SKETCH STORYTELLING

Yucheng Zhou

Fudan University
yczhou18@fudan.edu.cn

ABSTRACT

Sketch storytelling aims to generate a story for a given sketch. Although image captioning based on deep learning has great progress, describing the sketch in a story style is still a challenge. The reason is that there is currently no paired sketch-story data which is expensive to acquire. Therefore, it is necessary to train a sketch storytelling model without using any paired sketch-story data. To address these issues, we replace the natural image in image caption dataset with the sketch with the corresponding objects to generate pseudo sketch, which can obtain pseudo paired sketch-caption and sketch-image data. Due to these pseudo sketches are not drawn in a standardized way, we present a selective attention module to reduce noise for pseudo sketches. Furthermore, we propose four novel objectives include sketch-image matching, image-caption generation, sketch-caption generation, mask infilling, which help the model learn mappings between sketch and story from more perspectives. Consequently, we built a test set for sketch-story evaluation. The experimental results show that our model achieves state-of-the-art performance as compared to other methods.

***Index Terms*—** Image Captioning, Selective Attention, Cross-Domain Learning, Story Generation

1. INTRODUCTION

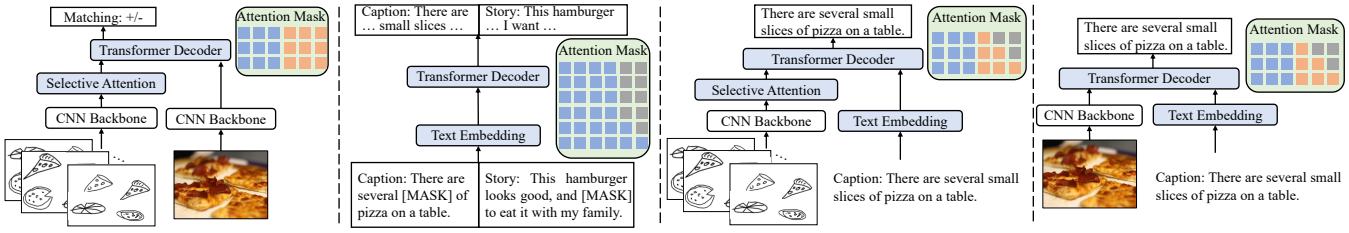
Sketch storytelling refers to generate a story for a given sketch. As image captioning becomes a major focus of crossing vision and language research, image description generation is a fundamental problem in many applications. Most existing image captioning models [1, 2] have been proposed and achieved promising results since their encoder-decoder architectures allow them to generate a description by encoding an image. To some extent, image captioning has shown the ability of recent deep neural network models for understanding and describing the image.

Despite their success, most existing image captioning models have only focused on describing the natural image. When the image captioning technology is applied to the education of children and drawing therapy, it is required to understand a sketch with some lines drawing by humans shown in Fig.1(right), which is not the complex sketches shown in Fig.1(left). However, there are not any sketch-caption data and it is expensive to acquire. To address the problem, we map the single-object sketches to an image in image-caption data by object-to-object pattern to generate pseudo sketch, which can obtain pseudo paired sketch-caption and sketch-image data. These single-object sketches are from the QuickDraw dataset [3]. However, there is noise in these pseudo sketches, because they are not drawn in a standardized way. To reduce the noise in pseudo sketch, we sample multiple pseudo sketches for a corresponding natural image and use a selective attention module to adaptively integrate their representation to reduce the bias from certain samples.

Fig. 1. **Left**: Sketch in ImageNet-Sketch [4]. **Right**: Sketch in QuickDraw [3].

In recent years, many researchers attempt to generate a cross-domain image caption. The methods proposed by [5, 6] mainly improve the decoder to integrate a style into a caption. However, these methods rely on the paired image and stylized caption data and it is expensive for acquiring these paired data. To make attempts at relaxing the reliance on paired data, Mathews et al. [7] divide the process of generating stylized captioning into two stages. The first is to generate the semantic terms of a given image and then the semantic terms are used to generate stylized image captioning. Although the Semstyle model proposed by Mathews et al. [7] can generate stylized image captioning, it still relies on the paired semantic terms and stylized caption data. To get rid of the constraints of paired data, the unsupervised image caption methods achieve mapping image to caption without any paired image-caption data [2, 8]. However, the performance of unsupervised methods is still not satisfactory.

