Diverse and Coherent Paragraph Generation from Images

Moitreya Chatterjee and Alexander G. Schwing

University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
metro.smiles@gmail.com, aschwing@illinois.edu

Abstract. Paragraph generation from images is an important task for video summarization, editing, and support of the disabled, which has gained popularity recently. Traditional image captioning methods fall short on this front, since they aren't designed to generate long informative descriptions. However, the naive approach of simply concatenating multiple short sentences, possibly synthesized from traditional image captioning systems, doesn't embrace the intricacies of paragraphs: coherent sentences, globally consistent structure, and diversity. To address those challenges, we propose to augment paragraph generation techniques with "coherence vectors," "global topic vectors," and modeling of the inherent ambiguity of associating paragraphs with images via a variational auto-encoder formulation. We demonstrate the effectiveness of the developed approach on two datasets, outperforming existing state-of-the-art techniques on both.

Keywords: Captioning, Review Generation, Variational Autoencoders

1 Introduction

Daily, we effortlessly describe fun events to friends and family, showing them pictures to underline the main plot. The narrative ensures that our audience can follow along step by step and picture the missing pieces in their mind with ease. Key to filling in the missing pieces is a consistency in our narrative which generally follows the arrow of time.

While computer vision, natural language processing and artificial intelligence techniques, more generally, have made great progress in describing visual content via image or video captioning [5,11,17,28,36], the obtained result is generally a single sentence of around 20 words, describing the main observation. Even if brevity caters to today's short attention span, 20 words are hardly enough to describe subtle interactions, let alone detailed plots of our experience. Those are much more meaningfully depicted in a paragraph of reasonable length.

To this end, visual paragraph generation methods [16,21,45,25], which have been proposed very recently, provide a longer narrative when describing a given image or video. However, as argued initially, coherence between successive sentences of the narrative is a key necessity to effectively convey the plot of our experience. Importantly, models for many of the aforementioned methods provide

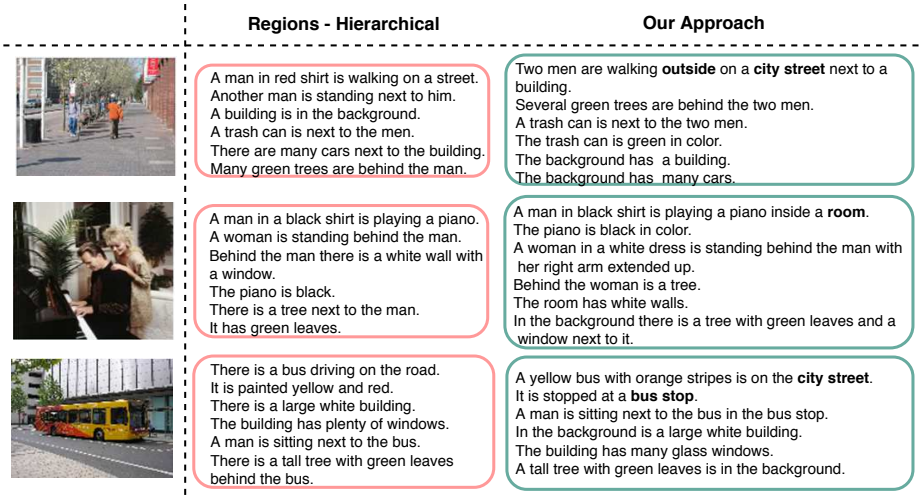

Fig. 1. Paragraphs generated with a prior state-of-the-art technique [21] and with our developed approach. Due to the introduced 'Coherence Vectors' we observe the generated paragraphs to be much more consistent than prior work [21]

.

no explicit mechanisms to ensure cross-sentence topic consistency, although a notable exception is the work of Liang *et al*. [25].

In particular, Liang *et al*. [25] propose to ensure consistency across sentence themes by training a standard paragraph generation module [21], coupled with an attention mechanism, under a Generative Adversarial Network (GAN) [13] setting which has an additional loss-term to enforce this consistency. However, difficulties associated with training GANs [3], leaves their method vulnerable to generating incoherent paragraphs.

Different from prior work, we explicitly focus on modeling the *diverse yet coherent possibilities* of successive sentences when generating a paragraph, while ensuring the 'big picture' underlying the image does not get lost in the details. To this end we develop a model that propagates, what we call "*Coherence Vectors*," which ensure cross-sentence topic smoothness, and a "*Global Topic Vector*," which captures the summarizing information about the image. Additionally, we observe improvements in the quality of the generated paragraphs, when our model is trained to incorporate diversity. Intuitively, the coherence vector embeds the theme of the most recently generated sentence. The topic vector of the next sentence is combined with the coherence vector from the most recently generated one and the global topic vector to generate a new topic vector, with the intention to ensure a smooth flow of the theme across sentences. Figure 1 illustrates a sampling of a synthesized paragraph, given an input image, using our method vis-à-vis prior work [21]. Notably, using our model we observe a smooth transition between sentence themes, while capturing summarizing information about the

image. For instance, generated paragraphs corresponding to the images in the first and the third rows in Figure 1 indicate that the images have been captured in a 'city' setting.

Following prior work we quantitatively evaluate our approach on the standard Stanford Image-Paragraph dataset [21], demonstrating state-of-the-art performance. Furthermore, different from all existing methods, we showcase the generalizability of our model, evaluating the proposed approach by generating reviews from the "Office-Product" category of the Amazon product review dataset [29] and by showing significant gains over all baselines.

