IDC HERZLIYA | Efi Arazi School of Computer Science

The Interdisciplinary Center,
Herzliya
Efi Arazi School of Computer Science
M.Sc. program - Research Track

Ordered Attention for Visual Storytelling

by
Tom Braude

September 20, 2020

M.Sc. dissertation, submitted in partial fulfillment of the requirements for the M.Sc. degree, research track, School of Computer Science
The Interdisciplinary Center, Herzliya

This work was carried out under the supervision of Prof. Ariel Shamir from the Efi Arazi School of Computer Science, The Interdisciplinary Center, Herzliya.

Acknowledgements

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1390/19) and the AWS Cloud Credit for Research program.

Abstract

We address the problem of visual storytelling, *i.e.*, generating a story for a given sequence of images. While each sentence of the story should describe a corresponding image, a coherent story also needs to be consistent and relate to both future and past images. To achieve this we develop ordered image attention (OIA). OIA models interactions between the sentence-corresponding image and important regions in other images of the sequence. To highlight the important objects, a message-passing-like algorithm collects representations of those objects in an order-aware manner. To generate the story's sentences, we then highlight important image attention vectors with an Image-Sentence Attention (ISA). Further, to alleviate common linguistic mistakes like repetitiveness, we introduce an adaptive prior. The obtained results improve the METEOR score on the VIST dataset by 1%. In addition, an extensive human study verifies coherency improvements and shows that OIA and ISA generated stories are more focused, shareable, and image-grounded.

0 Contents

0 List of Figures

0 List of Tables

1 Introduction

Visual Storytelling (VST) [5; 6] – the task of generating a story based on a sequence of images – goes beyond a basic understanding of visual scenes and can be applied in many real-world scenarios, *e.g.*, to support the visually impaired. Moreover, VST reflects on the creative ability of intelligent systems. Although similar in concept to other cognitive tasks such as image captioning and visual question answering, VST differs as it requires to reason over a *sequence* of images while simultaneously ensuring coherence across multiple generated sentences. To achieve this, VST methods need to address two major challenges: the first is visual and relates to grounding the story's text to the images. The second is linguistic and relates to the quality of the story. Both challenges can be described in terms of coherency: the story should be coherent by itself, and coherent with the images.

Prior research on VST started to address the aforementioned challenges. Early works expand captioning [7; 8; 9], focusing sentence generation mainly on the current image [10; 11]. This limits the ability to incorporate complex semantic information, which is necessary for visual reasoning. Prior work also makes limited use of temporal dependence and history, *e.g.*, sentences that have already been generated are not used. Consequently, the output lacks narrative consistency and is prone to linguistic errors such as *repetitiveness* and *incoherence* [12].

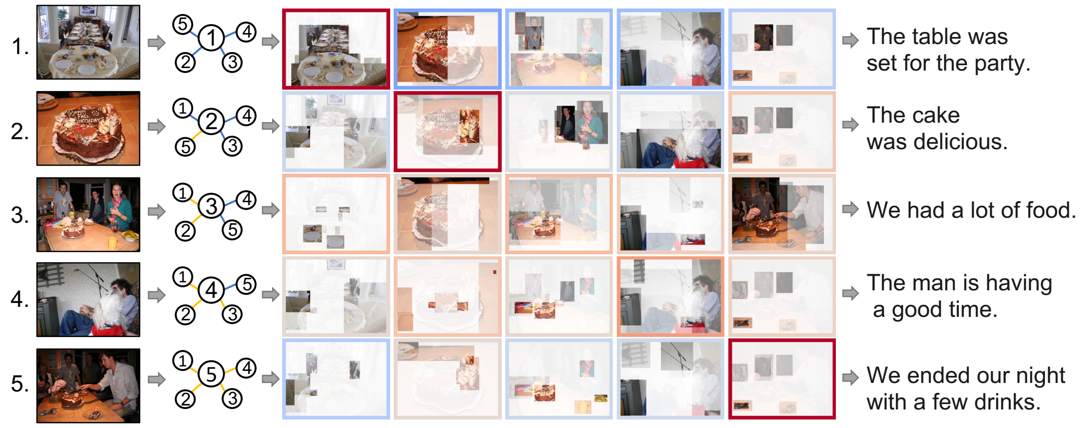

1.
The table was set for the party.
2.
The cake was delicious.
3.
We had a lot of food.
4.
The man is having a good time.
5.
We ended our night with a few drinks.

Figure 1.1: We propose Ordered Image Attention (OIA) to form the structure of a sentence and to encourage coherency. Each row shows the spatial attention of the five images created when generating a specific sentence. We find important objects by collecting directional interactions. The relative order to the sentence-corresponding image determines the connection type, illustrated as the blue and orange edges for preceding and proceeding connections. The attended images' border indicates the image attention importance formed by the Image-Sentence Attention (ISA). *E.g.*, red indicates a high attention score, meaning the image is essential for generating that sentence. Our model performs this step for all five images in parallel, creating a total of 25 spatial attention maps, that are fed into the decoder to create the sentences in order.

To mitigate these issues, later works strive to generate more meaningful stories via adversarial and reinforcement learning [1; 13], which remain delicate to train.

Importantly, images are not independent. For example, if the first image in a sequence shows a protest, the model may want to focus on signs in later images. Conversely, if the last image shows a ring on a finger, then the model should pay attention to wedding-related objects and activities in the preceding images. This is important for VST because sentences are created per image but are part of a story. Hence, objects that the model is focusing on in one image should be conditioned on the selection in other images.

To do this we develop a novel model which (1) implicitly reasons over objects, activities, and their temporal dependencies in each image; and which (2) improves the coherency of the narrative. To reason over objects and activities in each image, *i.e.*, to understand their dependencies and their temporal ordering, we introduce *ordered image attention* (OIA). As illustrated in Fig. 1.1, for each image, OIA accumulates representation information from objects detected within the corresponding image into an attended image representation. Importantly, accumulation factors depend on whether the image precedes or succeeds the image for which we are currently generating the sentence, which permits to establish an order. The attended image representations are subsequently summarized into a context embedding via an Image-Sentence Attention (ISA) unit, before being used for sentence decoding.

In addition, to alleviate common linguistic mistakes like repetitiveness and to promote coherence in the story, we incorporate information from the story generated up to the current sentence into the sentence generation decoder. Specifically, the decoding strategy decays the probability of a word if it has already been used in the story. The decoder also maintains a separate prior over the output probability distribution, independent from the language generation unit. This prior is based on counts of the words that were already predicted in the story. Both the prior, and the Recurrent Neural Net (RNN) decoder output are combined to predict the next word in the sentence.

Empirical results on the challenging VIST dataset [6] demonstrate that the proposed method generates stories with an improved narrative quality. The method outperforms prior state-of-the-art by 1% on the METEOR score. Examples of stories generated by the approach are shown in Fig. 1.1. We also present a user study demonstrating the advantage of the model in terms of coherency.

2 Related Work

Vision+Language has been an active area of research for many years, addressing tasks such as image/video captioning, paragraph generation, and visual question answering. We briefly review those related areas in the following.

2.1 Image Captioning

Bernard *et al.* [14] first explored annotating images with text. Since then, image/video captioning has seen a surge of research activity. Initial work utilized pre-trained image embeddings from a CNN network. The success of attention mechanisms for language translation quickly transferred to image captioning as well [8]. Later work leveraged advances in object detection and proposed a bottom-up/top-down attention approach to attend to specific objects in the image instead of fixed spatial regions [4]. Different from image captioning, for visual storytelling, both story coherency and visual grounding are important.

