbf{h}_t^P$ and $\mathbf{h}_{t-1}^S$ into a compact vector with the same dimension as $\mathbf{v}_j$; the function $\alpha(\cdot)$ is to compute the weight of each region and is implemented with a single linear layer. For notational simplicity, we use a_j to denote the

Table 1. The performance comparisons with four state-of-the-arts and the variants of our RTT-GAN on paragraph generation in terms of six language metrics. The human performance is included for providing a better understanding of all metrics.

Method	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Sentence-Concat	12.05	6.82	31.11	15.10	7.56	3.98
Template	14.31	12.15	37.47	21.02	12.03	7.38
Image-Flat [14]	12.82	11.06	34.04	19.95	12.20	7.71
Regions-Hierarchical [16]	15.95	13.52	41.90	24.11	14.23	8.69
RTT-GAN (Semi- w/o discriminator)	12.35	8.96	33.82	17.40	9.01	5.88
RTT-GAN (Semi- w/o sentence D)	11.22	10.04	35.29	19.13	11.55	6.02
RTT-GAN (Semi- w/o topic-transition D)	12.68	12.77	37.20	20.51	12.08	6.91
RTT-GAN (Semi-)	14.08	13.07	39.22	22.50	13.34	7.75
RTT-GAN (Fully- w/o discriminator)	16.57	15.07	41.86	24.33	14.56	8.99
RTT-GAN (Fully-)	17.12	16.87	41.99	24.86	14.89	9.03
RTT-GAN (Semi + Fully)	**18.39**	**20.36**	**42.06**	**25.35**	**14.92**	**9.21**
Human	19.22	28.55	42.88	25.68	15.55	9.66

normalized attentive weight of each region $\mathbf{R}_j$.

Given the visual representation $\mathbf{f}_t^v$, the sentence RNN is responsible for determining the number of sentences that should be in the paragraph and producing a topic vector of each sentence. Specifically, each hidden state $\mathbf{h}_t^S$ is first passed into a linear layer to produce a probability over the two states {CONTINUE=0, STOP=1} which determine whether t-th sentence is the last sentence. The updated $\mathbf{h}_t^S$ is treated as the topic vector of the sentence.

Word RNN with Language Attention. To generate meaningful paragraphs relevant to the image, the model is desired to recognize and describe substantial details such as objects, attributes, and relationships. The text phrases of semantic regions that express such local semantics are leveraged by a language attention module to help with the recurrent word generation. For example, the word RNN can conveniently copy precise concepts (e.g., baseball, helmet) from the local phrases. Following the copy mechanism [8] firstly proposed in natural language processing, we selectively incorporate the embeddings of local phrases based on the topic vector $\mathbf{h}_t^S$ and preceding word state $\mathbf{h}_{t,i-1}^w, i \in \{1, \ldots, \mathbf{N}_t\}$ by the word RNN to generate the next word representation $\mathbf{f}_{t,i}^l$. Since each local phrase $\mathbf{s}_j^r$ semantically relates to respective visual feature $\mathbf{v}_j$, we thus reuse the visual attentive weights $\{a_j\}_{j=1}^M$ to enhance the language attention. Representing each word with an embedding vector $\mathbf{w}_{i,j}^r$, the language representation $\mathbf{f}_{t,i}^l$ for each word prediction at i-th step is formulated as

$$\begin{aligned}\mathbf{f}_{t,i}^l &= \text{att}^l(\mathbf{S}^r, \mathbf{h}_t^S, \mathbf{h}_{t,i-1}^w) \\ &= \sum_{j=1}^{M}\sum_{i'=1}^{\mathbf{N}_j} \frac{\alpha(\mathbf{w}_{i',j}^r, \beta(\mathbf{h}_t^S, \mathbf{h}_{t,i-1}^w))}{\sum_{j'=1}^{M}\sum_{i''=1}^{\mathbf{N}_{j'}} \alpha(\mathbf{w}_{i'',j'}^r, \beta(\mathbf{h}_t^S, \mathbf{h}_{t,i''-1}^w))} a_j \mathbf{w}_{i',j}^r.\end{aligned} \tag{6}$$

Given the language representation $\mathbf{f}_{t,i}^l$ as the input at i-th step, the word RNN computes a hidden states $\mathbf{h}_{t,i}^w$ which is then used to predict a distribution over the words in the vocabulary. After obtaining all words of each sentence, these sentences are finally concatenated to form the generated paragraph.