In the next section, we discuss prior relevant work before discussing the details of our proposed approach in Section 3. Section 4 discusses the results of empirical evaluation. We finally conclude in Section 5, laying out avenues for future work.

2 Related Work

For a long time, associating language with visual content has been in the focus of research [24,38,4]. Early techniques in this area associated linguistic 'tag-words' with visual data. Gradually, we focused on generating entire sentences and paragraphs for visual data by bringing together techniques from both natural language processing and computer vision with the aim of building holistic AI systems that integrate naturally into common surroundings. Two tasks that spurred the growth of recent work in the language-vision area are *Image Captioning* [36,16,5,11,42], and *Visual Question Answering* [2,12,33,32,27,40,41,43]. More recently, image captioning approaches were extended to generate natural language descriptions at the level of paragraphs [21,16,25]. In the following, we review related work from the area of image captioning and visual paragraph generation, in greater detail, and point out the distinction with our work.

Image Captioning: *Image Captioning* is the task of generating textual descriptions, given an input image. Classical methods for image captioning, are usually non-parametric. These methods build a pool of candidate captions from the training set of image-caption pairs, and at test time, a fitness function is used to retrieve the most compelling caption for a given input image [24,30,4]. However the computationally demanding nature of the matching process imposes a bottleneck when considering a set of descriptions of a reasonable size.

To address this problem, Recurrent Neural Network (RNN)-based approaches have come into vogue [36,28,44,42,17,37,1,10] lately. These approaches, typically, first use a Convolutional Neural Network (CNN) [34,23] to obtain an encoding of the given input image. This encoding is then fed into an RNN which samples a set of words (from a dictionary of words) that agree the most with the image encoding. However, the captions generated through such techniques are short, spanning typically a single sentence of at most 20 words. Our approach differs from the aforementioned image captioning techniques, in that we generate a paragraph of multiple sentences rather than a short caption. Importantly, captioning techniques generally don't have to consider coherence across sentences, which is not true for paragraph generation approaches which we review next.

Visual Paragraph Generation: From a distance, the task of *Visual Paragraph Generation* resembles image captioning: given an image, generate a textual description of its content [21]. However, of importance for visual paragraph generation is the attention to detail in the textual description. In particular, the system is expected to generate a paragraph of sentences (typically 5 or 6 sentences per paragraph) describing the image in great detail. Moreover, in order for the paragraph to resemble natural language, there has to be a smooth transition across the themes of the sentences of the paragraph.

Early work in generating detailed captions, include an approach by Johnson *et al*. [16]. While generating compelling sentences individually, a focus on a theme of the story underlying a given image was missing. This problem was addressed by Krause *et al*. [21]. Their language model consists of a two-stage hierarchy of RNNs. The first RNN level generates sentence topics, given the visual representation of semantically salient regions in the image. The second RNN level translates this topic vector into a sentence. This model was further extended by Liang *et al*. [25] to encourage coherence amongst successive sentences. To this end, the language generation mechanism of Krause *et al*. [21], coupled with an attention mechanism, was trained in a Generative Adversarial Network (GAN) setting, where the discriminator is intended to encourage this coherence at training time. Dai *et al*. [8] also train a GAN for generating paragraphs. However, known difficulties of training GANs [3] pose challenges towards effectively implementing such systems. Xie *et al*. introduce regularization terms for ensuring diversity [39] but then this results in a constrained optimization problem, which does not admit a closed form solution and is thus hard to implement. Different from these approaches [25,8,39], we demonstrate that a change of the generation mechanism is better suited to obtain coherent sentence structure. To this end we introduce *Coherence Vectors* which ensure a gradual transition of themes between sentences.

Additionally, different from prior work, we also incorporate a summary of the topic vectors to sensitize the model to the 'main plot' underlying the image. Furthermore, to capture the inherent ambiguity of paragraph generation from images, *i.e*., multiple paragraphs can successfully describe an image, we cast our paragraph-generation model as a Variational Autoencoder (VAE) [18,15,7,14], enabling our model to generate a set of diverse paragraphs, given an image.

3 Our Proposed Method for Paragraph Generation

As mentioned before, coherence of sampled sentences is important for automatic generation of human-like paragraphs from visual data, while not losing sight of the underlying 'big picture' story illustrated in the image. Further, another valuable element for an automated paragraph generation system is the diversity of the generated text. In the following we develop a framework which takes into account these properties. We first provide an overview of the approach in Section 3.1, before discussing our approach to generate coherent paragraphs in Section 3.2 and finally our technique to obtain diverse paragraphs in Section 3.3.

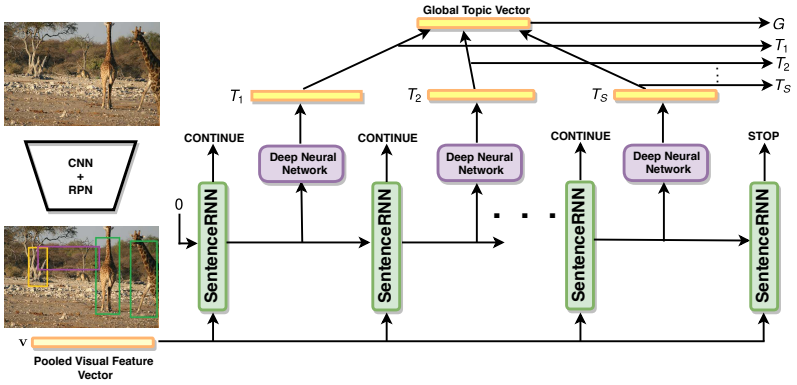

Fig. 2. Overview of the Topic Generation Net of our proposed approach illustrating the construction of the individual and 'Global Topic Vector'.