2.2 Multimodal Attention

Multimodal problems are characterized by input data that comes from different domains, *e.g.*, visual and linguistic. This raises two challenges: 1) how to model

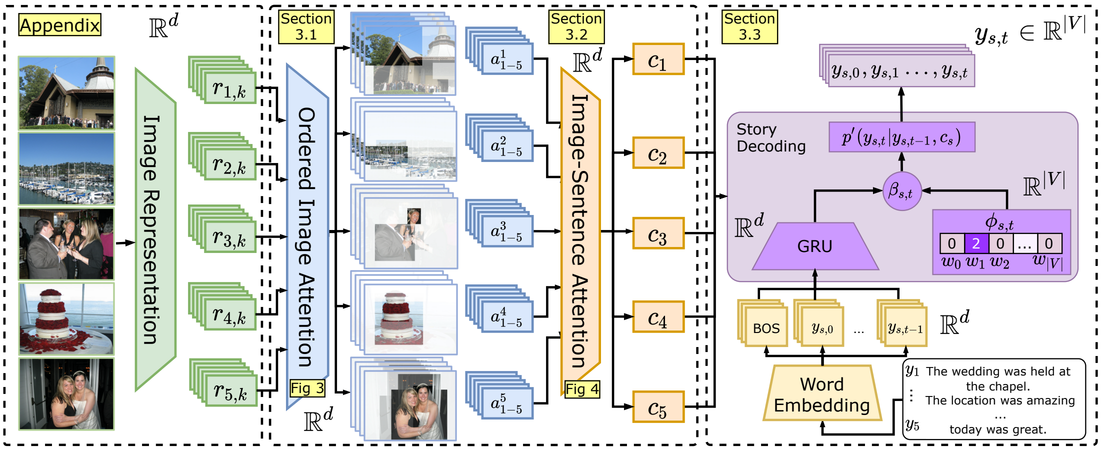


Figure 2.1: Our architecture for Visual Storytelling synthesis.

interactions between different domains, and 2) how to manage the large input data. Considering those challenges, attention has been a prominent tool as it models interactions to select the important elements. In early work, Xu *et al*. [8] used interaction-based attention with the image at each caption generation step. This idea was later extended to visual question answering [15]. To imitate multi-step reasoning, Yang *et al*. [16] stacked attention modules sequentially. Later, many works concentrated on better vector-fusion modeling [17; 18; 19; 20]. Importantly, Lu *et al*. [21] suggested attending to the visual and textual modalities separately. Afterward, Kim *et al*. [22] proposed a bilinear module that efficiently generates attention for every pair. Following Lu *et al*. [21], Schwartz *et al*. [23; 24] suggested a general framework that extends attention to any number of utilities via local and interaction-based factors. We improve upon those ideas by suggesting an ordered attention. This ensures that interaction modeling is affected by the image position in a sequence.

2.3 Visual Storytelling

Huang *et al.* [6] introduced the Visual Storytelling task. Initially, Gonzalez *et al.* [10] adapted work by Vinyals *et al.* [7] used for captioning. Kim *et al.* [25] presented a Seq2Seq [26] approach with a decoding sampling strategy aimed to reduce the amount of repetition based on a word list. We improve their strategy by using a data-driven approach, penalizing each word differently based on its average counts. Wang *et al.* [1] employ adversarial learning to improve output stories. Huang *et al.* [13] utilize a reinforcement learning (RL) approach based on inter-image relations. Later works by Li *et al.* [27] and Zhang *et al.* [28] rely on preprocessing the data to better ground visual elements to the text while Yang *et al.* [29] and Hsu *et al.* [30] enrich the data with an external word common-sense knowledge graph. Our approach captures inter-image relations via ordered attention and is trained in an end-to-end manner alleviating the computational drawbacks of preprocessing or RL. Recently, state-of-the-art results were obtained by generating scene graphs for each image in the sequence [31]. Conversely, our image representations are dependant on all the images in the sequence.

3 Method

The goal of visual storytelling is to generate a story, composed of N *ordered* sentences $\{y_s|1 \leq s \leq N\}$, given an *ordered* sequence of images $I = \{I_s|1 \leq s \leq N\}$. Each sentence $y_s = (y_{s,0}, \dots, y_{s,t}, \dots)$ is composed of words $y_{s,t} \in \mathcal{Y}$ from vocabulary $\mathcal{Y}$.

The order in which the images are given is essential as it defines the plot line of the story. The story should be focused, *i.e.*, each sentence should be related to the remainder of the story. Importantly, the sentences should form a coherent body of text describing the set of images, and not only a set of related information. For instance, the story *"The church was beautiful. The bride and groom walk down the aisle. The cake was amazing."* is less coherent than: *"We went to the church for the wedding today. The bride and groom were excited for the day. Both cut the cake together."*

Overview: To address this challenge, we develop the model illustrated in Fig. 2.1. It infers conditional probabilities $p'(y_{s,t}|y_{s,t-1}, c_s)$ for the t-th word $y_{s,t} \in \mathcal{Y}$ in sentence y_s given the previous word $y_{s,t-1}$ and the context embedding c_s for sentence s. The context embedding c_s summarizes region representations $r_{i,k}$ (Sec. 3.1 of all K object regions across all N images I_i ($i \in [1, N]$, $k \in [1, K]$) via Ordered Image Attention (OIA) (Sec. 3.2) and Image-Sentence Attention (ISA) (Sec. 3.3). Specifically, when generating sentence s, OIA computes an attended image representation a_i^s for every image I_i by attending to the K region representations $r_{i,k}$ (Sec. 3.2). These attended image representations a_i^s are subsequently summarized into the context embedding c_s via an image-sentence attention (Sec. 3.3).

Below we first discuss our initial image representation. We then describe the computation of the attended image representation a_i^s (Sec. 3.2), before detailing computation of the context embedding c_s (Sec. 3.3) and computation of the conditional probabilities $p'(y_{s,t}|y_{s,t-1}, c_s)$ (Sec. 3.4).

3.1 Image Representation

An initial pre-processing step represents each of the input images I_i via K regional features $r_{i,k} \in \mathbb{R}^d, 1 \leq k \leq K$. For this we use bottom-up attention features [4]. Specifically, for each image I_i we first extract the top K region features $e_{i,k} \in \mathbb{R}^m$. Hereby, $e_{i,k}$ is an m-dimensional feature vector extracted from a pre-trained image classification network [2] along with their respective bounding boxes $b_{i,k} \in \mathbb{R}^4$, and classes $c_{i,k} \in \mathbb{N}$. The final d-dimensional representation $r_{i,k} \in \mathbb{R}^d$, of each region is defined by a combination of the extracted semantic features. Formally,

$$r_{i,k} = W_r[W_e e_{i,k} + W_b b_{i,k} + E_c(c_{i,k})], \tag{3.1}$$

where $W_r \in \mathbb{R}^{d \times d}$, $W_e \in \mathbb{R}^{d \times m}$, $W_b \in \mathbb{R}^{d \times 4}$, and E_c are trainable parameters shared between all images. We set $K = 36$ in our proposed model. Biases and normalization are omitted for readability.

3.2 Ordered Image Attention (OIA)

Ordered Image Attention (OIA) is designed to 1) form a structure across ordered images and to 2) select the relevant objects per image. For this we model preceding and proceeding interactions separately using different attention factors. We calibrate each factor's importance with trainable scalars, which forms a graph of dependencies between the images. For each sequence of N images, the model infers a total of N^2 attention maps, one per image for each sentence. We detail this module next.

3.2.1 Attention Belief

For each image $I_i = \{r_{i,1}, \ldots r_{i,K}\}$ we consider a set of K regions, represented by their feature vectors $r_{i,k} \in \mathbb{R}^d$, where d is the objects' embedding dimension. Suppose we are currently generating sentence y_s $(1 \leq s \leq N)$. To do this we first compute an attended image representation a_i^s as follows

$$a_i^s = \sum_{k=1}^{K} b_{i,k}^s r_{i,k}, \tag{3.2}$$

where $b_{i,k}^s \geq 0$ is the attention belief highlighting the importance of the k-th object in the i-th image when generating the s-th sentence. Importantly, for every image I_i we require $b_{i,k}^s$ to be a valid probability distribution, *i.e.*, we also enforce $\sum_{k=1}^{K} b_{i,k}^s = 1 \; \forall s, i$.

The object attention belief $b_{i,k}^s$ is dependent on all the input data, *i.e.*, other objects and images. To avoid complex computation, we factorize the belief $b_{i,k}^s$ into two pairwise dependencies that preserve the order, and a local term. For the pairwise terms we use $\mu_{j\to i}^{\text{bwd}}$, which is a message from a preceding image I_j, or $\mu_{j\to i}^{\text{fwd}}$, which is a message from a subsequent image I_j. We also use $\mu_{i\to i}$ for self-messages. Additionally, we include a local factor $\Psi_i(r_{i,k})$ that considers the object representation. Unlike the messages mentioned before, the local factor does not rely on interactions with other objects. We aggregate all the messages along with the local factor as illustrated in Fig. 3.1. For normalization we employ a softmax.

