Abstract

In visual storytelling, a short story is generated based on a given image sequence. Despite years of work, most visual storytelling models remain limited in terms of the generated stories’ fixed length: most models produce stories with exactly five sentences because five sentence stories dominate the training data. The fix-length stories carry limited details and provide ambiguous textual information to the readers. Therefore, we propose to “stretch” the stories, which create the potential to present in-depth visual details. This paper presents Stretch-VST, a visual storytelling framework that enables the generation of prolonged stories by adding appropriate knowledge, which is selected by the proposed scoring function. We propose a length-controlled Transformer to generate long stories. This model introduces novel positional encoding methods to maintain story quality with lengthy inputs. Experiments confirm that long stories are generated without deteriorating the quality. The human evaluation further shows that Stretch-VST can provide better focus and detail when stories are prolonged compared to the state of the art. The demo video is available on Youtube, and the live demo can be found on website.

1 Introduction

Visual storytelling (VIST) is an interdisciplinary task that takes a sequence of photos as input and produces a corresponding short story as output (Huang et al., 2016). Prior work explores either end-to-end or hierarchical methods for visual storytelling, but machine-generated stories still fall far short of human-generated stories. One obvious limitation is the inability to generate stories with diverse length, especially to prolong a story. In real-world applications, when pictures accompany textual stories, the number of sentences is often much greater than the number of images. Recent visual storytelling frameworks demonstrate the potential in prolonging visual stories, such as KG-Story (Hsu et al., 2020), a state-of-the-art framework that uses a knowledge graph to generate one additional sentence and attach it to 5-sentence visual stories for improved coherence. However, current models, including KG-Story, are incapable of further “stretching” stories beyond five or six sentences. In short, generating prolonged visual stories faces three main hurdles: First, as VIST—the only existing visual storytelling dataset—is mostly constructed as 5-photo sequences paired with 5-sentence stories, models trained on it easily overfit to the dominant length. Second, in visual storytelling, the quality of the textual story must be maintained when asking the model for more context. Third, the model’s generation function must generate stories with the desired number of sentences. That is, control of the continuation and termination of natural language generation depends on a given length factor.

To meet these challenges, we introduce Stretch-VST, a modification of the KG-Story framework that greatly increases the number of sentences in visual stories while maintaining the quality thereof. Story coherence and detail are improved by using cohesive and relevant information to generate additional sentences. Illustrated in Fig. 1, Stretch-VST has three main stages: First, it extracts representative terms (e.g., actions or objects) from each image. Second, it finds relations between consecutive images using a knowledge graph, after which a scoring model selects the most suitable subset of terms (“term set” hereafter) given its length, term semantics, and cohesion. The length of the term set for the resultant term sequence hence depends...
Figure 1: Stretch-VST extracts representative key terms (e.g., objects, people, and actions) from each image, and uses knowledge graphs to further expand the term set. For any arbitrary subset of terms, Stretch-VST can generate a story for it: the longer the term set, the longer the output story. The framework generates stories from 5 to 9 sentences long, and selects the best story with the lowest term perplexity (PPL score).

3 Methodology

With variable-length visual storytelling, Stretch-VST brings two major contributions for VIST: enriching the ingredients as desired (Sect. 3.1) and enabling story generation according to the term sequence length (Sect. 3.2).

3.1 Expanding and Scoring Term Sequences

Prolonging Term Sequences Drawing from KG-Story (Hsu et al., 2020), we utilize their Transformer-based model to distill the representative terms (e.g., nouns and frames) for each image. Stretch-VST manipulates term sequence lengths to increase the story lengths. For every two consecutive images, we choose whether to insert a relation into the term sequence; hence, the sequence length ranges from 5 to 9, as illustrated in Fig. 1. Given 5 images, we define the image-extracted original term sequence as \{m_1, ..., m_t, ..., m_5\}, where \{m_1, ..., m_{N_1}\} denotes first image’s term set, \(m_t\) denotes the i-th term from image t and \(N_k\) is the number of terms from image k. From consecutive images, we explore all possible relations \((m_t, r, m_{t+1})\) and \((m_t, r_1, m_{middle}, r_2, m_{t+1})\), where \(m_{middle}\) denotes a knowledge graph entity that bridges \(m_t\) and \(m_{t+1}\). The chosen relation is inserted into the original term sequence. For every 5 term sets generated from the images, the model can insert an additional 0 to 4 term sets, resulting in 5 to 9 term sets in total. Moreover, if no relation can be found between two consecutive images, we also attempt to find a relation in the reverse direction, as well on the score. Finally, a length-controlled Transformer is used to generate the story given the term sequence.