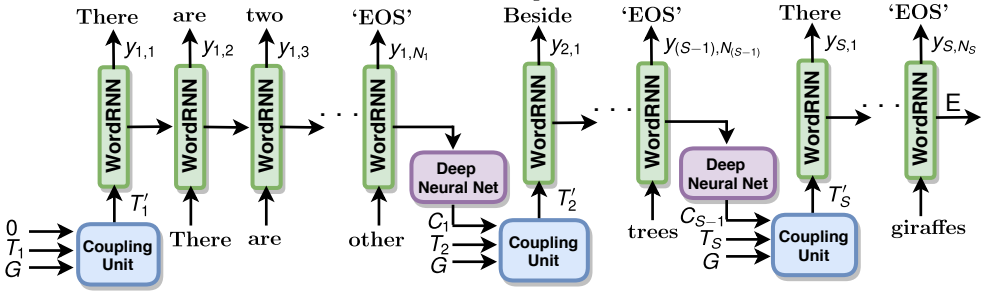

Fig. 3. Overview of the Sentence Generation Net.

3.1 Overview

To generate a paragraph $y = (y_1, \ldots, y_S)$ consisting of S sentences y_i, $i \in \{1, \ldots, S\}$, each with N_i words $y_{i,j}$, $j \in \{1, \ldots, N_i\}$, for an image x, we make use of a deep net composed out of two modules which are coupled hierarchically: the *Topic Generation Net* and the *Sentence Generation Net*.

The *Topic Generation Net* illustrated in Figure 2 seeks to extract a set of S topic vectors, $T_i \in \mathbb{R}^H \; \forall i \in \{1, \ldots, S\}$, given an appropriate visual representation of the input image x. The topic generation net is a parametric function which, recursively at every timestep, produces a topic vector T_i and a probability measure u_i indicating if more topics are to be generated. We implement this function using a recurrent net, subsequently also referred to as the *SentenceRNN*. We then leverage the topic vectors T_i to construct a *Global Topic Vector* $G \in \mathbb{R}^H$, which captures the underlying image summary. This global topic vector is constructed via a weighted combination of the aforementioned topic vectors T_i.

Figure 2 illustrates a detailed schematic of the topic generation net. Formally we use $(G, \{(T_i, u_i)\}_{i=1}^S) = \Gamma_{w_T}(x)$ to denote the input and output of the net $\Gamma_{w_T}(\cdot)$, where the vector w_T subsumes the parameters of the function. The global

topic vector G, and the individual topic vectors and probabilities $\{(T_i, u_i)\}_{i=1}^{S}$ are the output which also constitute the input to the second module.

The second module of the developed approach, called the *Sentence Generation Net*, is illustrated in Figure 3. Based on the output of the topic generation net, it is responsible for producing a paragraph y, one sentence y_i at a time.

Formally, the sentence generation module is also modeled as a parametric function which synthesizes a sentence y_i, one word $y_{i,j}$ at a time. More specifically, a recurrent net $\Gamma_{w_s}(\cdot,\cdot)$ is used to obtain the predicted word probabilities $\{p_{i,j}\}_{j=1}^{N_i} = \Gamma_{w_s}(T_i, G)$, where w_s subsumes all the parameters of the net, and $p_{i,j} \in [0,1]^V \ \forall j \in \{1,\ldots,N_i\}$ is a probability distribution over the set of V words in our vocabulary. We realize the function, $\Gamma_{w_s}(\cdot,\cdot)$ using a recurrent net, subsequently referred to as the *WordRNN*.

In order to incorporate cross-sentence coherence, rather than directly using the topic vector T_i in the WordRNN, we first construct a modified topic vector T_i', which better captures the theme of the i^{th} sentence. For every sentence i, we compute $T_i' \in \mathbb{R}^H$ via a *Coupling Unit*, by combining the topic vector T_i, the global vector G and a previous sentence representation C_{i-1}, called a *Coherence Vector*, which captures properties of the sentence generated at step $i-1$. Note that the synthesis of the first sentence begins by constructing T_1', which is obtained by coupling T_1 with the global topic vector G, and an all zero vector.

Visual Representation: To obtain an effective encoding of the input image, x, we follow Johnson *et al.* [16]. More specifically, a Convolutional Neural Network (CNN) (VGG-16 [34]) coupled with a Region Proposal Network (RPN) gives fixed-length feature vectors for every detection of a semantically salient region in the image. The obtained set of vectors $\{v_1,\ldots,v_M\}$ with $v_i \in \mathbb{R}^D$ each correspond to a region in the image. We subsequently pool these vectors into a single vector, $v \in \mathbb{R}^I$ – following the approach of Krause *et al.* [21]. This pooled representation contains relevant information from the different semantically salient regions in the image, which is supplied as input to our topic generation net. Subsequently, we use v and x interchangeably.

3.2 Coherent Paragraph Generation

The construction of coherent paragraphs adopts a two-step approach. In the first step, we derive a set of individual and a global topic-vector starting with the pooled representation of the image. This is followed by paragraph synthesis.