Formally we compute the attention belief $b^s_{i,k}$ by distinguishing three cases. If $i = s$ we have

$$\begin{aligned} b^s_{i,k} \;\propto\; & \exp(\alpha^s_i \Psi_i(r_{i,k}) + \alpha^s_{i,i}\mu_{i\to i}(r_{i,k}) + \\ & \sum_{j<i} \alpha^s_{i,j}\mu^{\text{bwd}}_{j\to i}(r_{i,k}) + \sum_{j>i} \alpha^s_{i,j}\mu^{\text{fwd}}_{j\to i}(r_{i,k})). \end{aligned} \tag{3.3}$$

If $i < s$ we use

$$\begin{aligned} b^s_{i,k} \;\propto\; & \exp(\alpha^s_i \Psi_i(r_{i,k}) + \\ & \alpha^s_{i,i}\mu_{i\to i}(r_{i,k}) + \alpha^s_{i,s}\mu^{\text{bwd}}_{s\to i}(r_{i,k})). \end{aligned} \tag{3.4}$$

If $i > s$ we obtain

$$\begin{aligned} b^s_{i,k} \;\propto\; & \exp(\alpha^s_i \Psi_i(r_{i,k}) + \\ & \alpha^s_{i,i}\mu_{i\to i}(r_{i,k}) + \alpha^s_{i,s}\mu^{\text{fwd}}_{s\to i}(r_{i,k})). \end{aligned} \tag{3.5}$$

In all three cases $\alpha^s_i, \alpha^s_{i,i}, \alpha^s_{i,j} \in \mathbb{R}$ are scalars used to calibrate the importance of different messages for a given sentence. These scalars form a dependency structure between images for each of the generated sentence indices. Intuitively, when we generate the first sentence, the attention belief might depend more on subsequent images, to correctly identify the story event, *e.g.*, a wedding, a parade, *etc*. Thus, the scalars will promote interaction with later images. An analysis of these scalars is provided in the Sec. 4.3. Next, we define the different types of messages.

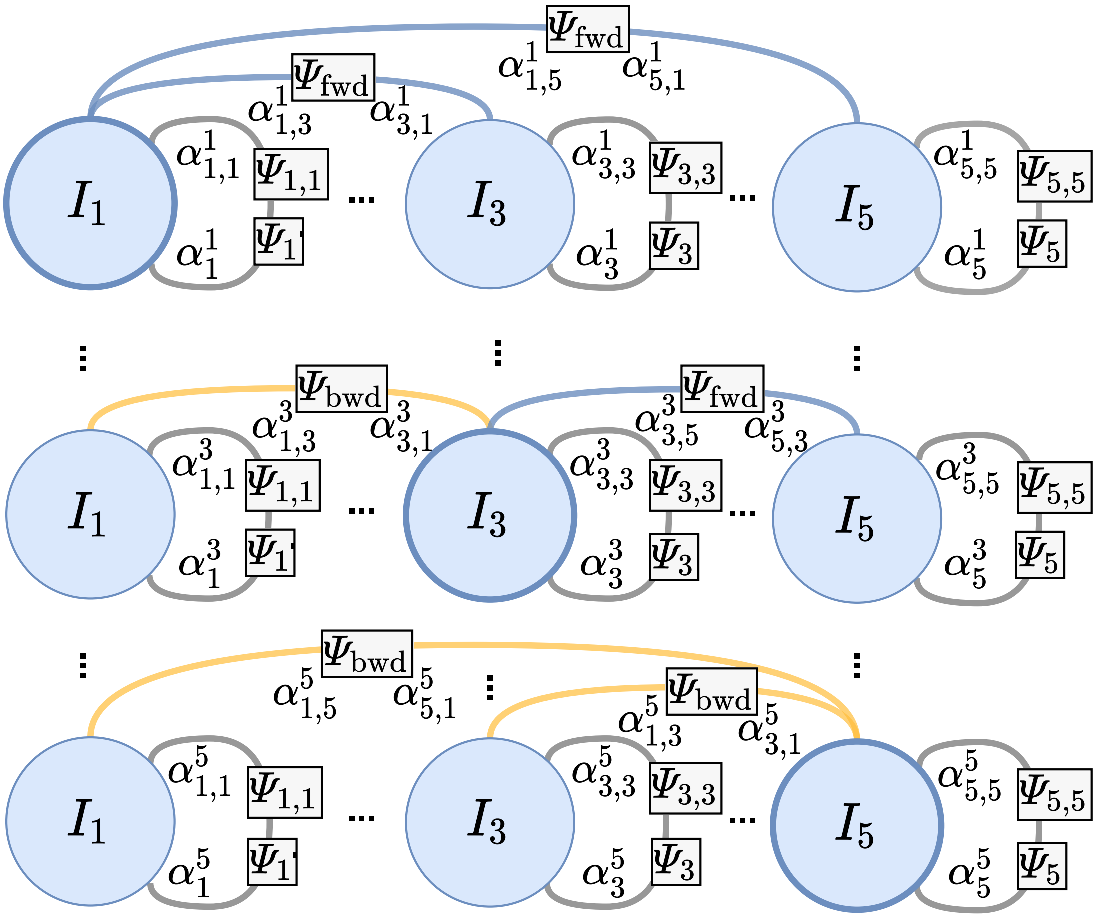


Figure 3.1: Illustration of Ordered Image Attention. Each node represents an image attention belief. For each sentence, we connect all the images with the sentence-corresponding image. The relative position to this image determines whether the connection is modeled with the Ψ_{bwd} factor (for preceding images) or the Ψ_{fwd} factor (for subsequent images). We infer the attention belief by collecting interactions and local object information within the image. We use scalars to calibrate the importance of each factor. In total, we generate 25 attention maps, one per image for every sentence.

3.2.2 Pairwise Messages and Factors

A message aggregates interaction scores from an image to an object. The three messages $\mu^{\text{bwd}}_{j\to i}, \mu^{\text{fwd}}_{j\to i}$ and $\mu_{i\to i}(r_{i,k})$ are computed as follows:

$$\mu^{\text{bwd}}_{j\to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{\text{bwd}}(r_{i,k}, r_{j,k'}), \tag{3.6}$$

$$\mu^{\text{fwd}}_{j\to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{\text{fwd}}(r_{i,k}, r_{j,k'}), \text{ and} \tag{3.7}$$

$$\mu_{i\to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{i,i}(r_{i,k}, r_{i,k'}). \tag{3.8}$$

Importantly, these messages collect three different types of order-dependent interaction factors: (1) A backward image interaction, namely $\Psi_{\text{bwd}}(r_{i,k}, r_{j,k'})$. This interaction models relations to the preceding j-th image in the sequence. (2) A forward image interaction, namely $\Psi_{\text{fwd}}(r_{i,k}, r_{j,k'})$. This interaction models relations to the subsequent j-th image in the sequence. (3) The self interaction factor, namely $\Psi_{i,i}(r_{i,k}, r_{i,k'})$, which takes into account interactions between objects within the image. We formally define the different factors next.

3.2.3 Interaction factors

A commonly used practice to capture interactions across attention mechanisms is to first embed the elements into a joint Euclidean space followed by a dot-product [23; 24; 32; 33]. While we follow the same practice, we define three types of interaction factors to preserve the order. Consider two objects, $r_{i,k} \in I_i$ from the sentence-corresponding image and $r_{j,k'} \in I_j$ from the interacting image. We describe three types of interactions: for interactions with subsequent images (*i.e.*,

$j > i$) we use

$$\Psi_{\text{fwd}}(r_{i,k}, r_{j,k'}) = \left(\frac{L_{\text{fwd}} r_{i,k}}{\|L_{\text{fwd}} r_{i,k}\|_2}\right)^{\top} \left(\frac{R_{\text{fwd}} r_{j,k'}}{\|R_{\text{fwd}} r_{j,k'}\|_2}\right). \tag{3.9}$$

For interactions with preceding images (*i.e.*, $j < i$) we use

$$\Psi_{\text{bwd}}(r_{i,k}, r_{j,k'}) = \left(\frac{L_{\text{bwd}} r_{i,k}}{\|L_{\text{bwd}} r_{i,k}\|_2}\right)^{\top} \left(\frac{R_{\text{bwd}} r_{j,k'}}{\|R_{\text{bwd}} r_{j,k'}\|_2}\right). \tag{3.10}$$

For interactions within the image (*i.e.*, $j = i$) we have

$$\Psi_{i,i}(r_{i,k}, r_{i,k'}) = \left(\frac{L_{i,i} r_{i,k}}{\|L_{i,i} r_{i,k}\|_2}\right)^{\top} \left(\frac{R_{i,i} r_{i,k'}}{\|R_{i,i} r_{i,k'}\|_2}\right). \tag{3.11}$$

Note, $L_{\text{fwd}}, R_{\text{fwd}}, L_{\text{bwd}}, R_{\text{bwd}}, L_{i,i}, R_{i,i} \in \mathbb{R}^{d \times d}$ are trainable shared weights across the entire image sequence. Also, the object from the sentence-corresponding image will always be on the left side of the factor equation. Thus, the factor embeddings preserve the order.