The proposed work generates a variable number of sentences, and finds the optimal subset of terms given the story length. The human evaluation shows that Stretch-VST generates better stories when prolonging stories, provides more detailed information comparing 5-sentence stories, and is more robust in cohering story context when the images are incoherent.
as relations between cross images. That is, we include \((m_i^{t+1}, r, m_j^t), (m_i^t, r, m_j^{t+1})\), and also these for two-hop relations. Furthermore, we also applied an image-grounded relation filtering, which is to ensure the predicted terms appear in the image. This prevents the model from generating irrelevant terms. Note that KG-Story is unable to expand or manipulate the size of the term set, and can only produce 6-sentence stories.

**Rating Prolonged Term Sequences** We implement a Transformer with a masked language model objective (Devlin et al., 2019). We use spaCy\(^3\), Open Sesame (Swayamdipta et al., 2017), and the FrameNet parser (Baker et al., 1998) to convert the story text to term sequences. We iteratively mask one position in the overall term sequence to train the Transformer model. Then, for every possible term, we calculate the average perplexity of it with a mask at each position. The term sequence with the best (lowest) average perplexity is used in the next stage to generate stories as

\[
P(m') = F(m'_1, \ldots, m'_N), \quad PPL(m') = P(m')^{-\frac{1}{N}}, \quad score = \frac{1}{N_m} \sum_{i=1}^{N_m} PPL(m_i),
\]

where \(m'\) is the masked term, \(N_M\) is the number of term sets, \(N_m\) is the number of terms in the sequence, \(F\) is the Transformer language model, and \(PPL\) denotes perplexity.

**3.2 Generating Stories From Term Sequences**

Most story generation models generate only 5-sentence stories, regardless of the input length; story quality usually decays when generating longer stories (Guo et al., 2018). To this end, we propose a length-controlled Transformer model structure with unique positional encoding and history embedding to reflect the prolonged input length, prevent story decay, and maintain topic coherence. The model flowchart is shown in Fig. 2.

**Length-Controlled Transformer** To generate a story depending on the term sequence length, a Transformer (Vaswani et al., 2017) is used as a next-sentence generator to generate a story sentence by sentence. Generating sentence \(s_x\), the model is given a history embedding \(H_c\) and all images’ term sets \(M^1, \ldots, M^{N_M}\), where \(H_x = LSTM(s_0, \ldots, s_{x-1})\), denotes a history embedding for all previous sentences, generated from a LSTM layer; \(M^t = \{m^t_1, \ldots, m^t_{N_m}\}\) denotes the set of \(N_m\) terms belonging to image \(t\). Given an expanded term sequence with \(N_M\) term sets, the model generates \(N_M\) times to obtain a story consisting of \(N_M\) sentences.

**Positional Encoding** In 5-sentence VIST training dataset, most stories only contain sentence position up to 5. When generating such stories, naive absolute positional encoding (Vaswani et al., 2017) doesn’t handle positions larger than 5, thus, story quality decays accordingly. To this end, we introduce term positional encoding and beginning-inside-ending (BIE) positional encoding to reflect diverse input lengths. Term positional encoding is implemented in the Transformer encoder to inform the model of the current term position. While generating sentence \(x\), the model sets input term set \(M^x\)’s position to 1 and masks \(M^1, \ldots, M^{x-1}, M^{x+1}, \ldots, M^{N_M}\) as 0. In addition, BIE positional encoding is implemented in the Transformer decoder to focus on the beginning and the end of the story while generalizing the sentences in between. Specifically, we assign position 1 and 3 to the first and last sentence, and position 2 to the sentences in the middle.

**4 System Interface**

Fig. 3 illustrates the user interface of Stretch-VST. We create a webpage for users to (A) search a story by story ID or (B) search for stories by keyword. In Fig. 4(a), our user interface displays five images of the selected album and the visual story with recommended length generated by Stretch-

\(^3\text{SpaCy: https://spacy.io/}\)
Figure 3: User interface of Stretch-VST. User can (A) select an story ID from the drop-down menu or (B) search a stories by keywords.

Figure 4: (a) The panel will show 5 images and visual story with recommended length. User can (b) drag the bar-slider and select the desired length of visual story.

VST. The recommended story length is decided by our scoring model (Sect. 3.1). Users can also drag the bar-slider to select the desired length of visual story (Fig. 4(b)). For the keyword search, the user interface displays several images and story snippets for search results, and the searching algorithm is an elastic search (Fig. 5(a)). Likewise, the panel will display the images, visual story, and the recommended story length (Fig. 5(b)), and users can also select the desired story length.

