Stretch-VST: Getting Flexible With Visual Stories

Chi-Yang Hsu[1,3]*, **Yun-Wei Chu**[2]*, **Tsai-Lun Yang**[3],
Ting-Hao (Kenneth) Huang[1], **Lun-Wei Ku**[3]
Pennsylvania State University [1], Purdue University [2],
Institute of Information Science, Academia Sinica[3]
{cxh5438, txh710}@psu.edu
{chu198}@purdue.edu
{a7532ariel, lwku}@iis.sinica.edu.tw

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[1], and the live demo can be found on website[2].

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

* denotes equal contribution

[1]Demo video: https://youtu.be/-uF8IV6T1NU

[2]Live demo website: https://doraemon.iis.sinica.edu.tw/acldemo/index.html

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).

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.

2 Related Work

Visual storytelling was proposed by Huang et al. (2016). Two lines of work explore this task: one focuses on model architecture for better story generation (Hsu et al., 2018; Gonzalez-Rico and Pineda, 2018; Kim et al., 2018; Huang et al., 2019; Jung et al., 2020; Wang et al., 2020), and the other uses adversarial training to generate more diverse stories (Chen et al., 2017; Wang et al., 2018a,b; Hu et al., 2020). However, these methods often overfit to the number of sentences in the stories. Stretch-VST modifies both the source and generation modules to generate variable-length stories. On the source side, we use knowledge graphs to expand the term set to represent the input image sequence. Integrating a knowledge graph into language generation is beneficial (LoBue and Yates, 2011; Bowman et al., 2015; Hayashi et al., 2020; Zhang et al., 2017; Zhou et al., 2018; Yang et al., 2019; Guan et al., 2019). On the generation side, some explore the use of relative positional encoding (Takase and Okazaki, 2019), adding embedding layers, and manipulating the beam search process (Kikuchi et al., 2016). However, these methods control only the number of words and not the number of sentences.

3 Methodology

With variable-length visual sorytelling, 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 $\{m_1^1, ..., m_i^t, ..., m_{N_5}^5\}$, where $\{m_1^1, ..., m_{N_1}^1\}$ denotes first image's term set, m_i^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_i^t, r, m_j^{t+1}) and $(m_i^t, r_1, m_{middle}, r_2, m_j^{t+1})$, where m_{middle} denotes a knowledge graph entity that bridges m_i^t and m_j^{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

as relations between cross images. That is, we include (m_i^{t+1}, r, m_j^t), (m_i^t, r, m_j^{t+n}), 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 generate 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

$$\mathrm{P}(m') = \mathrm{F}(m'|m_1^1, ..., m_{N_m}^{N_M}), \quad (1)$$

$$\mathrm{PPL}(m') = \mathrm{P}(m')^{-\frac{1}{N_m}}, \quad (2)$$

$$score = \frac{1}{N_m}\sum_{i=1}^{N_m}\mathrm{PPL}(m_i), \quad (3)$$

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_x and all

[3] SpaCy: https://spacy.io/

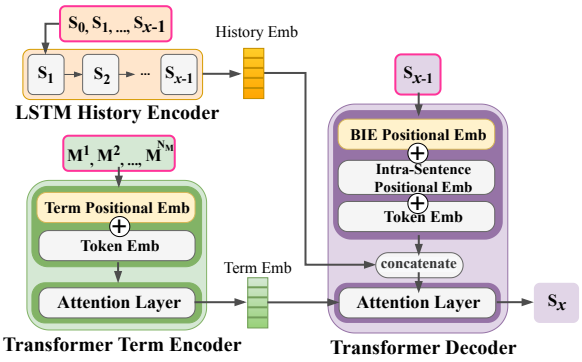

Figure 2: Flowchart for length-controlled Transformer. When generating sentence s_x, the model is input $(s_0, s_1, ..., s_{x-1})$, $(M^1, M^2, ..., M^{N_M})$, and s_{x-1}.

images' term sets $M^1, ..., M^{N_M}$, where $H_x = LSTM(s_0, ..., s_{x-1})$, denotes a history embedding for all previous sentences, generated from a LSTM layer; $M^t = \{m_1^t, ..., m_{N_m}^t\}$ 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 $M^1, ..., M^{x-1}, M^{x+1}, ..., M^{N_M}$ 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-