Topic Generation: The *Topic Generation Net* $(G, \{(T_i, u_i)\}_{i=1}^{S}) = \Gamma_{w_T}(x)$ constructs a set of relevant topics T_i for subsequent paragraph generation given an image x. Figure 2 provides a schematic of the proposed topic generation module. At first, the pooled visual representation of the image, v, is used as input for the *SentenceRNN*. The SentenceRNN is a single layer Gated Recurrent Unit (GRU) [6], parameterized by w_T. It takes an image representation v as input and produces a probability distribution u_i, over the labels 'CONTINUE' or 'STOP,' while its hidden state is used to produce the topic vector $T_i \in \mathbb{R}^H$ via a

2-layer densely connected deep neural network. A 'CONTINUE' label ($u_i > 0.5$), indicates that the recurrence should proceed for another time step, while a 'STOP' symbol terminates the recurrence.

However, automatic description of an image via paragraphs necessitates tying all the sentences of the paragraph to a 'big picture' underlying the scene. For example, in the first image in Figure 1, the generated paragraph should ideally reflect that it is an image captured in a 'city' setting. To encourage this ability we construct a *Global Topic Vector* $G \in \mathbb{R}^H$ for a given input image (see Figure 2).

Intuitively, we want this global topic vector to encode a holistic understanding of the image, by combining the aforementioned individual topic vectors as follows:

$$G = \sum_{i=1}^{n} \alpha_i T_i \quad \text{where} \quad \alpha_i = \frac{||T_i||_2}{\sum_i ||T_i||_2}. \tag{1}$$

Our intention is to facilitate representation of 'meta-concepts' (like 'city') as a weighted combination of its potential constituents (like 'car,' 'street,' 'men,' *etc.*). The synthesized global vector and the topic vectors are then propagated to the sentence generation net which predicts the words of the paragraph.

Sentence Generation: Given the individual topic vectors T_i and the global topic vector G, the *Sentence Generation Net* synthesizes sentences of the paragraph by computing word probabilities $\{p_{i,j}\}_{j=1}^{N_i} = \Gamma_{w_s}(T_i, G)$, conditioned on the previous set of synthesized words (see Figure 3). One sentence is generated for each of the S individual topic vectors $T_1, \ldots, T_S$. Synthesis of the i^{th} sentence commences by combining via the *Coupling Unit* the topic vector T_i, the global topic vector G, and the consistency ensuring *Coherence Vector* $C_{i-1} \in \mathbb{R}^H$.

The *Coupling Unit* produces a modified topic vector ($T_i' \in \mathbb{R}^H$), which is propagated to the *WordRNN* to synthesize the sentence. The *WordRNN* is a 2-layer GRU, which generates a sentence, y_i, one word at a time, conditioned on the previously synthesized words. The j^{th} word of the i^{th} sentence is obtained by selecting the word with the highest posterior probability, $p_{i,j}$, over the entries of the vocabulary V. A sentence is terminated when either the maximum word limit per sentence is reached or an 'EOS' token is predicted. In the following, we describe the mechanism for constructing the coherence vectors, and the coupling technique referenced above.

Coherence Vectors: An important element of human-like paragraphs is coherence between the themes of successive sentences, which ensures a smooth flow of the line of thought in a paragraph.

As shown in Figure 3, we encourage topic coherence across sentences by constructing *Coherence Vectors*. In the following we describe the process of building these vectors. In order to compute the coherence vector for the $(i-1)^{\text{th}}$ sentence, we extract the hidden layer representation ($\in \mathbb{R}^H$) from the WordRNN, after having synthesized the last word of the $(i-1)^{\text{th}}$ sentence. This encoding carries information about the $(i-1)^{\text{th}}$ sentence, and if favorably coupled with the topic vector T_i of the i^{th} sentence, encourages the theme of the i^{th} sentence to be

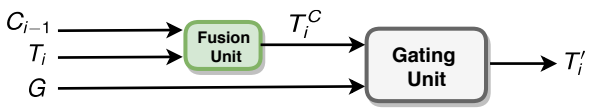

Fig. 4. The internal architecture of the 'Coupling Unit'.

coherent with the previous one. However, for the aforementioned coupling to be successful, the hidden layer representation of the $(i-1)^{\text{th}}$ sentence still needs to be transformed to a representation that lies in the same space as the set of topic vectors. This transformation is achieved by propagating the final representation of the $(i-1)^{\text{th}}$ sentence through a 2-layer deep net of fully connected units, with the intermediate layer having H activations. We used Scaled Exponential Linear Unit (SeLU) activations [20] for all neurons of this deep net. The output of this network is what we refer to as 'Coherence Vector,' $C_{(i-1)}$.

Coupling Unit: Having obtained the coherence vector C_{i-1} from the $(i-1)^{\text{th}}$ sentence, a *Coupling Unit* combines it with the topic vector of the next sentence, T_i, and the global topic representation G. This process is illustrated in Figure 4.