3.3. Paragraph Discriminators

The paragraph discriminators $\{D^s, D^r\}$ aim to distinguish between real paragraphs and synthesized ones based on the linguistic characteristics of a natural paragraph description. In particular, the sentence discriminator D^s evaluates the plausibility of individual sentences, while the topic-transition discriminator D^r evaluates the topic coherence of all sentences generated so far. With such multi-level assessment, the model is able to generate long yet realistic descriptions. Specifically, the sentence discriminator D^s is an LSTM RNN that recurrently takes each word embedding within a sentence as the input, and produces a real-value plausibility score of the synthesized sentence. The topic-transition discriminator D^r is another LSTM RNN which recurrently takes the sentence embeddings of all preceding sentences as inputs and computes the topic smoothness value of the current constructed paragraph description at each recurrent step.

3.4. Implementation Details

The discriminators D^s and D^r are both implemented as a single-layer LSTM with hidden dimension of 512. For the generator, the paragraph RNN is a single-layer LSTM with hidden size of 512 and the initial hidden and memory cells set to zero. Similarly, the sentence RNN and word RNN are single-layer LSTMs with hidden dimension of 1024 and 512, respectively. Each input word is encoded as a embedding vector of 512 dimension. The visual feature vector $\mathbf{v}_j$ of each semantic region has dimension of 4096.

Table 2. Ablation studies on the effectiveness of key components in the region-based attention mechanism of our RTT-GAN.

Method	METEOR	CIDEr
RTT-GAN (Fully- w/o phrase att)	16.08	15.13
RTT-GAN (Fully- w/o att)	15.63	14.47
RTT-GAN (Fully- 10 regions)	14.13	13.26
RTT-GAN (Fully- 20 regions)	16.92	16.15
RTT-GAN (Fully-)	**17.12**	**16.87**

The adversarial framework is trained following the Wasserstein GAN (WGAN) [1] in which we alternate between the optimization of $\{D^s, D^r\}$ with Eq.(2) and the optimization of G with Eq.(4). In particular, we perform one gradient descent step on G every time after 5 gradient steps on $\{D^s, D^r\}$. We use minibatch SGD and apply the RMSprop solver [28] with the initial learning rate set to 0.0001. For stable training, we apply batch normalization [12] and set the batch size to 1 (i.e., "instance normalization"). In order to make the parameters of D^s and D^r lie in a compact space, we clamp the weights to a fixed box $[-0.01, 0.01]$ after each gradient update. In the semi-supervised setting where only single-sentence captioning for images and standalone paragraph corpus are available, we set the maximal number of sentences in the generated paragraph to 6 for all images. In the fully-supervised setting, the groundtruth sentence number in each visual paragraph is used to train the sentence-RNN for learning how many sentences are needed. We train the models to converge for 40 epochs. The implementations are based on the public Torch7 platform on a single NVIDIA GeForce GTX 1080.

4. Experiments

4.1. Experimental Settings

To generate a paragraph for an image, we run the paragraph generator forward until the STOP sentence state is predicted or after $S_{\max} = 6$ sentences, whichever comes first. The word RNN is recurrently forwarded to sample the most likely word at each time step, and stops after choosing the STOP token or after $N_{\max} = 30$ words. We use beam search with beam size 2 for generating paragraph descriptions. Training details are presented in Section 3.4, and all models are implemented in Torch platform. In terms of the fully-supervised setting, to make a fair comparison with the state-of-the-art methods [14, 16], the experiments are conducted on the public image paragraph dataset [16], where 14,575 image-paragraph pairs are used for training, 2,487 for validation and 2,489 for testing. In terms of semi-supervised setting, our RTT-GAN is trained with the single sentence annotations provided in MSCOCO image captioning dataset [2] which contains 123,000 images. The image-paragraph validation set is used for validating the semi-supervised paragraph generation. The paragraph generation performance is also evaluated on 2,489 paragraph testing samples. For both fully-supervised and semi-supervised settings, we use the word vocabulary of image-paragraph dataset as [16] does and the 14,575 paragraph descriptions on public image paragraph dataset [16] are adopted as the standalone paragraph corpus for training discriminators. We report six widely used automatic evaluation metrics, BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and CIDEr. The model checkpoint selection is based on the best combined METEOR and CIDEr score on the validation set. Table 1 reports the performance of all baselines and our models.