5 Experimental Results

5.1 Evaluation Methods and Baselines

Per the literature (Wang et al., 2018a), human evaluation is the most reliable way to evaluate the quality of visual stories; automatic metrics often do not align faithfully to human judgment (Hsu et al., 2019). Therefore, we conducted human evaluations to assess the quality of stories generated by Stretch-VST. We randomly selected 250 stories and evaluated each by five different workers on Amazon Mechanical Turk. Each worker was presented with the image sequence and its corresponding stories generated by different models and asked to rank the stories. In addition, we also conduct a questionnaire asking annotators “what makes the story better”, based on the 6 criteria set by VIST dataset (Huang et al., 2016). These criteria include focus, coherence, shareability, humanness, grounding, and detail. We used the same datasets and knowledge graphs as Hsu et al. (2020), and compared the proposed method with three baselines for visual storytelling: AREL (Wang et al., 2018a), GLAC (Kim et al., 2018), and KG-Story (Hsu et al., 2020). Note that we did not compare the results with KG-Story in Sect. 5.3 and 5.4, as its generation model neither handles diverse inputs nor controls the length.

5.2 Generating Optimal-Length Stories

First, we evaluate the ability of Stretch-VST to generate better and longer stories. Given 5 candidate sequences with distinct lengths from 5 to 9, we
selected the best sequence of terms with the lowest perplexity as the material to tell the visual story, as described in Sect. 3.1. The resulting average number of sentences in the generated stories was 6.22; that is, the proposed model tends to add one or two relations to enrich the original story.

The average ranking results, shown in the first row of Table 1 are better than baseline models. This indicates the proposed stories are superior to those from the baseline. Figure 6 shows the questionnaire result for the best-ranked stories. For Stretch-VST and KG-Story’s best-ranked stories, the Stretch-VST story counts are generally higher in all aspects; specifically, Detailed, Coherence, and Focused are significantly higher. As our stories contain more sentences than KGStory, the stories are undoubtedly more detailed. Additionally, the increase of stories’ coherence indicates the advantage of our multiple term set insertion as compared to KGStory’s single insertion. The stretching stories are beneficial to detailed and coherence, we also found that story prolongation is beneficial to topic-focus. We presume the increase number of relevant sentences can improve the focus. Note that we did not use automatic metrics for evaluation because these metrics do not indicate the quality of visual stories (Wang et al., 2018; Hsu et al., 2019).

5.3 Robustness to Incoherent Images

Next, we evaluated the robustness of the proposed method story coherence by deleting the second and fourth of the five input images. The second column of Table 1 shows that Stretch-VST brings together the diverse contents to generate the best story context even when the input is disrupted. Figure 7(b) is an example of such input disruption. Although removing two images creates an incoherence in the photo sequence, Stretch-VST makes the best of the knowledge graph to fill this gap and generate a coherent story.

5.4 Robustness to Overstretched Stories

Without changing the input image sequences, does forcing a model to generate longer stories decrease the story quality? As no existing method generates longer visual stories with a fixed number of input images, we selected a strong Transformer baseline that incorporates the length-controlling mechanism proposed in (Kikuchi et al., 2016) as a baseline for comparison. The baseline model takes the term sequence and the desired length as the encoder input. After forwarding the encoder output to the decoder, we obtain the baseline story from the decoder’s output. The result in Fig. 8 shows that Stretch-VST is

Table 1: Average rankings (1 to 4, lower is better) and number 1st ranked stories (larger is better) rated by human judges, along with average number of sentences and tokens per story. ($\rho$ value < 0.05, N=250)
better at generating longer sentence story than our baseline model.

6 Conclusion

We propose a novel method for generating length-controlled visual stories which includes an enhanced knowledge-graph reasoning module and a length-controlled Transformer architecture. Using human evaluations, we show that the method tells longer and better stories.

7 Ethical Considerations

Although our research aims to produce stories that are vivid, engaging, and innocent, we are aware of the possibilities of utilizing a similar approach to generate inappropriate text (e.g., violent, racial, or gender-insensitive stories). The proposed visual storytelling technology enables people to generate stories rapidly based on photo sequences at scale, which could also be used with malicious intent, for example, to concoct fake stories using real images. Finally, as the proposed methods use external knowledge graphs, they reflect the issues, risks, and biases of such information sources. Mitigating these potential risks will require continued research.

8 Acknowledgements

This research is supported by Ministry of Science and Technology, Taiwan under the project contract 108-2221-E-001-012-MY3 and 108-2923-E-001-001-MY2 and the Seed Grant from the College of Information Sciences and Technology (IST), Pennsylvania State University. We also thank the crowd workers for participating in this project.

References


Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In EMNLP.


Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In AAAI.


Peter LoBue and A. Yates. 2011. Types of commonsense knowledge needed for recognizing textual entailment. In ACL.


Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. In NAACL-HLT.


Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and X. Huang. 2020. Storytelling from an image stream using scene graphs. In AAAI.


Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, J. Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In IJCAI.