More specifically, we first combine C_{i-1} and T_i into a vector $T_i^C \in \mathbb{R}^H$ which is given by the solution to the following optimization problem:

$$T_i^C = \arg\min_{\hat{T}_i^C} \quad \alpha||T_i - \hat{T}_i^C||_2^2 + \beta||C_{i-1} - \hat{T}_i^C||_2^2 \quad \text{with} \quad \alpha, \beta \geq 0.$$

The solution, when α, β both are not equal to 0, is given by:

$$T_i^C = \frac{\alpha T_i + \beta C_{i-1}}{\alpha + \beta}.$$

We refer the interested reader to the supplementary for this derivation. Intuitively, this formulation encourages T_i^C to be 'similar' to both the coherence vector, C_{i-1} and the current topic vector, T_i – thereby aiding cross-sentence topic coherence. Moreover, the closed form solution of this formulation permits an efficient implementation as well.

This obtained vector, T_i^C is then coupled with the global topic vector G, via a gating function. We implement this gating function using a single GRU layer with vector T_i^C as input and global topic vector G as its hidden state vector. The output of this GRU cell, T_i', is the final topic vector which is used to produce the i^{th} sentence via the WordRNN.

Loss Function and Training: Both Topic Generation Net and Sentence Generation Net are trained jointly end-to-end using labeled training data, which consists of pairs (x, y) of an image x and a corresponding paragraph y. If one image is associated with multiple paragraphs, we create a separate pair for each. Our training loss function $\ell_{\text{train}}(x, y)$ couples two cross-entropy losses, a binary cross-entropy sentence-level loss on the distribution u_i for the i^{th} sentence

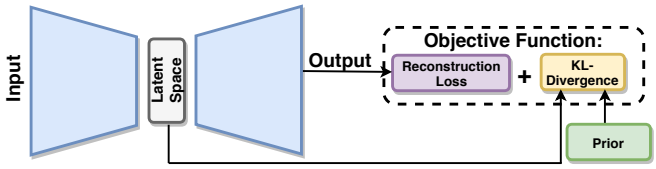

Fig. 5. General Framework of our VAE Formulation.

$(\ell_{\mathrm{s}}(u_i, \mathbb{1}_{i \leq S}))$, and a word-level loss, on the distribution $p_{i,j}$ for the j^{th} word of the i^{th} sentence $(\ell_{\mathrm{w}}(p_{i,j}, y_{i,j}))$. Assuming S sentences in the ground-truth paragraph, with the i^{th} sentence having N_i words, our loss function is given by:

$$\ell_{\mathrm{train}}(x, y) = \lambda_s \sum_{i=1}^{S} \ell_{\mathrm{s}}(u_i, \mathbb{1}_{i=S}) + \lambda_w \sum_{i=1}^{S} \sum_{j=1}^{N_i} \ell_{\mathrm{w}}(p_{i,j}, y_{i,j}), \tag{2}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Armed with this loss function our method is trained via the Adam optimizer [19] to update the parameters w_T and w_s.

3.3 Diverse Coherent Paragraph Generation

The aforementioned scheme for generating paragraphs lacks in one key aspect: it doesn't model the ambiguity inherent to a *diverse* set of paragraphs that fit a given image. In order to incorporate this element of diversity into our model, we cast the designed paragraph generation mechanism into a *Variational Autoencoder* (VAE) [18] formulation, a generic architecture of which is shown in Figure 5. Note that we prefer a VAE formulation over other popular tools for modeling diversity, such as GANs, because of the following reasons: (1) GANs are known to suffer from training difficulties unlike VAEs [3]; (2) The intermediate sampling step in the generator of a GAN (for generating text) is not differentiable and thus one has to resort to Policy Gradient-based algorithms or Gumbel softmax, which makes the training procedure non-trivial. The details of our formulation follow.

VAE Formulation: The goal of our VAE formulation is to model the log-likelihood of paragraphs y conditioned on images x, *i.e.*, $\ln p(y|x)$. To this end, a VAE assumes that the data, *i.e.*, in our case paragraphs, arise from a low-dimensional manifold space represented by samples z. Given a sample z, we reconstruct, *i.e.*, decode, a paragraph y by modeling $p_\theta(y|z, x)$ via a deep net. The ability to randomly sample from this latent space provides diversity. In the context of our task the decoder is the paragraph generation module described in Section 3.2, augmented by taking samples from the latent space as input. We subsequently denote the parameters of the paragraph generation module by $\theta = [w_T, w_{\mathrm{s}}]$. To learn a meaningful manifold space we require the decoder's posterior $p_\theta(z|y, x)$. However computing the decoder's posterior $p_\theta(z|y, x)$ is known to be challenging [18]. Hence, we commonly approximate this distribution

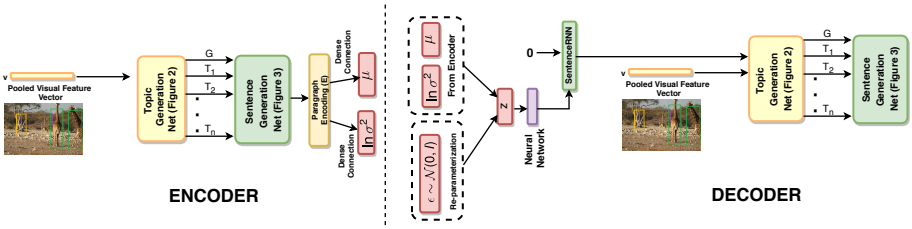

Fig. 6. Architecture of the Encoder and Decoder of our VAE formulation.