Figure 4. Visualization of our region-based attention mechanism. For each sentence generation, RTT-GAN selectively focuses on semantic regions of interest in the spatial visual attention, and attentively leverage the word embeddings of their local phrases to enhance the word prediction. In the top row, we illustrate the regions with highest attention confidences during the spatial visual attention and its corresponding words (highlighted in red) with highest attention confidences during the language attention in each step.

4.2. Comparison with the State-of-the-arts

We obtain the results of all four baselines from [16]. Specifically, *Sentence-Concat* samples and concatenates five sentence captions from the model trained on MS COCO captions, in which the first sentence uses beam search and the rest are samples. *Image-Flat* [14] directly decodes an image into a paragraph token by token. *Template* predicts the text via a handful of manually specified templates. And *Region-Hierarchical* [16] uses a hierarchical recurrent neural network to decompose the paragraphs into the corresponding sentences. Same with all baselines, we adopt VGG-16 net [27] to encode the visual representation of an image. Note that our RTT-GAN and *Region-Hierarchical* [16] use the same dense captioning model [13] to extract semantic regions. *Human* shows the results by collecting an additional paragraph for 500 randomly chosen images as [16]. As expected, humans produce superior descriptions over any automatic method and the large gaps on CIDEr and METEOR verify that CIDEr and METEOR metrics align better with human judgment than BLEU scores.

Fully-supervised Setting. We can see that our *RTT-GAN (Fully-)* model significantly outperforms all base-

lines on all metrics; particularly, 3.35% over *Region-Hierarchical* and 5.81% over *Image-Flat* in terms of CIDEr. It clearly demonstrates the effectiveness of our region-based attention mechanism that selectively incorporate visual and language cues, and the adversarial multi-level discriminators that provide a better holistic, semantic regularization of the generated sentences in a paragraph. The inferiority of *Image-Flat* compared to the hierarchical networks of *RTT-GAN* and *Region-Hierarchical* demonstrates the advantage of performing hierarchical sentence predictions for a long paragraph description.

Semi-supervised Setting. The main advantage of our RTT-GAN compared to prior works is the capability of generating realistic paragraph descriptions coordinating with the natural linguistic properties, given only singe sentence annotations. It is demonstrated by the effectiveness of our semi-supervised model *RTT-GAN (Semi-)* that only uses the single sentence annotations of MSCOCO captions, and imposes the linguistic characteristics on the rest sentence predictions using adversarial discriminators that are trained with the standalone paragraph corpus. Specifically, *RTT-GAN (Semi-)* achieves comparable performance with the fully-supervised *Regions-Hierarchical* without using any groundtruth image-paragraph pairs. After augmenting the image paragraph dataset with the synthesized paragraph descriptions by *RTT-GAN (Semi-)*, *RTT-GAN (Semi+Fully)* dramatically outperforms *RTT-GAN (Fully-)* and other baselines, e.g. 6.84% increase over *Regions-Hierarchical* on CIDEr. We also show some qualitative results of generated paragraphs by our *RTT-GAN (Semi-)* in Figure 5. These promising results further verify the capability of our RTT-GAN on reasoning a long description that has coherent topics and plausible sentences without the presence of ground-truth image paragraph pairs.

4.3. The Importance of Adversarial Training

After eliminating the discriminators during the model optimization in both fully- and semi-supervised settings (i.e. *RTT-GAN (Fully- w/o discriminator)* and *RTT-GAN (Semi- w/o discriminator)*), we observe consistent performance drops on all metrics compared to the full models, i.e. 1.80% and 4.11% on CIDEr, respectively. *RTT-GAN (Semi- w/o discriminator)* can be regarded as a image captioning model due to the lack of adversarial loss, similar to *Sentence-Concat*. It justifies that the sentence plausibility and topic coherences with preceding sentences are very critical for generating long, convincing stories. Moreover, the pure word prediction loss largely hinders the model's extension to unsupervised or semi-supervised generative modeling. Training adversarial discriminators that explicitly enforce the linguistic characteristics of a good description can effectively impose high-level and semantic constraints on sentence predictions by the generator.

Table 3. Human voting results for the plausibility of generated personalized paragraphs by the variants of our RTT-GAN.

Semi- w/o discriminator	Semi-	Semi + Fully
12.6%	40.5 %	46.9%

Furthermore, we break down our design of discriminators in order to compare the effect of the sentence discriminator and recurrent topic-transition discriminator, as *RTT-GAN (Semi- w/o sentence D)* and *RTT-GAN (Semi- w/o topic-transition D)*, respectively. It can be observed that although both discriminators help bring the significant improvement, the sentence discriminator seems to play a more critical role by addressing the plausibility of each sentence.