3.2.4 Local factor

Differently from the previous interactions the following factor captures how important an object is based solely on the object representation. Given an object $r_{i,k} \in I_i$, we define the local factor as,

$$\Psi_i(r_{i,k}) = v^{\top} \operatorname{ReLU}(V r_{i,k}), \tag{3.12}$$

where $v \in \mathbb{R}^d, V \in \mathbb{R}^{d \times d}$ are trainable weights.

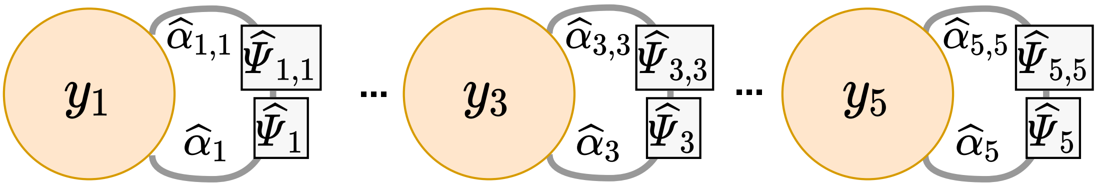


Figure 3.2: Illustration of ISA. The attention selects the attended image representation per sentence. We model interactions between attended images of the same sentence to compute each image's importance. Note, each node represents a sentence attention belief over the attended images.

3.3 Image-Sentence Attention (ISA)

In a next step we summarize the attended image representations a_i^s produced by OIA to compute the context embedding c_s for the sentence s that we wish to generate. For this we use the Image-Sentence Attention (ISA) unit. It picks the relevant image context for generating the specific sentence. Formally we obtain the context embedding via

$$c_s = \sum_{i=1}^{N} \hat{b}_{s,i} a_i^s, \tag{3.13}$$

where attention factors

$$\hat{b}_{s,i} \propto \exp\left(\hat{\alpha}_s \hat{\Psi}_i(a_i^s) + \hat{\alpha}_{s,s}\hat{\mu}_{s\to s}(a_i^s)\right), \tag{3.14}$$

and where $\hat{\alpha}_s, \hat{\alpha}_{s,s} \in \mathbb{R}$ are scalars. To avoid spurious correlations between sentences, we consider only self interactions and a local factor. This is illustrated in Fig. 3.2. The self-message of the attended image representation a_i^s is

$$\hat{\mu}_{s\to s}(a_i^s) = \sum_{j=1}^{N} \hat{\Psi}(a_i^s, a_j^s). \tag{3.15}$$

Finally, the self and local factors are defined with a different set of weights following Eq. (3.11) and Eq. (3.12) respectively.

3.4 Story Decoding

The goal at each timestep of decoding is to compute the conditional probability $p(y_{s,t}|y_{s,t-1}, c_s)$ where $y_{s,t} \in \mathcal{Y}$ is the t-th word in sentence y_s, $\mathcal{Y}$ is the vocabulary and c_s is the context embedding detailed in Sec. 3.3. For this we use a GRU recurrent unit, tasked with generating probabilities over the vocabulary conditioned on the context embedding c_s and the previously generated token $y_{s,t-1}$:

$$\begin{aligned} p(y_{s,t} = w|y_{s,t-1}, c_s) \propto \exp(&\beta_{s,t} \cdot g_w(y_{s,t-1}, h_{s,t-1}, c_s) \\ &+(1-\beta_{s,t}) \cdot f_w(\phi_{s,t})), \end{aligned} \tag{3.16}$$

where g_w is the output of a GRU unit for the word w. We set the GRU hidden dimension to d. $h_{s,t-1} \in \mathbb{R}^d$ is the hidden state at timestep $t-1$ for sentence s. $f : \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}^{|\mathcal{Y}|}$ is a learned prior over the vocabulary based on a bag-of-words prior histogram $\phi_{s,t}$, which we describe in the next paragraph. The purpose of f is to reduce text repetitions. f_w denotes the value of f for a word w. We also incorporate a calibration gate $\beta_{s,t} : \mathbb{R}^d \to [0, 1]$ for functions f and g using

$$\beta_{s,t} = \sigma\left(v_\beta^\top \tanh(G_g h_{s,t} + G_f W_1(\phi_{s,t}))\right). \tag{3.17}$$

Here, $G_g \in \mathbb{R}^{d\times d}$ and $G_f \in \mathbb{R}^{\gamma\times d}$ are trained projections of the GRU hidden state and the bottleneck layer respectively, $v_\beta \in \mathbb{R}^d$ are learned weights and σ is the sigmoid function. W_1 is obtained from the prior as discussed next.

3.4.1 Bag-of-words (BOW) prior

Remembering history during storytelling permits to stay on topic and advance the story in the desired direction. Although quite intuitive, mimicking this ability is not trivial. *E.g.*, most approaches for VST generate all the sentences in parallel. Converting the parallel sentence generation into a sequential one implies a major computational overhead during training.

To address this, we propose a simple yet effective learnable framework that does not require sequential training while still exploiting information found in prior sentences. The history is represented via a bag-of-words histogram $\phi_{s,t}$, which includes all words that have been used until timestep t for the s-th sentence. During training, we initialize $\phi_{s,t=0}$ with the ground truth history counts found in the previous $s-1$ sentences. We update the statistics at each timestep with the predicted word $y_{s',t}$ for $s' < s$, and produce the next state of the counter $\phi_{s,t+1}$. At inference we generate sentences sequentially and update $\phi_{s,t}$ with the predicted words. $\phi_{s,t}$ is fed through a shallow bottleneck network to obtain the prior f, composed of two layers $W_1 \in \mathbb{R}^{|\mathcal{Y}| \times \gamma}$ and $W_2 \in \mathbb{R}^{\gamma \times |\mathcal{Y}|}$ without activation, where γ is the bottleneck dimension:

$$f(\phi_{s,t}) = W_2(W_1(\phi_{s,t})). \tag{3.18}$$

Also note the use of $W_1(\phi_{s,t})$ in the gate (Eq. (3.17)).

3.4.2 Intra-repetition regularization

To regularize intra-repetitions, we decay the probability of previously used words during sentence generation. A critical aspect of this approach is to exclude words that appear frequently in the language (*e.g.*, was, were, am). For this we preprocess the training set to calculate the average story frequency $\rho(w)$ of a word w via $\rho(w) = \frac{\text{\# appearances of word } w}{\text{\# stories } w \text{ was used}}$. The final count for word w at timestep t is calculated as $\phi'_{s,t}(w) = \max[0, (\phi_{s,t}(w) - \rho(w) + 1)]$. Intuitively, a word will not be penalized before it is used more than the prior belief average $\rho(w)$. The final probability for word w being used is given by

$$p'(y_{s,t} = w|y_{s,t-1}, c_s) = \frac{p(y_{s,t} = w|y_{s,t-1}, c_s)}{\pi \cdot \phi'_{s,t}(w) + 1}, \tag{3.19}$$

where $\pi \geq 0$ is a constant hyper-parameter. A penalty of 2 proved to work best on the validation set.

4 Results

4.1 Training Setup

4.1.1 Dataset

To train and test the model we use the VIST dataset [6]. This dataset is composed of stories. Each story has 5 images and $N = 5$ corresponding sentences. All images were collected from Flickr albums. Sequences of images belong to the same album. Each image sequence is annotated with 5 ground-truth reference stories. On average, around 2.5 stories are based on the images, and the rest are rewrites. The overall numbers are 40,098 training stories, 4,988 validation stories, and 5,050 test stories.

4.1.2 Training

We extracted the image features using a pre-trained F-RCNN model with a ResNet152 backbone [2; 4; 34]. We set the number of extracted objects $K = 36$. Bounding box coordinates were normalized between 0 and 1. Words that appear less than 3 times in the training set are represented by an <*UNK*> token. The vocabulary size is 12,210 words. Word representations were initialized using GloVe embeddings [35]. We set the decay parameter $\pi = 2$ and the image representation dimension $d = 512$. We set the dropout parameter to 0.3. We use cross-entropy loss to maximize likelihood of ground-truth stories. At decoding time we employ a beam search algorithm, with beam width set to 3. We use Adam [36] optimizer with a learning-rate of 0.0004, which is decayed by a factor of 0.8 if the validation score (METEOR) does not improve after 4 epochs. The total amount of trainable parameters is 13,092,194. Training converges after ∼20 epochs. Each epoch needs 20 minutes on an Nvidia V100 GPU.