To address these issues, we propose four novel objectives, including sketch-image matching, image-caption generation, sketch-caption generation, mask infilling. Sketch-image matching takes two images from natural image and sketch and classify whether they have the same meaning, which could strengthen objects alignment between natural image and sketch. For image-caption generation and sketch-caption



(a) Sketch-Image Matching (b) Mask Infilling (c) Sketch-Caption Generation (d) Image-Caption Generation

Fig. 2. Four objectives in our model.

generation, the model learns the mapping from visual space to language space to improve its cross-modal generative ability. Mask infilling has been verified that it can effectively improve the ability of language model generation [9]. We take mask infilling to recover the caption and story to enhance understanding and generation for the story. Using objectives of different perspectives to improve the model's performance is proven to be very effective [10, 11]. It should be noted that these objectives can be used to generate sentences in other styles because they are independent of the specific style.

In the experiments, we employ the ROCStories [12] as the story corpus and MSCOCO dataset [13] as the image-caption dataset. Moreover, we built a test set (STST) to evaluate the performance for sketch storytelling. The experimental results show that our model achieves state-of-the-art performance as compared to other methods. In particular, we compare our model and other methods on the FlickrStyle10K dataset [6] to demonstrate that the four objectives we proposed are independent of the specific style. The evaluation on FlickrStyle10K is an unpaired setting and we built a test set (STStyle) for evaluation. In addition, we compare the performance of our model on the different ablation settings to demonstrate the effectiveness of our method. Finally, we further analyze the performance on different numbers of pseudo sketch samples for the selective attention module.

2. METHOD

2.1. Feature Extraction and Selective Attention

As shown in Fig.2, given image $\boldsymbol{I}$ and sketch $\boldsymbol{I}_s$, we use a pre-trained convolutional neural network (CNN) backbone to extract their features map $\boldsymbol{f}_i$ and $\boldsymbol{f}_s$. As same as [14], the dimension of $\boldsymbol{f}_i$ and $\boldsymbol{f}_s$ are $\mathbb{R}^{C\times H\times W}$, and $C = 2048$, $H = H/32$, $W = W/32$. We take a 1×1 convolution to reduce the channel dimension of the feature maps from C to d, which is same as embedding dimension of language model. Then, we transform their dimension to $\mathbb{R}^{HW\times d}$, like a sequence. Because there are noises in sketches, we sample multiple sketches mapped by an image and pass them into the selective attention module to get a more accurate representation. In the selective attention module, we employ Multi-Layer Perceptrons (MLP) to calculate the weight of each sketch and integrate the features by the weight,

$$\alpha_i = \mathrm{MLP}(\mathrm{Pooling}(\boldsymbol{f}_s^i)) \tag{1}$$

$$\hat{\boldsymbol{f}}_s = \boldsymbol{f}_s^0 \times \alpha_0 + \boldsymbol{f}_s^1 \times \alpha_1 + ... + \boldsymbol{f}_s^n \times \alpha_n \tag{2}$$

where $\hat{\boldsymbol{f}}_s$ refers to a denoise sketch representation through selective attention module.

2.2. Sketch-Image Matching

Since the sketches are not drawn in a standardized way, the sketches lack some details of the original images, such as the background. To strengthen objects alignment relation between natural image and sketch, we propose the sketch-image matching objective to take a classification objective to classify whether they are with the same meaning, shown in Fig.2(a). Given an image feature $\boldsymbol{f}_i$ and a sketch feature $\boldsymbol{f}_s$, we take them into a sequence "[CLS] [SoS] $\boldsymbol{f}_s$ [EoS] [SoI] $\boldsymbol{f}_i$ [EoI]". The hidden features corresponding [CLS] pass a MLP to infer whether the sketch and image are matching.

$$\boldsymbol{p} = \text{Transfromer-Decoder}(\boldsymbol{f}_i, \boldsymbol{f}_s) \tag{3}$$

$$\mathcal{L}^{(SIM)} = -\log \boldsymbol{p}[y = c] \tag{4}$$

where Transfromer-Decoder is our model, and $\boldsymbol{p}$ is a probability distribution over class. c is gold label, and $\mathcal{L}^{(SIM)}$ is classification loss obtained by cross-entropy loss functions.