4.4. The Importance of Region-based Attention

We also evaluate the effectiveness of the spatial visual attention and language attention mechanisms to facilitate the paragraph prediction, as reported in Table 2. *RTT-GAN (Fully- w/o att)* directly pools the visual features of all regions into a compact representation for sequential sentence prediction, like *Region-Hierarchical*. *RTT-GAN (Fully- w/o phrase att)* represents the variant that removes the language attention module. It can be observed that the attention mechanism effectively facilitates the prediction of RTT-GAN by selectively incorporating appropriate visual and language cues. Particularly, the advantages of explicitly leveraging words from local phrases suggest that transferring visual-language knowledges from more fundamental tasks (e.g. detection) is beneficial for generating high-level and holistic descriptions.

As an exploratory experiment, we investigate generating paragraphs from a smaller number of regions (10 and 20) than 50 used in previous models, denoted as *RTT-GAN (Fully- 10 regions)* and *RTT-GAN (Fully- 20 regions)*. Although these results are worse than our full model, the performance of using only top 10 regions is still reasonably good. Figure 4 gives some visualization results of our region-based attention mechanism. For generating the sentence at each step, our model selectively focuses on distinct regions and their distinct corresponding words in local phrases to facilitate the sentence prediction.

4.5. Personalized Paragraph Generation

Different from prior works, our model supports the personalized paragraph generation which produces diverse descriptions by manipulating first sentences. It can be conveniently achieved by initializing the paragraph RNN with the sentence embedding of a predefined first sentence. The generator can sequentially output diverse and topic-coherent sentences to form a personalized paragraph for an image. We present qualitative results of our model in Figure 6. Some interesting properties of our predictions include its usage of coreference in the first sentence and its ability to capture topic relationships with preceding sentences. Given

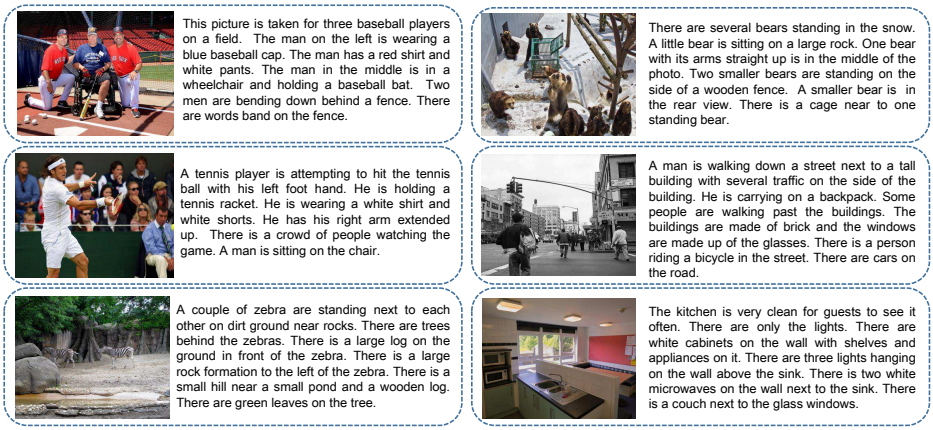

Figure 5. Example paragraph generation results of our model in the semi-supervised setting (*RTT-GAN (Semi-)*). For each image, a paragraph description with all six sentences is generated.

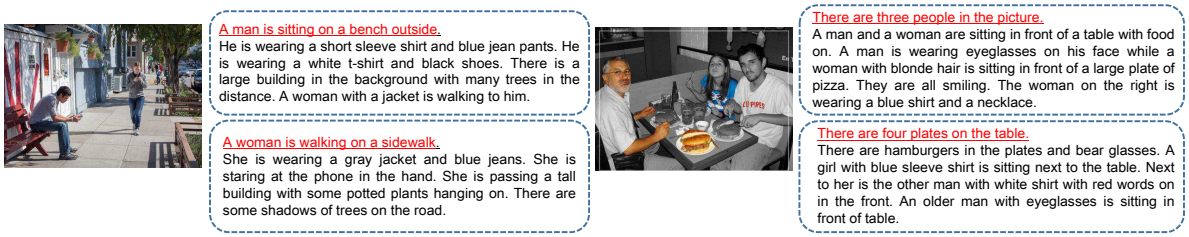

Figure 6. Personalized paragraph generations of our model (i.e. *RTT-GAN (Semi + Fully)*) by manipulating the first sentence. With two different first sentences for each image, our model can effectively generate two distinct paragraphs with different topics.