4.2 Quantitative analysis

4.2.1 Evaluation metrics

As suggested by the creators of VIST [6], METEOR [37] correlates best with human judgement. The METEOR metric assesses unigram precision and recall based on matchings between candidates and references. Following their example, we use METEOR as the primary metric. We also compute BLEU [38], which measures the effective overlap between a reference sentence and a candidate sentence. ROUGE [39] (Recall Oriented Understudy of Gisting Evaluation), a recall-based metric that measures the longest common subsequence of tokens, and CIDEr [40] the Consensus-based Image Description Evaluation which measures the similarity of a sentence to the consensus over the test split and compare to prior work where available. For evaluation we use the evaluation script of Yu *et al.* [41][1].

4.2.2 Comparison to state-of-the-art

In Tab. 4.1 we compare the method to recent baselines. Early methods did not take into account visual-spatial information, which harms the performance (*e.g.*,

[1] `http://github.com/lichengunc/vist_eval` - Codebase for commonly used evaluation scripts.

Method	M	B-1	B-2	B-3	B-4	R	C	Img Feat
seq2seq [6]	31.4	-	-	-	3.5	-	6.84	FC
h-attn-rank [41]	33.9	-	29.8	-	-	29.8	7.4	FC
Contextualize, Show & Tell [10]	34.4	60.1	36.5	21.1	12.7	29.2	7.1	FC
AREL [1]	35.0	63.8	39.1	23.2	14.1	29.5	9.4	FC
KnowledgeableStoryteller [29]	35.2	66.4	39.2	23.1	12.8	29.9	**12.1**	FC
HSRL [13]	35.2	-	-	-	12.3	29.5	8.4	Spatial
StoryAnchor [28]	35.5	65.1	40.0	23.4	14.0	30.0	9.9	FC
SGVST [31]	35.8	65.1	40.1	23.8	14.7	29.9	9.8	F-RCNN
Ours - ResNet	36.3	66.3	41.5	23.7	14.5	30.0	9.8	Spatial
Ours - Full	**36.8**±0.1	**68.4**±0.7	**42.7**±0.3	**25.2**±0.2	**15.3**±0.2	**30.2**±0.1	10.1±0.2	F-RCNN

Table 4.1: Quantitative results on the VIST dataset for METEOR, BLEU-1…4, ROUGE-L and CIDEr. The primary metric is METEOR. The 'Img Feat' column describes the pretrained image features. All models utilize a ResNet [2] backbone except CS&T which employs an Inception v3 model [3]. FC and Spatial refer to features extracted from the penultimate layer and the preceding one accordingly. F-RCNN are bottom up features [4].

35.5% *vs.* 36.8% on METEOR) [1; 6; 10]. Wang *et al.* [31] utilize image representations similar to our approach but do not consider relations between different images, resulting in a 1% drop on METEOR, showing that ordered structure encoding with OIA is beneficial. SGVST and StoryAnchor [28; 41] use different methods for mapping the image sequence to distinct topics. Differently, our approach is trained end-to-end. Finally, Yang *et al.* [29] utilize an external commonsense dataset to enrich the input. Their CIDEr score is significantly higher, yet this improvement does not translate to all other metrics. The approach improves upon the current state-of-the-art by a margin (36.8% *vs.* 35.8% on METEOR). Note, the ROUGE-L metric is based on finding the longest subsequence matched to human generated stories. However, this score is almost identical for all prior works, indicating that this metric doesn't capture story generation improvements. We also report the performance with spatial ResNet152 features [2], which outperforms the state-of-the-art as well. This shows that the method is stable irrespective of image features.

4.2.3 Ablation study

In Tab. 4.2 we show the importance of different components via an ablation study. In 'w/o OIA,' we replace the OIA module (Sec. 3.2) with simple averaging of the K object representations of image I_i, resulting in a 0.8% drop on METEOR. Similarly, in 'w/o ISA,' we replace the ISA unit (Sec. 3.3) with averaging, leading to a 0.9% drop on METEOR. In 'w/o attention,' we removed both OIA and ISA, which dropped the METEOR score to 35.8%. For the method referred to as 'no-direction,' we use the same factor for preceding and proceeding interaction (*i.e.*, $L_{\text{bwd}} = L_{\text{fwd}}$ and $R_{\text{bwd}} = R_{\text{fwd}}$). Here, METEOR results drop by 0.7%. Hence, ordered interactions are beneficial. Next, we assess the decoding components (Sec. 3.4). We first remove the intra-repetition regularization (*i.e.*, $\rho(w)$), which causes METEOR score to drop by 0.8%. Removing the popular words count ($\phi'_{s,t}$), results in a 0.7% drop on METEOR. The METEOR score drops by 0.6% when we remove the BOW prior. Next, we evaluate the effect of the decoding strategy for reducing repetitions directly.

Model	METEOR	B-4	#Params
w/o OIA	36.0	14.1	11M
w/o ISA	35.9	14.2	11M
w/o attention	35.8	13.6	11M
no-direction	36.1	14.5	12M
w/o rep. regularization	36.0	14.2	13M
w/o count norm	36.1	14.6	13M
w/o BOW prior	36.2	14.5	13M
Full model	**36.8**	**15.3**	13M

Table 4.2: Components ablation analysis.

In Tab. 4.3, we show the ability to reduce repetitions. As proposed by Bertoldi *et al.* [42], text repetitiveness is measured by the repetition rate of non-singleton n-grams within each story. In our experiment, we use up to 4-grams. The use of intra-repetition regularization reduces text repetition (0.14 to 0.04). Combined with the trainable bag-of-words prior module, we further improve this measure (0.008 *vs*. 0.14). We also report sentence repetitiveness, *i.e.*, the average number of repeated sentences in a story.

Model		**Text Rep.**	**Sent. Rep.**
AREL [1]		0.16	0.4
BOG prior	**Intra-repetition reg.**		
No	No	0.14	0.33
Yes	No	0.10	0.18
No	Yes	0.04	0.04
Yes	Yes	**0.008**	**0.0**

Table 4.3: Story generation ablation analysis.

In Tab. 4.4 we show an ablation of the different factors. We found that each factor contributes to the model's performance, and the directional factors (*i.e.*, Ψ_{fwd} and Ψ_{bwd}) have the biggest impact.

Model			**Metric**						
Local	**Self**	**Directional**	**R**	**C**	**B-1**	**B-2**	**B-3**	**B-4**	**M**
×	✓	✓	30.0	9.3	67.4	42.4	24.2	14.5	36.2
✓	×	✓	29.8	9.2	67.8	42.3	24.2	14.4	36.0
✓	✓	×	29.9	8.5	67.6	42.2	24.0	14.2	35.9
✓	✓	✓	**30.2**	**10.1**	**68.4**	**42.7**	**25.2**	**15.3**	**36.8**

Table 4.4: Factor ablation analysis.

4.3 Factors Importance Analysis

In Fig. 4.1, we illustrate for each sentence, the value of the importance calibration scalars (*i.e.*, α_i^s and $\alpha_{i,s}^s$ in Eq. 2,3, and 4). Intuitively, these values indicate the importance of different image-to-image messages. We focus our analysis on the sentence-corresponding image (*i.e.*, $i = s$ in Sec. 3.1). We observe that the self-message scalars (*i.e.*, $\mu_{i \to i}$) of the sentences in the middle of the sequence, *i.e.*, sentences (2,3, and 4), are low. This indicates that the images in the middle of the sequence rely more on the other images. The beginning and the ending of the story depend more on the local factors. Notably, the most substantial influence is given to the following image (*i.e.*, $\alpha^i_{i,i+1}$). This means that while generating the current sentence, the OIA decision is based mostly on the next image. This is intuitive as it helps to advance the narrative in a desired direction.

4.4 Human Evaluation

The subjective nature of the VST task calls for a human evaluation. We use a sample of 150 image sequences and test different story qualities by asking 3 MTurk annotators to rank or compare them to other methods. We compare our results to the AREL baseline since none of the more recent baselines are publicly available. Note that we also compare coherency against a model without ordered-factors, which already improves upon the prior state-of-the-art.

In Fig. 4.2 we provide the results when asking annotators to pick the most human-like story. We use the majority vote to decide the best model per story. The generated stories outperform the AREL baseline (73.87% *vs.* 22.53%). Surprisingly, in many cases, the annotators found the generated stories to be more human-like than the ground truth stories (41% *vs.* 48.57%).