2.3. Mask Infilling

The mask infilling model has demonstrated its significance in many pre-trained language models, such as BART [9]. In this objective shown in Fig.2(b), we randomly sample a number of text spans, with span lengths drawn from a Poisson distribution ($\lambda = 3$), and each span is replaced with a single [MASK] token. For caption and story, we set prompts before them, like "Caption:" for a caption and "Story:" for a story. The loss $\mathcal{L}^{(MI)}$ is calculated by the cross-entropy loss function in the mask recovery part.

2.4. Sketch-Caption and Image-Caption Generation

To establish the mapping relationship from visual space to language space, we propose two objectives, including sketch-caption generation and image-caption generation. For the two

different modal features of image and text, we take a prompt before caption to integrate them into a sequence, like "[CLS] [SoS] $\boldsymbol{f}_s$ [EoS] Caption:" for a sketch-caption and "[CLS] [SoI] $\boldsymbol{f}_i$ [EoI] Caption:" for a image-caption. The calculation of the loss $\mathcal{L}^{(S2C)}$ and $\mathcal{L}^{(I2C)}$ are the same as that of $\mathcal{L}^{(MI)}$.

2.5. Joint Training

We train our model with all objectives (i.e., Sketch-Image Matching, Image-Caption Generation, Sketch-Caption Generation, Mask Infilling) jointly by minimizing the four loss functions as:

$$\mathcal{L} = \mathcal{L}^{(SIM)} + \mathcal{L}^{(MI)} + \mathcal{L}^{(S2C)} + \mathcal{L}^{(I2C)} \tag{5}$$

3. EXPERIMENTS

In the experiments, we investigate performance by comparing our model with other methods to verify the effectiveness of our model. Meanwhile, we demonstrate the four objectives are independent of a specific style by comparing our method with others methods. Furthermore, we verify the effectiveness of four objectives and the selective attention module by an ablation study and analyze the impact of different numbers of pseudo sketch samples for the selective attention module.

3.1. Dataset

In the experiments, we train our model on three datasets: MSCOCO [13], QuickDraw [3], ROCStories [12]. For MSCOCO, we use its subset that only includes images of 69 categories, because only these categories overlap with that of QuickDraw. The ROCStories includes 98,159 stories for story generation. For testing, we built two test sets (STST and STStyle). STST includes 3500 sketches and their story sentence. STStyle includes 1000 sketches and their stylized sentences. Since the FlickrStyle10K dataset [6] only provides 7000 training images and their stylized captions, we use the captions of 6000 images that do not overlap with the samples used to built STStyle as a corpus to verify the independence of our method for a specific style. For the model evaluation, we used the open-source code[2] to report the results, which includes BLEU [15], METEOR [16], ROUGE [17], CIDEr [18]. In addition, we use BERTScore [19] to evaluate the generation results.

3.2. Experimental Settings

The pre-trained CNN backbone we used is ResNet152 [20], and each image and sketch is resize to 224×224. The transformer decoder is a pre-trained GPT-2 [21]. We use Adam optimizer to optimize the model with a learning rate of 1×10^{-5} and a linear warm-up. The maximum training epoch, warm-up step, and batch size are set to 10, 2000, and 16. The maximum sequence length and dropout are set to 128 and 0.1. The weight decay and gradient clipping are set to 0.01 and 1.0. In the inference, we set the maximum sentence length to 30, and use beam search with the beam size is 3.

[2]https://github.com/tylin/coco-caption

Table 1. The results on STST.

Methods	BLEU-4	METEOR	ROUGE-L	CIDEr	BERT
UIC [8]	7.81	17.25	13.69	9.22	39.58
DLN [22]	8.87	23.34	18.94	14.48	42.51
TSGAN [2]	9.15	23.70	19.65	15.22	44.87
Up-Down [1]	11.37	26.45	27.62	17.77	50.26
VLP [23]	14.11	32.33	31.89	26.21	52.80
Ours	**24.53**	**44.73**	**40.67**	**45.81**	**59.88**

3.3. Main Results

Since the methods [5, 6, 7] rely on paired image and stylized caption data or paired semantic terms and stylized caption data, we first compare our method with three unsupervised methods, because there are not any sketch-story data for model training. The methods of UIC [8] and TSGAN [2] are selected due to their state-of-the-art results in unsupervised image captioning. The DLN [22] is a state-of-the-art model on the unsupervised stylish image caption. These unsupervised methods are trained on unpaired sketch and story data. In addition, we compare a supervised image caption model (Up-Down [1]) trained on paired image-caption data. Due to the rapid development of cross-modal pre-trained models, VLP [23] has demonstrated excellent capability on image caption generation and achieve state-of-the-art results in image captioning. Up-Down and VLP are fine-tuned on sketch-caption data and image-caption data.