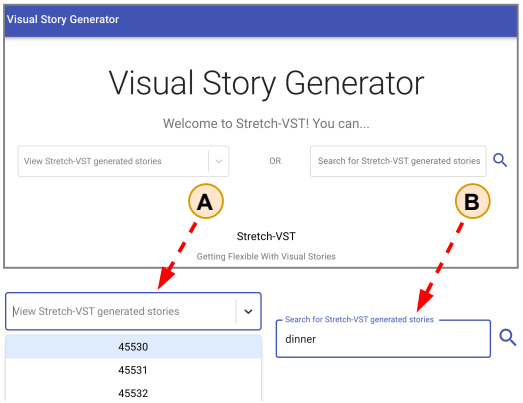

Figure 3: User interface of Stretch-VST. User can (A) select an story ID from the drop-down menu or (B) search a stories by keywords.

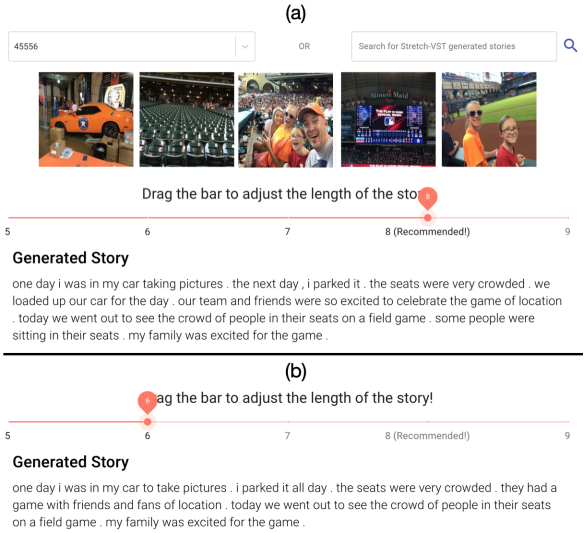

Figure 4: (a) The panel will show 5 images and visual story with recommended length. User can (b) drag the bar-slider and select the desired length of visual story.

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 story length (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

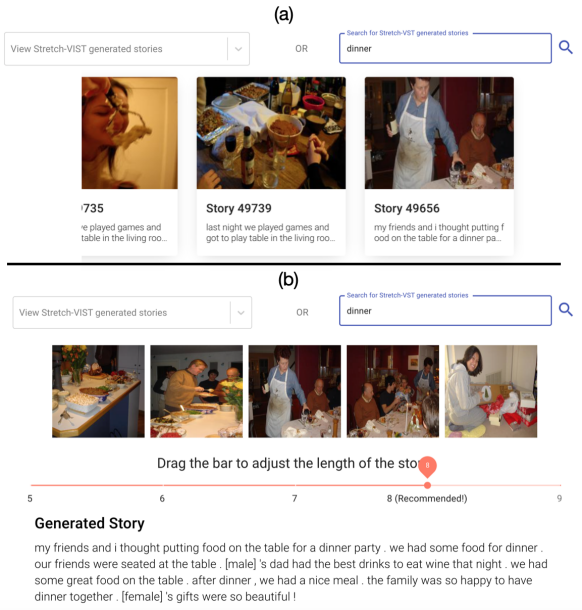

Figure 5: (a) The panel will provide several snippets of visual stories that contain the keyword (e.g, *dinner* in the story). (b) Selected a snippet, the panel will show the visual story with the recommended length. User can also drag the bar-slider to select desired story length.

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

	Rank	#1st rank	#Sentences	#Tokens
VIST(Sect. 4.1)				
AREL	2.47	274	5.00	41.99
GLAC	2.60	258	5.00	35.32
KG-Story	2.51	297	5.81	44.13
Stretch-VST	**2.41**	**421**	6.22	69.74
VIST w/ incoherent image (Sect. 4.2)				
AREL	2.04	364	3.00	25.41
GLAC	2.08	375	3.00	22.37
Stretch-VST	**1.87**	**511**	3.83	41.56

Table 1: Average rankings (1 to 4, lower is better) and number 1^{st} ranked stories (larger is better) rated by human judges, along with average number of sentences and tokens per story. (ρ value < 0.05, N=250)

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 compare to KGStory's single insertion. While the prolonging 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., 2018b; Hsu et al., 2019). Figure 7(a) compares stories generated from Stretch-VST to stories from the baselines.