using a probability $q_\phi(z|y,x)$, which constitutes the encoder section of the model, parameterized by ϕ. Further, let $p(z)$ denote the prior distribution of samples in the latent space. Using the aforementioned distributions, the VAE formulation can be obtained from the following identity:

$$\ln p(y|x) - KL(q_\phi(z|y,x), p_\theta(z|y,x)) = \mathbb{E}_{q_\phi(z|y,x)}[\ln p_\theta(y|z,x)] - KL(q_\phi(z|y,x), p(z)),$$

where $KL(\cdot,\cdot)$ denotes the KL divergence between two distributions. Due to the non-negativity of the KL-divergence we immediately observe the right hand side to be a lower bound on the log-likelihood $\ln p(y|x)$ which can be maximized w.r.t. its parameters ϕ and θ. The first term on the right hand side optimizes the reconstruction loss, *i.e.*, the conditional likelihood of the decoded paragraph (which is equivalent to optimizing the loss in Equation 2), while the second term acts like a distributional regularizer (ensuring smoothness). Training this system end-to-end via backpropagation is hard because of the intermediate, non-differentiable, step of sampling z. This bottleneck is mitigated by introducing the *Re-parameterization Trick* [18]. Details of the encoder and decoder follow.

Encoder: The encoder architecture is shown in Figure 6. Given the image x and a ground-truth paragraph y we encode the sample (x, y) by passing it through the topic and sentence generation nets. We then extract the hidden state vector ($E \in \mathbb{R}^H$) from the final WordRNN of the Sentence Generation net. This vector is passed through a 1-layer densely connected net, the output layer of which has $2H$ neurons. We assume the conditional distribution underlying the encoder, $q_\phi(z|y,x)$ to be a Gaussian, whose mean μ is the output of the first H neurons, while the remaining H neurons give a measure of the log-variance, *i.e.*, $\ln \sigma^2$.

Decoder: The decoding architecture is also shown in Figure 6. While decoding, we draw a sample $z \sim \mathcal{N}(0, I)$ ($z \in \mathbb{R}^H$, for training, we additionally shift and scale it by: $z = \mu + \sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$) and pass it to the SentenceRNN, via a single-layer neural net with I output neurons. The hidden state of this RNN is then forward propagated to the SentenceRNN unit, which also receives the pooled visual vector v. Afterwards, the decoding proceeds as discussed before.

4 Experimental Evaluations

Datasets: We first conduct experiments on the *Stanford image-paragraph dataset* [21], a standard in the area of visual paragraph generation. The dataset consists of 19,551 images from the Visual Genome [22] and MS COCO dataset [26]. These images are annotated with human-labeled paragraphs, 67.50 words long, with each sentence having 11.91 words, on average. The experimental protocol divides this dataset into 14,575 training, 2,487 validation, and 2,489 testing examples [21]. Further, in order to exhibit generalizability of our approach, different from prior work, we also undertake experiments on the much larger, *Amazon Product-Review dataset* ('Office-Products' category) [29] for the task of generating reviews. This is a dataset of images of common categories of office-products, such as printer, pens, *etc*. (see Figure 7), crawled from amazon.com. There are 129,970 objects in total, each of which belongs to a category of office products. For every object, there is an associated image, captured in an uncluttered setting with sufficient illumination. Accompanying the image, are multiple reviews by users of the product. Further, each review is supplemented by a star rating, an integer between 1 (poor) and 5 (good). On an average there are 6.4 reviews per star rating per object. A review is 71.66 words long, with 13.52 words per sentence, on average. We randomly divide the dataset into 5,000 test, and 5,000 validation examples, while the remaining examples are used for training.

Baselines: We compare our approach to several recently introduced and our own custom designed baselines. Given an image, 'Image-Flat' directly synthesizes a paragraph, token-by-token, via a single RNN [17]. 'Regions-Hierarchical' on the other hand, generates a paragraph, sentence by sentence [21]. Liang *et al*. [25] essentially train the approach of Krause *et al*. [21] in a GAN setting ('RTT-GAN'), coupled with an attention mechanism. However, Liang *et al*. also report results on the Stanford image-paragraph dataset by using additional training data from the MS COCO dataset, which we refer to as 'RTT-GAN (Plus).' We also train our model in a GAN setting and indicate this baseline as 'Ours (GAN).' Additionally, we create baselines for our model without coherence vectors, essentially replacing them with a zero vector for every time-step. We refer to this baseline as 'Ours (NC).' In another setting, we only set the global topic vector to zero for every time-step. We refer to this baseline as 'Ours (NG).'

Evaluation Metrics: We report the performance of all models on 6 widely used language generation metrics: BLEU-{1, 2, 3, 4} [31], METEOR [9], and CIDEr [35]. While the BLEU scores largely measure just the n-gram precision, METEOR, and CIDEr are known to provide a more robust evaluation of language generation algorithms [35].