the first sentences, subsequent sentences give some details about scene elements mentioned earlier in the description and also connect to other related content. We also report the human evaluation results in Table 3 on randomly chosen 100 testing images, where three model variants are compared, i.e. *RTT-GAN (Semi- w/o discriminator)*, *RTT-GAN (Semi-)*, *RTT-GAN (Semi + Fully)*. For each image, given two first sentences with distinct topics, each model produces two personalized paragraphs accordingly. Regarding to each first sentence of the image, we present three paragraphs by three models in a random order to judges, and ask them to select the most convincing ones. The results in Table 3 indicate that 87.4% of the judges think the paragraphs generated by the models (i.e. *RTT-GAN (Semi-)*, *RTT-GAN (Semi + Fully)*) that incorporate two adversarial discriminators, look more convincing than those by *RTT-GAN (Semi-w/o discriminator)*.

4.6. Extension to Video Domain

As in Table 4, we also extend our RTT-GAN to the task of video paragraph generation and evaluate it on TACoS-MultiLevel dataset [25] that contains 185 long videos filmed in an indoor environment, following [34]. To model spatial appearance, we also extract 50 semantic regions for the frames in every second. To capture the motion patterns, we enhance the feature representation with motion features. Similar to [34], we use the pre-trained C3D [29] model on the Sport1M dataset [15], which outputs a fixed-length feature vector every 16 frames. We then perform a mean pooling over all features to generate a compact motion representation, which are used as additional inputs in every visual attention step. Our model significantly outperforms all state-of-the-arts, demonstrating its good generalization capability in video domain.

Table 4. Results of video paragraph generation on TACoS-MultiLevel in terms of BLEU-4, METEOR, CIDEr metrics.

Method	BLEU-4	METEOR	CIDEr
CRF-T [26]	25.3	26.0	124.8
CRF-M [25]	27.3	27.2	134.7
LRCN [4]	29.2	28.2	153.4
h-RNN [34]	30.5	28.7	160.2
RTT-GAN (ours)	**33.8**	**30.9**	**165.3**

5. Conclusion and Future Work

In this paper, we proposed a Recurrent Topic-Transition GAN (RTT-GAN) for visual paragraph generation. Thanks to the adversarial generative modeling, our RTT-GAN is capable of generating diverse paragraphs when only first sentence annotations are given for training. The generator incorporates visual attention and language attention mechanisms to recurrently reason about fine-grained semantic regions. In future, we will extend our generative model into other vision tasks that require joint visual and language modeling.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3, 6

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

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

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 1, 2, 8

[5] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010. 2

[6] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. CVPR, 2017. 2

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2, 3

[8] J. Gu, Z. Lu, H. Li, and V. O. Li. Incorporating copying mechanism in sequence-to-sequence learning. *ACL*, 2016. 5

[9] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *ICML*, 2017. 3

[10] Z. Hu, Z. Yang, R. Salakhutdinov, and E. P. Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017. 2

[11] T. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. B. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *NAACL- HLT*, 2016. 1

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015. 6

[13] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, 2016. 2, 4, 6

[14] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 5, 6

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014. 8

[16] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. *CVPR*, 2017. 1, 2, 4, 5, 6

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016. 1

[18] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011. 1, 2

[19] J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *ACL*, 2015. 1, 2

[20] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *arXiv preprint arXiv:1703.03054*, 2017. 2

[21] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*, pages 999–1007, 2015. 1

[22] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan. Proposal-free network for instance-level object segmentation. *arXiv preprint arXiv:1509.02636*, 2015. 1

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[24] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015. 1, 2

[25] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition*, pages 184–195. Springer, 2014. 8

[26] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, pages 433–440, 2013. 8

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[28] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 6

[29] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. *CoRR, abs/1412.0767*, 2:7, 2014. 8

[30] 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. In *ICML*, volume 14, pages 77–81, 2015. 1

[31] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. Review networks for caption generation. In *NIPS*, pages 2361–2369, 2016. 1, 2

[32] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015. 2

[33] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016. 1, 2

[34] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, pages 4584–4593, 2016. 1, 2, 8

[35] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: sequence generative adversarial nets with policy gradient. *arXiv preprint arXiv:1609.05473*, 2016. 3, 4

[36] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, pages 443–457. Springer, 2016. 1