As shown in Table 1, both our models achieve better results on each metric as compared to other methods. The reason is that our methods can effectively transfer knowledge between different domains, even if there is no pairing data, such as sketch and story. We can observe that the methods that use paired data are better than others without paired data. It demonstrates that even if the data is not completely in the same domain, it can still improve performance because it will include some common knowledge.

3.4. Independent of Specific Style

To verify the claim that our method is independent of a specific sentence style, we train and evaluate the model on captions with other styles. We keep the setting the same as Sec.3.3, except for the training corpus and test set. As shown in Table 2, we can observe that our model still achieves better results as compared with other methods. It demonstrates our method can be effectively used for other styles.

Table 2. The results on STStyle.

Methods	BLEU-4	METEOR	ROUGE-L	CIDEr	BERT
UIC [8]	9.68	10.65	15.36	12.69	39.89
DLN [22]	9.57	13.11	19.11	15.66	44.92
TSGAN [2]	10.37	15.31	20.03	17.54	45.18
Up-Down [1]	13.56	18.88	30.65	19.84	53.72
VLP [23]	16.69	26.35	36.73	29.77	57.51
Ours	**25.66**	**34.93**	**41.86**	**47.42**	**60.40**

Table 3. Ablation study of our approach. "*w/o SIM*" denotes removing the sketch-image matching objective in our model, "*w/o S2C*" denotes removing the sketch-caption generation objective in our method, '*w/o I2C*" denotes removing the image-caption generation objective in our approach, and "*w/o SA*" denotes removing the selective attention module.

Methods	BLEU-4	METEOR	ROUGE-L	CIDEr	BERT
Ours	**24.53**	**44.73**	**40.67**	**45.81**	**59.88**
◊ w/o SIM	24.21	43.65	39.64	44.21	58.94
◊ w/o S2C	23.61	42.99	39.47	42.69	57.40
◊ w/o I2C	21.94	42.26	38.57	41.48	55.64
◊ w/o SA	22.92	42.84	38.57	42.17	57.20

3.5. Ablation study

As shown in Table 3, we conduct an ablation study for our model to investigate each component of our method. We investigate the effectiveness of three objectives by removing the corresponding objectives. Moreover, we verify the effectiveness of the selective attention module by removing it. The results show that the performance drops a lot when the image-caption generation (I2C) objective is removed. In addition, the sketch-caption generation (S2C) and sketch-image matching are also important for our methods, because the performance drops a bit when one of the two objectives is removing. The selective attention (SA) module shows its significance in our model, because it plays an important role in improving the performance of the model.

3.6. Qualitative Analysis

Fig. 3 gives some qualitative examples sampled randomly in the result of our model. The difference between caption generation and story generation is that the prompt is "Caption:" or "Story:". We can see that the caption results include the main objects in the sketch and describe their relation, which shows the ability to visual understand and language generation. We can see that the sample "the cars stop at the red stop sign.", which includes the commonsense from the caption. It indicates that the mapping from sketch to caption is effectively learned by the model. We can see that the generated stories of our model contain some human subjective emotions and actions, like "getting ready" and "drive it away".

3.7. Analysis of Selective Attention

To investigate the impact on the different numbers of pseudo sketch samples for the selective attention module, we train

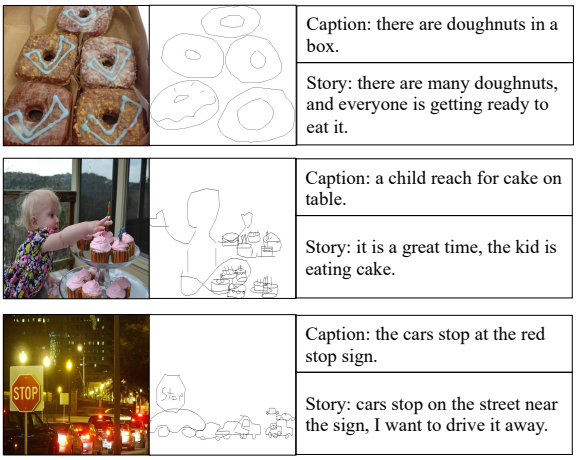


Fig. 3. Captions and stories generated from randomly sampled sketches, and the corresponding image.