Implementation Details: For the Stanford dataset, we set the dimension of the pooled visual feature vector, v, to be 1024. For the Amazon dataset, however, we use a visual representation obtained from VGG-16 [34]. Since, these images are generally taken with just the principal object in view (see Figure 7), a standard

Table 1. Comparison of captioning performance on the Stanford Dataset

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Image-Flat [17]	34.04	19.95	12.2	7.71	12.82	11.06
Regions-Hierarchical [21]	41.9	24.11	14.23	8.69	15.95	13.52
RTT-GAN [25]	41.99	24.86	14.89	9.03	17.12	16.87
RTT-GAN (Plus) [25]	42.06	25.35	14.92	9.21	18.39	20.36
Ours (NC)	42.03	24.84	14.47	8.82	16.89	16.42
Ours (NG)	42.05	25.05	14.59	8.96	17.26	18.23
Ours	42.12	25.18	14.74	9.05	17.81	19.95
Ours (with GAN)	42.04	24.96	14.53	8.95	17.21	18.05
Ours (with VAE)	**42.38**	**25.52**	**15.15**	**9.43**	**18.62**	**20.93**
Human (as in [21])	42.88	25.68	15.55	9.66	19.22	28.55

CNN suffices. We extract representations from the penultimate fully connected layer of the CNN, giving us a vector of 4,096 dimensions. Hence, we use a single-layer neural network to map this vector to the input vector of 1,024 dimensions. For both SentenceRNN and WordRNN, the GRUs have hidden layers (H) of 512 dimension. For the Amazon dataset, we condition the first SentenceRNN, with an H-dimensional embedding of the number of stars. We set λ_s, λ_w to be 5.0, and 1.0 respectively, the maximum number of sentences per paragraph, $S_{\max}$, to be 6, while the maximum number of words per sentence is set to be 30, for both datasets. In the coupling unit, α is set to 1.0, and β is set to 1.5 for the Stanford dataset, while for the Amazon dataset the corresponding values are 1.0 and 3.0. The learning rate of the model is to 0.01 for the first 5 epochs and is halved every 5 epochs after that, for both datasets. These hyper-parameters are chosen by optimizing the performance, based on the average of METEOR and CIDEr scores, on the validation set for both datasets. We use the same vocabulary as Krause*et al.* [21], for the Stanford dataset, while a vocabulary size of the $11,000$ most frequent words is used for the Amazon dataset. Additional implementational details can be found on the project website[1]. For purposes of comparison, for the Amazon dataset, we run our implementation of all baselines, with their hyper-parameters picked based on a similar protocol, while for the Stanford dataset we report performance for prior approaches directly from [25].

Results: Tables 1 and 2 show the performance of our algorithm vis-à-vis other comparable baselines. Our model, especially when trained in the VAE setting, outperforms all other baselines (on all 6 metrics). Even the models trained under the regular (non-VAE) setup outperform most of the baselines and are comparable to the approach of Liang *et al.* [25], an existing state-of-the-art for this task. Our performance on the rigorous METEOR and CIDEr metrics, on both datasets, attest to our improved paragraph generation capability. The capacity to generate diverse paragraphs, using our VAE setup, pays off especially well on the Amazon dataset, since multiple reviews with the same star rating are associated with

[1] https://sites.google.com/site/metrosmiles/research/research-projects/capg_revg

Table 2. Comparison of captioning performance on the Amazon Dataset

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Image-Flat [17]	40.31	30.63	25.32	15.64	10.97	9.63
Regions-Hierarchical [21]	45.74	34.8	27.54	16.67	14.23	12.02
RTT-GAN [25]	45.93	36.42	28.28	17.26	16.29	15.67
Ours (NC)	45.85	35.97	27.96	16.98	15.86	15.39
Ours (NG)	45.88	36.33	28.15	17.17	16.04	15.54
Ours	46.01	36.86	28.73	17.45	16.58	16.05
Ours (with GAN)	45.86	36.25	28.07	17.06	15.98	15.43
Ours (with VAE)	**46.32**	**37.45**	**29.42**	**18.01**	**17.64**	**17.17**

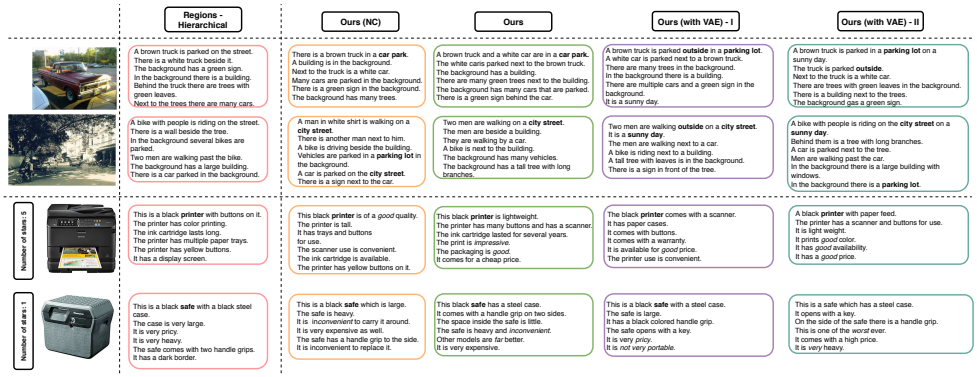

Fig. 7. Paragraphs generated under different settings with our developed approach, vis-à-vis Regions-Hierarchical [21]. The first, and second images are from the Stanford dataset, while the third and fourth images are from the Amazon dataset.

an object, creating an inherent ambiguity. Noticeably, our model is worse off in terms of performance, when trained under the GAN setting. This observation is along the lines of prior work [8]. We surmise that this results from the inherent difficulty of training GANs properly [3] in conjunction with the fact that the GAN-based setup isn't trained directly with maximum-likelihood.