In Fig. 4.3, we assess coherency. An important aspect of our work are the directional factors for coherency. To validate their effectiveness, we compared to a model that does not incorporate direction into the attention representation (*i.e.*, we use the same factor for preceding and proceeding interactions). The comparison shows a significant coherency improvement (64.2% *vs.* 28.7%). Also, a comparison against the AREL baseline demonstrates a more significant improvement (70.24% *vs.* 25.32%).

To further evaluate the quality of the stories, we follow the criteria set by the

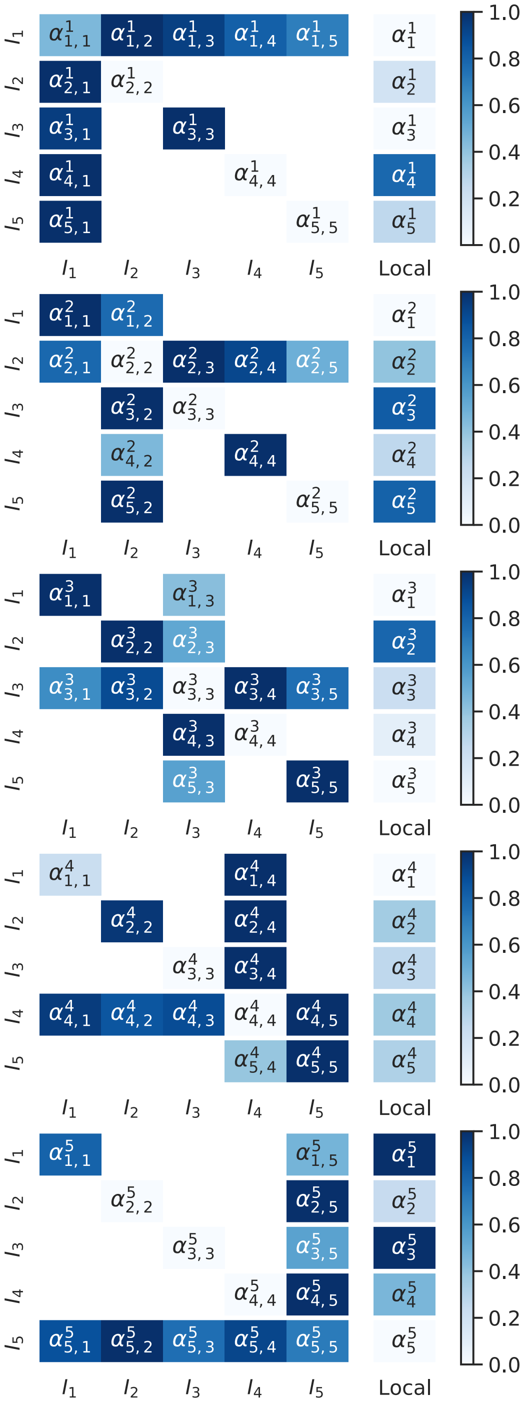


Figure 4.1: OIA scalar values (*i.e.*, α_i^s and $\alpha_{i,s}^s$ in Sec. 3.1.1). The top map corresponds to the first sentence (*i.e.*, $s = 1$) and bottom one to the last sentence (*i.e.*, $s = 5$)

.

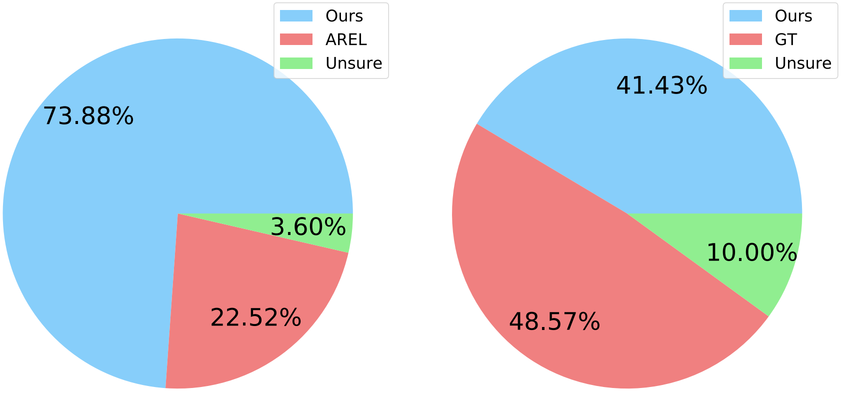


Figure 4.2: Human-like property comparison.

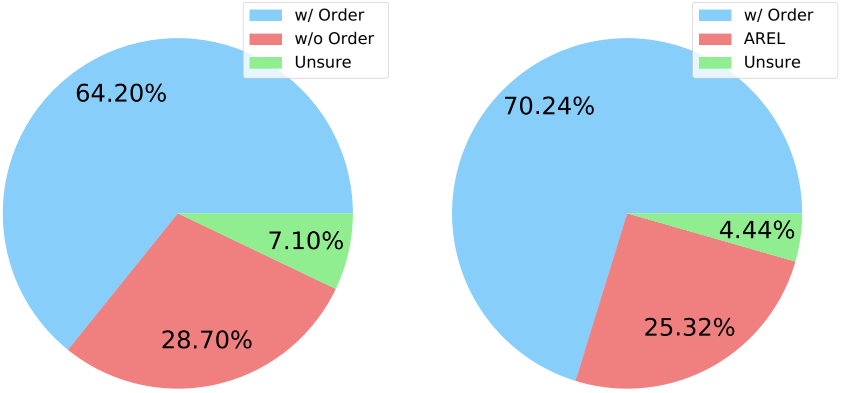


Figure 4.3: Coherence property comparison.

Visual Storytelling Challenge[1] and conduct a survey where judges are asked to rate six categories between 1-5:

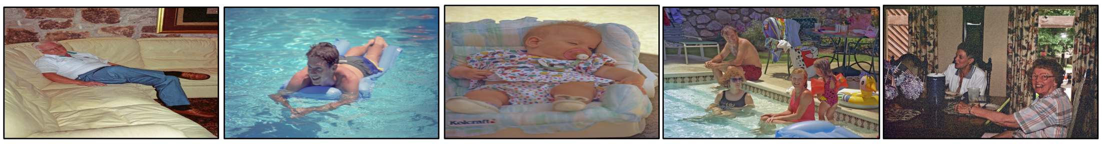
1. *Focused*: the story contains information that is "naturally" relevant to the rest of the story;
2. *Coherence*: the sentences in the story are related and consistent;
3. *Share*: the inclination to share the story;
4. *Human-like*: the story was likely written by a human;
5. *Grounded*: the story directly reflects concrete entities in the image; and
6. *Detailed*: the story provides an appropriate level of detail.

To obtain the final score, we average the annotators' scores per sample, followed by averaging across the entire sample set. From Tab. 4.5 we observe: the model improved on all the criteria compared to the AREL model. Importantly, the generated stories are comparable to the ground-truth stories, indicating success in reducing the shortcomings found in prior methods. Nonetheless, the level of detail is still lacking, supporting the observation of Holtzman *et al.* [43] that current decoding strategies tend to generate well-formed yet somewhat generic text.

Method	Focused	Coherent	Share	Human-like	Grounded	Detailed
AREL	3.49	3.18	3.18	3.26	3.32	3.15
Ours	3.67	3.52	3.20	3.56	3.54	3.32
GT	**3.72**	**3.57**	**3.34**	**3.64**	**3.56**	**3.53**

Table 4.5: Human evaluation results for rating survey (scores are between 1-5).