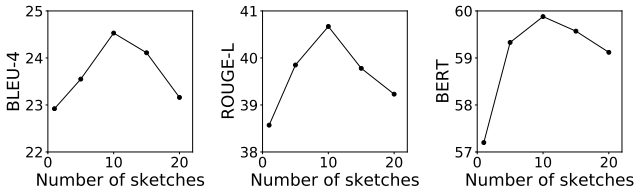


Fig. 4. The impact of different numbers of pseudo sketch samples for the selective attention module.

five models when the numbers of pseudo sketch samples are 1, 5, 10, 15, 20. Specifically, when we train the model, we control the number of pseudo sketch samples corresponding to a natural image. When the number of pseudo sketch samples is 1, it means that the selective attention module is removed. The results show that the performance is best when the number of pseudo sketch samples is 10. It is observed, with more pseudo sketch samples added, the performance of our approach first improves and then drops, which is in accordance with intuition. Because there are too few pseudo sketch samples, these sketches may contain a lot of noise. However, if there are too many pseudo sketch samples, the selective attention module needs to judge more samples, which is not conducive to obtaining accurate results.

4. CONCLUSION

In this paper, we propose a novel model to generate the story for a given sketch. The model can adaptively integrate multiple sketches by a selective attention module. Moreover, we present four training objectives, including sketch-image matching, image-caption generation, sketch-caption generation, mask infilling, which can help model learn mappings between sketch and story from more perspectives. In the experiments, we built a test set (STST) for sketch storytelling evaluation. The experimental results show our method can achieve state-of-the-art results as compared to other methods.

5. REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR 2018*, 2018, pp. 6077–6086.

[2] Yucheng Zhou, Wei Tao, and Wenqiang Zhang, "Triple sequence generative adversarial nets for unsupervised image captioning," in *ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. 2021, pp. 7598–7602, IEEE.

[3] David Ha and Douglas Eck, "A neural representation of sketch drawings," in *ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. 2018, OpenReview.net.

[4] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing, "Learning robust global representations by penalizing local predictive power," in *NeurIPS 2019*, 2019, pp. 10506–10518.

[5] Alexander Patrick Mathews, Lexing Xie, and Xuming He, "Senticap: Generating image descriptions with sentiments," in *AAAI, 2016, Phoenix, Arizona, USA*. 2016, pp. 3574–3580, AAAI Press.

[6] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng, "Stylenet: Generating attractive visual captions with styles," in *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 955–964.

[7] Alexander Patrick Mathews, Lexing Xie, and Xuming He, "Semstyle: Learning to generate stylised image captions using unaligned text," in *CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 8591–8600.

[8] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo, "Unsupervised image captioning," in *CVPR*, 2019, pp. 4125–4134.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL 2020*, 2020, pp. 7871–7880.

[10] Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang, "Eventbert: A pre-trained model for event correlation reasoning," *arXiv preprint arXiv:2110.06533*, 2021.

[11] Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang, "Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph," in *NAACL*, 2021, pp. 5822–5834.

[12] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen, "A corpus and cloze evaluation for deeper understanding of commonsense stories," in *NAACL*, 2016, pp. 839–849.

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015.

[14] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang, "E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning," in *ACL/IJCNLP 2021*, 2021, pp. 503–513.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[16] Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[17] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[18] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR 2016*, 2016, pp. 770–778.

[21] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[22] Cheng-Kuan Chen, Zhu Feng Pan, Ming-Yu Liu, and Min Sun, "Unsupervised stylish image description generation via domain layer norm," in *AAAI 2019*, 2019, pp. 8151–8158.

[23] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao, "Unified vision-language pre-training for image captioning and VQA," in *AAAI 2020, New York, NY, USA, February 7-12, 2020*. 2020, pp. 13041–13049, AAAI Press.