Qualitative results: Figure 7 presents a sampling of our generated paragraphs. The first example in the figure (the first row) shows that our model can generate coherent paragraphs, while capturing meta-concepts like 'car-park' or 'parking lot,' from images with complex scenes. Regions-Hierarchical [21] faces challenges to incorporate these 'meta-concepts' into the generated paragraphs. For several of the instances in the Amazon dataset (such as the images in the third and fourth rows), both our method and Regions-Hierarchical [21] successfully detect the principal object in the image. We speculate that this is due to easy object recognition for images of the Amazon dataset, and to a lesser extent due to an improved paragraph generation algorithm. Additionally in the VAE setting, we are able to generate two distinctly different paragraphs with the same set of

inputs, just by sampling a different z each time (the two rightmost columns in Figure 7), permitting our results to be diverse. Moreover, for the Amazon dataset (third and fourth rows in Figure 7) we see that our model learns to synthesize 'sentiment' words depending on the number of input stars. We present additional visualizations in the supplementary.

Ablation study: In one setting, we judge the importance of coherence vectors, by just using the global vector and setting the coherence vectors to 0, in the sentence generation net. The results for this setting ('Ours (NC)') are shown in Tables 1,2, while qualitative results are shown in Figure 7. These numbers reveal that just by incorporating the global topic vector it is feasible to generate reasonably good paragraphs. However, incorporating coherence vectors makes the synthesized paragraphs more human-like. A look at the second column of Figure 7 shows that even without coherence vectors we are able to detect global topics like 'car-park' but the sentences seem to exhibit sharp topic transition, quite like Regions-Hierarchical approach. We rectify this by introducing coherence vectors.

In another setting, we set the global topic vector to 0, at every time-step, while retaining the coherence vectors. The performance in this setting is indicated by 'Ours (NG)' in Tables 1,2. The results suggest that incorporating the coherence vectors is much more critical for improved performance.

5 Conclusions and Future Work

In this work, we developed 'coherence vectors' which explicitly ensure consistency of themes between generated sentences during paragraph generation. Additionally, the 'global topic vector' was designed to capture the underlying main plot of an image. We demonstrated the efficacy of the proposed technique on two datasets, showing that our model when trained with effective autoencoding techniques can achieve state-of-the-art performance for both caption and review generation tasks. In the future we plan to extend our technique for the task of generation of even longer narratives, such as stories.

Acknowledgments: This material is based upon work supported in part by the National Science Foundation under Grant No. 1718221, Samsung, and 3M. We thank NVIDIA for providing the GPUs used for this research.

References

1. Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional Image Captioning. In: Proc. CVPR (2018)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proc. ICCV (2015)
3. Arjovsky, M., etal.: Wasserstein gan. arXiv preprint 2017
4. Chatterjee, M., Leuski, A.: A novel statistical approach for image and video retrieval and its adaption for active learning. In: Proc. ACM Multimedia (2015)
5. Chen, X., Lawrence Zitnick, C.: Mind's eye: A recurrent visual representation for image caption generation. In: Proc. CVPR (2015)

6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint 2014
7. Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A.C., Bengio, Y.: A recurrent latent variable model for sequential data. In: Proc. NIPS (2015)
8. Dai, B., etal.: Towards diverse and natural image descriptions via a conditional gan. arXiv preprint 2017
9. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proc. ninth workshop on statistical machine translation (2014)
10. Deshpande, A., Aneja, J., Wang, L., Schwing, A.G., Forsyth, D.A.: Diverse and Controllable Image Captioning with Part-of-Speech Guidance. In: https://arxiv.org/abs/1805.12589 (2018)
11. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proc. CVPR (2015)
12. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question. In: Proc. NIPS (2015)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. NIPS (2014)
14. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. arXiv preprint 2015
15. Jain, U., Zhang, Z., Schwing, A.: Creativity: Generating diverse questions using variational autoencoders. In: Proc. CVPR (2017)
16. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proc. CVPR (2016)
17. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proc. CVPR (2015)
18. Kingma, D.P., etal.: Auto-encoding variational bayes. arXiv preprint 2013
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint 2014
20. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Proc. NIPS (2017)
21. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: Proc. CVPR (2017)
22. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proc. NIPS (2012)
24. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proc. NIPS (2004)
25. Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P.: Recurrent Topic-Transition GAN for Visual Paragraph Generation. In: Proc. ICCV (2017)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proc. ECCV (2014)
27. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proc. ICCV (2015)
28. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint 2014

29. McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proc. ACM SIGIR (2015)
30. Pan, J.Y., Yang, H.J., Duygulu, P., Faloutsos, C.: Automatic image captioning. In: Proc. ICME (2004)
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proc. ACL (2002)
32. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Proc. NIPS (2015)
33. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Proc. CVPR (2016)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint 2014
35. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proc. CVPR (2015)
36. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proc. CVPR (2015)
37. Wang, L., Schwing, A.G., Lazebnik, S.: Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. In: Proc. NIPS (2017)
38. Xiao, Y., Chua, T.S., Lee, C.H.: Fusion of region and image-based techniques for automatic image annotation. In: Proc. International Conference on Multimedia Modeling (2007)
39. Xie, P.: Diversity-Promoting and Large-Scale Machine Learning for Healthcare. http://www.cs.cmu.edu/~pengtaox/thesis_proposal_pengtaoxie.pdf (2018), [Online; accessed 25-July-2018]
40. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: Proc. ICML (2016)
41. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Proc. ECCV (2016)
42. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proc. ICML (2015)
43. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proc. CVPR (2016)
44. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proc. CVPR (2016)
45. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: Proc. CVPR (2016)