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

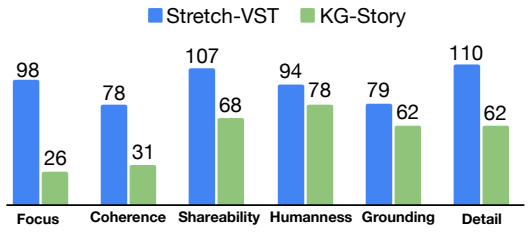

Figure 6: Aspect-wise votes for Stretch-VST and KG-Story's first-place stories collected via the questionnaire.

Figure 7: (a) Example visual stories generated by baselines and Stretch-VST. (b) Stories with fewer images from baseline models and Stretch-VST.

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

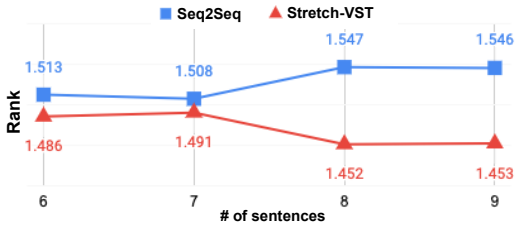

Figure 8: Average rankings between Stretch-VST and baseline for prolonged stories.

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

Collin F. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *COLING-ACL*.

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

Zhiqian Chen, Xuchao Zhang, Arnold P. Boedihardjo, Jing Dai, and Chang-Tien Lu. 2017. Multimodal storytelling via generative adversarial imitation learning. *ArXiv*, abs/1712.01455.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Diana Gonzalez-Rico and Gibran Fuentes Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. *ArXiv*, abs/1806.00738.

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

Jiaxian Guo, S. Lu, Han Cai, W. Zhang, Y. Yu, and J. Wang. 2018. Long text generation via adversarial training with leaked information. In *AAAI*.

Hiroaki Hayashi, Zecong Hu, Chenyan Xiong, and Graham Neubig. 2020. Latent relation language models. In *AAAI*.

Chao-Chun Hsu, Szu-Min Chen, Ming-Hsun Hsieh, and Lun-Wei Ku. 2018. Using inter-sentence diverse beam search to reduce redundancy in visual storytelling. *ArXiv*, abs/1805.11867.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Kenneth Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual storytelling. *ArXiv*, abs/1912.01496.

Ting-Yao Hsu, Huang Chieh-Yang, Yen-Chia Hsu, and Ting-Hao Kenneth Huang. 2019. Visual story post-editing. In *ACL*.

J. Hu, Yu Cheng, Zhe Gan, J. Liu, Jianfeng Gao, and Graham Neubig. 2020. What makes a good story? designing composite rewards for visual storytelling. *ArXiv*, abs/1909.05316.

Qiuyuan Huang, Zhe Gan, A. Çelikyilmaz, Dapeng Wu, J. Wang, and X. He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. *ArXiv*, abs/1805.08191.

Ting-Hao Kenneth Huang, Francis Ferraro, N. Mostafazadeh, Ishan Misra, Aishwarya Agrawal, J. Devlin, Ross B. Girshick, X. He, P. Kohli, Dhruv Batra, C. L. Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *HLT-NAACL*.

Y. Jung, Dahun Kim, S. Woo, Kyungsu Kim, Sungjin Kim, and I. Kweon. 2020. Hide-and-tell: Learning to bridge photo streams for visual storytelling. *ArXiv*, abs/2002.00774.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and M. Okumura. 2016. Controlling output length in neural encoder-decoders. *ArXiv*, abs/1609.09552.

Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and B. Zhang. 2018. Glac net: Glocal attention cascading networks for multi-image cued story generation. *ArXiv*, abs/1805.10973.

Peter LoBue and A. Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *ACL*.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *ArXiv*, abs/1706.09528.

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

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Jing Wang, J. Fu, J. Tang, Zechao Li, and Tao Mei. 2018a. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *AAAI*.

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

Xin Eric Wang, Wenhu Chen, Y. Wang, and William Yang Wang. 2018b. No metrics are perfect: Adversarial reward learning for visual storytelling. *ArXiv*, abs/1804.09160.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *ACL*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

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