[1] http://visionandlanguage.net/workshop2018



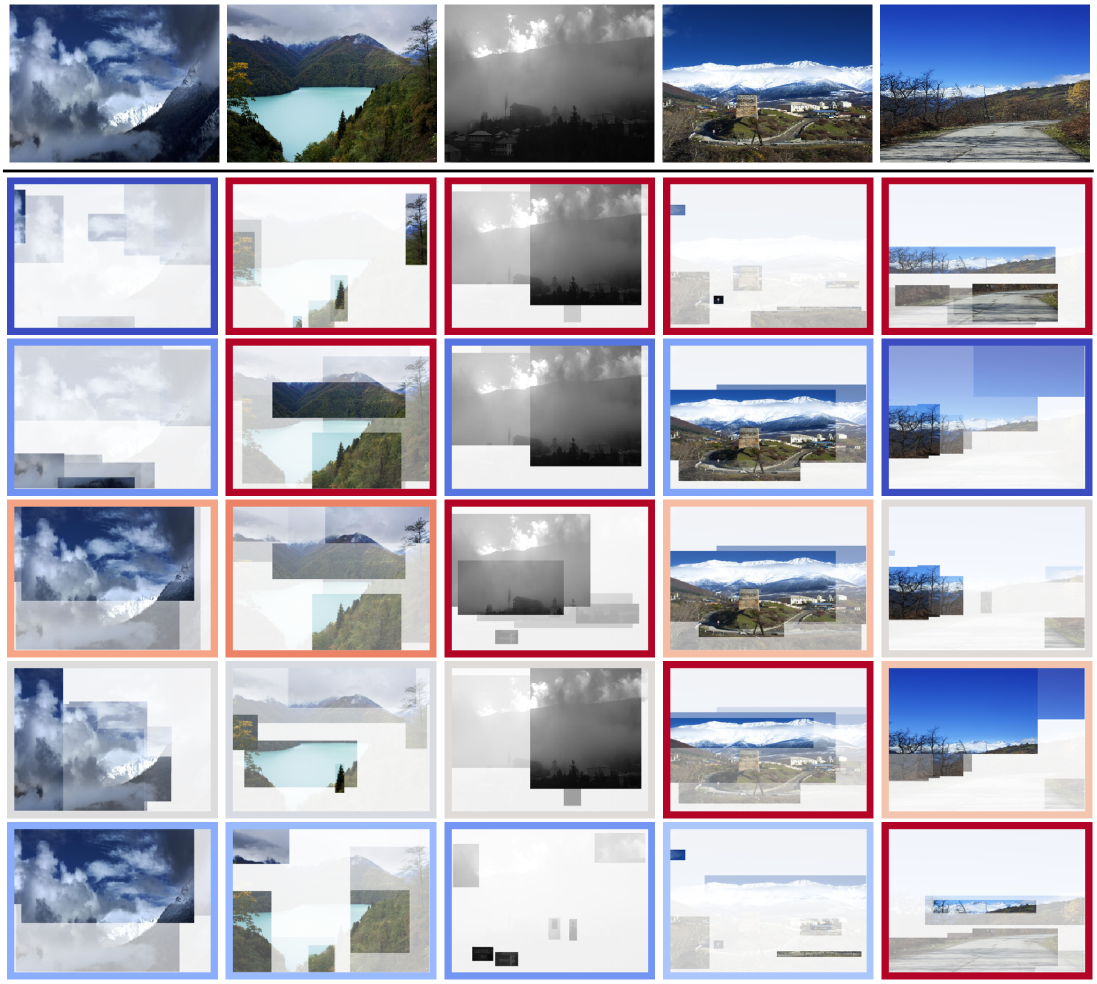
AREL	The kids had a great time at the pool. The little boy was excited to see the kids. We had a great time at the park. We had a great time at the pool. We had a great time at the park.
No History	The kids had a great time at the beach. The baby was happy to see the baby. We had a great time at the park. The had a great time at the pool. We had a great time at the park.
With History	**The family went to the pool. The baby was very happy. The kids had a great time. The kids played in the pool. The little girl is having a good time.**

Figure 4.4: An illustration of an image sequence along with three different stories generated by: (1) AREL baseline [1], (2) No History: a model without intra-repetition regularization and BOW prior (see Sec. 3.4); and (3) With History: the final model. Repeated sentences are highlighted with a yellow colored marker. Repeated words in a sentence are emphasized in red color.

4.5 Qualitative evaluation

In Fig. 4.4, we show the ability of the method in reducing repetitions. We observe the AREL baseline to repeat the same sentences, for example, "...had a great time at...". We also observe this repetitiveness when we remove the bag-of-words prior and the intra-sentence regularization (*i.e.*, No History column). Nevertheless, the method remains on topic, *i.e.*, family in the pool.

In Fig. 4.5 we sketch the attention maps along with the generated story. The first sentence, "We went to the mountains," sets the theme for the story, which requires the processing of subsequent images. Notably, the ISA module picked the proceeding images. In contrast, for the second sentence, the attention focuses mostly on the second image resulting in a description of the lake observed exclusively in this image. The third sentence relates to the scenery. Hence the attention focuses on preceding and proceeding images.



Ground Truth	The clouds compliment the mountain peak. They find a lovely forested mountain with a lake. The misty clouds roll in and obscure the scene. The height of the mountains can be seen by the snow covering them. On the road again moving towards another place.
Ours	**We went to the mountains for a hike. The view of the lake was amazing. The scenery was breathtaking. We saw some old buildings. The view of the mountain was spectacular.**

Figure 4.5: Illustration of OIA and ISA attention maps, the ground-truth story and the final generated story. Each row corresponds to a story sentence and shows objects OIA highlights. The attended images' border specifies the relevancy to sentence generation, from red (important) to blue (not important).

5 Discussion

5.1 Metrics

Different evaluation metrics have been used to assess the quality of generated stories. In their introduction of the Visual Storytelling task and dataset [6], the authors assess the correlation between the two metrics (BLEU and METEOR) and human judgements. These metrics were originally developed to measure the quality of human translation. As noted in Hu *et al.* [44], these metrics compare the n-gram overlap between the reference sentence and the generated sentence. As such, they treat each word in the sentence equally, without considering the semantic relevance of the words to the image. For storytelling semantics are very important, two sequences can be completely different in language yet have the very high semantic similarity. For example, the sentences “The bride and groom are ready...” (e.g. reference) and “The wedding was about to begin” (e.g. generated sentence) have a low overlap yet are semantically very similar. Intuitively, the generated sentence should obtain a high score compared to the reference, yet BLEU and METEOR fail to capture the similarity. Hu et al. chose METEOR as the automatic metric as it has a higher correlation with human judgement than BLEU (ρ=0.2 vs. ρ=0.08), yet the low correlation of both is a clear indicator to their shortcoming in assessing the overall performance. Later

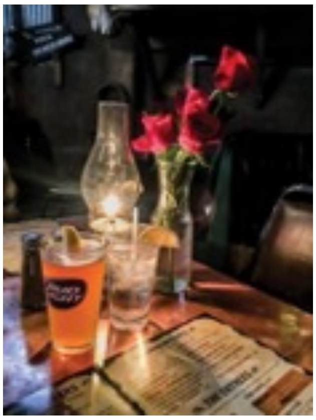


Figure 5.1: Example of image for metrics discussion

works evaluate the output based on CIDEr (Consensus-based Image Description Evaluation) [40] as well. CIDEr is the prominent metric used for assessing the quality of image caption, where the goal is to describe the image. The metric assumes image description is an objective task and as such both reference and generated text should concur on the main concepts to describe in an image. For example in Fig. 5.1, most would agree that given the following image, a correct description should include concepts such as drinks, glass, flowers, table, menu etc. The sentence “Glasses with drinks and a vase with flowers on top of a table” would therefore achieve a high score in CIDEr. Though highly descriptive, this plain description is unlikely to begin a story based on the image. A more fitting beginning to the story could be ”My wife and I went out to eat” yet this sentence would score very low in CIDEr as it does not describe any of the objects in the image. To address the shortcoming described above, many works, including ours, perform a qualitative analysis. Though rigorous qualitative analysis is a good estimator for the performance of the model, it is unscalable and more importantly harms the ability to compare different research approaches. It is evident from the critique above that a better automatic metric is necessary to benchmark model performance. This theoretical metric should be able to capture the semantics of the generated text as well as the relation (again semantic) between the image and the generated text. Although some work has been performed to define such metrics, it is still an open question, currently being researched.

We present a novel approach for VST, which encourages coherency of generated story. We incorporate structure between images with a new attention method that selects the important objects in an ordered image sequence. Human evaluation and quantitative analysis demonstrate that the approach outperforms existing methods. Further, we perform ablation and qualitative analysis to show effectiveness.

5.2 Future Work

In this work we focused on incorporating order into an attention mechanism and investigating its effect by employing simple models. We did not experiment with large transformer-based [32] models. This type of models have achieved great success recently in vision+language tasks such as captioning and VQA yet have not been tested for Visual Storytelling. Although the VIST dataset is a great initial step towards storytelling, there are many shortcomings in it. First, many sequences are not true stories, for instance, 5 images of fireworks. For these types of sequences, the task of storytelling is hard for humans as well. Consequently, we suggest that filtering out such sequences could benefit future models. Another limitation of the dataset is the creative freedom given to the judges. This freedom leads to widely different stories between judges given the same sequence, many times lacking any common patterns. We believe that constraining the task per judge with specific topics, themes (i.e. positive story, sad story), points of view (i.e. first/third person) or other constrains, could benefit the dataset generation as it will improve the consistency between different judges. Video storytelling is extremely similar to VST, as such, we believe that many of the lessons learned in this research could be applied to the task given a high-quality dataset. We are not aware of such dataset at the time of this publication.

5.3 Conclusions

We present a novel approach for VST, which encourages coherency of generated story. We incorporate structure between images with a new attention method that selects the important objects in an ordered image sequence. Human evaluation and quantitative analysis demonstrate that the approach outperforms existing methods. Further, we perform ablation and qualitative analysis to show effectiveness.

5 References

[1] X. Wang, W. Chen, Y. fang Wang, and W. Y. Wang, "No metrics are perfect: Adversarial reward learning for visual storytelling," in *ACL*, 2018. vi, 2, 6, 20, 22, 27

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2015. viii, 8, 18, 20

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2015. viii, 20

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2017. viii, 4, 8, 18, 20

[5] C. C. Park and G. Kim, "Expressing an image stream with a sequence of natural sentences," in *NeurIPS*, 2015. 1

[6] T.-H. 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*, 2016. 1, 3, 6, 18, 19, 20, 29

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2014. 1, 6

[8] 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*, 2015. 1, 4, 5

[9] X. Chen and C. L. Zitnick, “Mind’s eye: A recurrent visual representation for image caption generation,” in *CVPR*, 2015. 1

[10] D. Gonzalez-Rico and G. F. Pineda, “Contextualize, show and tell: A neural visual storyteller,” in *Storytelling Workshop, NAACL*, 2018. 1, 6, 20

[11] B. Wang, L. Ma, W. Zhang, W. Jiang, and F. Zhang, “Hierarchical photo-scene encoder for album storytelling,” in *AAAI*, 2019. 1

[12] Y. Modi and N. Parde, “The steep road to happily ever after: an analysis of current visual storytelling models,” in *Workshop on Shortcomings in Vision and Language, NAACL*, 2019. 1

[13] Q. Huang, Z. Gan, A. Çelikyilmaz, D. Wu, J. Wang, and X. He, “Hierarchically structured reinforcement learning for topically coherent visual story generation,” in *AAAI*, 2018. 2, 6, 20

[14] K. Barnard, P. D. Sahin, D. A. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, “Matching words and pictures,” *JMLR*, 2003. 4

[15] H. Xu and K. Saenko, “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering,” in *ECCV*, 2016. 5

[16] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, “Stacked attention networks for image question answering,” 2015. 5

[17] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *EMNLP*, 2016. 5

[18] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *ICLR*, 2017. 5

[19] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *ICCV*, 2017. 5

[20] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," in *NeurIPS*, 2018. 5

[21] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NeurIPS*, 2016. 5

[22] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018. 5

[23] I. Schwartz, A. G. Schwing, and T. Hazan, "High-order attention models for visual question answering," in *NeurIPS*, 2017. 5, 12

[24] I. Schwartz, S. Yu, T. Hazan, and A. G. Schwing, "Factor graph attention," in *CVPR*, 2019. 5, 12

[25] T. Kim, M.-O. Heo, S. Son, K.-W. Park, and B.-T. Zhang, "Glac net: Glocal attention cascading networks for multi-image cued story generation," in *CoRR*, 2018. 6

[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NeurIPS*, 2014. 6

[27] J. Li, H. Shi, S. Tang, F. Wu, and Y. Zhuang, "Informative visual storytelling with cross-modal rules," in *MM*, 2019. 6

[28] B. Zhang, H. Hu, and F. Sha, “Visual storytelling via predicting anchor word embeddings in the stories,” in *ICCV*, Workshop on Closing the Loop Between Vision and Language, 2020. 6, 20

[29] P. Yang, F. Luo, P. Chen, L. Li, Z. Yin, X. He, and X. Sun, “Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling,” in *IJCAI*, 2019. 6, 20

[30] C.-C. Hsu, Z.-Y. Chen, C.-Y. Hsu, C.-C. Li, T.-Y. Lin, T.-H. Huang, and L.-W. Ku, “Knowledge-enriched visual storytelling,” in *AAAI*, 2020. 6

[31] R. Wang, Z. Wei, P. Li, Q. Zhang, and X. Huang, “Storytelling from an image stream using scene graphs,” in *AAAI*, 2019. 6, 20

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017. 12, 31

[33] P. Gao, Z. Jiang, H. You, P. Lu, S. C. Hoi, X. Wang, and H. Li, “Dynamic fusion with intra-and inter-modality attention flow for visual question answering,” in *CVPR*, 2019. 12

[34] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *IEEE PAMI*, 2015. 18

[35] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*, pp. 1532–1543, 2014. 18

[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *CoRR*, 2014. 19

[37] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *ACL*, 2005. 19

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *ACL*, 2001. 19

[39] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *ACL*, 2004. 19

[40] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *CVPR*, 2014. 19, 30

[41] L. Yu, M. Bansal, and T. L. Berg, “Hierarchically-attentive rnn for album summarization and storytelling,” in *EMNLP*, 2017. 19, 20

[42] N. Bertoldi, M. Cettolo, and M. Federico, “Cache-based online adaptation for machine translation enhanced computer assisted translation,” in *MT Summit*, 2013. 21

[43] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” in *ICLR*, 2020. 26

[44] J. Hu, Y. Cheng, Z. Gan, J. Liu, J. Gao, and G. Neubig, “What makes a good story? designing composite rewards for visual storytelling,” in *AAAI*, vol. 34, pp. 7969–7976, 2020. 29

[45] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *ICCV*, 2015.

[46] J. Wang, J. Fu, J. Tang, Z. Li, and T. Mei, “Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training,” in *AAAI*, 2018.

[47] 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,” in *ArXiv*, 2015.

[48] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in *IEEE Trans. Signal Processing*, vol. 45, 1997.

[49] K. Cho, B. van Merrienboer, Çaglar Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[50] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP*, 2015.

[51] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *CVPR*, 2018.

[52] H. Schwenk, L. Barrault, A. Conneau, and Y. LeCun, "Very deep convolutional networks for text classification," in *EACL*, 2016.

[53] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, 2018.

[54] H. AlAmri, V. Cartillier, A. Das, J. Wang, S. Lee, P. Anderson, I. Essa, D. Parikh, D. Batra, A. Cherian, T. K. Marks, and C. Hori, "Audio-visual scene-aware dialog," in *ArXiv*, vol. abs/1901.09107, 2019.

[55] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural Computation*, 1997.

[56] A. Karpathy and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[57] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *CVPR)*, 2016.

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

[59] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *CoRR*, 2014.

[60] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *ICCV*, 2015.

[61] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.

תקציר 6

בעבודה זו נעסוק בבעיית visual storytelling, בה מיוצר סיפור המבוסס על רצף תמונות. על כל כל משפט בסיפור יש לתאר את התמונה במתאימה, בנוסף סיפור קוהרנטי צריך להיות עקבי ולהתייחס הן לתמונות עתידיות והן לתמונות מהעבר. על מנת להשיג מטרה זו, אני מציעים מנגנון Ordered Image Attention (OIA). OIA ממדל קשרים בין התמונה התואמת למשפט לבין אזורים חשובים משאר התמונות ברצף. בכדי להדגים אובייקטים חשובים בתמונה, אנו מציעים אלגוריתם הדומה להעברת מסרים אשר מאגד ייצוגים כאלו לאובייקטים באופן אשר מודע למיקום התמונה ברצף. בכדי ליצור את משפטי הסיפור, אנו מציעים מנגנון נוסף אשר מייצר ייצוגים ווקטריים בשם Image Sentence Attention (ISA). בנוסף, בכדי למנוע שגיאות לשוניות נפוצות כגון חזרתיות וחוסר קוהרנטיות, אני מציגים prior אדפטיבי. התוצאות המתקבלות משפרות ב1% את התוצאות הקיימות. בכדי לאשש את תוצאותינו אנו גם במצעים מחקר איכותני מקיף וכך מראים שאכן הסיפורים הנוצרים ע"י הפתרון המוצע הם יותר מפוקסים וקוהרנטיים.

המרכז הבינתחומי בהרצליה

בית-ספר אפי ארזי למדעי המחשב
התכנית לתואר שני (M.Sc.) - מסלול מחקרי

תשומת לב עם סדר תמונות עבור סיפור סיפורים וויזואלי

מאת
תום בראודה

עבודת תזה המוגשת כחלק מהדרישות לשם קבלת תואר מוסמך M.Sc.
במסלול המחקרי בבית ספר אפי ארזי למדעי המחשב, המרכז הבינתחומי הרצליה

מאי 2021

עבודה זו בוצעה בהדרכתו של פרופ' אריאל שמיר, בית ספר אפי ארזי למדעי המחשב,
המרכז הבינתחומי, הרצליה.