

Sketchy Scene Captioning: Learning Multi-Level Semantic Information from Sparse Visual Scene Cues

Lian Zhou, Yangdong Chen, Yuejie Zhang[†]

School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai 200438, China

{16110240019, 19110240010, yjzhang}@fudan.edu.cn

Abstract

To enrich the research about sketch modality, a new task termed Sketchy Scene Captioning is proposed in this paper. This task aims to generate sentence-level and paragraph-level descriptions for a sketchy scene. The sentence-level description provides the salient semantics of a sketchy scene while the paragraph-level description gives more details about the sketchy scene. Sketchy Scene Captioning can be viewed as an extension of sketch classification which can only provide one class label for a sketch. To generate multi-level descriptions for a sketchy scene is challenging because of the visual sparsity and ambiguity of the sketch modality. To achieve our goal, we first contribute a sketchy scene captioning dataset to lay the foundation of this new task. The popular sequence learning scheme, e.g., Long Short-Term Memory neural network with visual attention mechanism, is then adopted to recognize the objects in a sketchy scene and infer the relations among the objects. In the experiments, promising results have been achieved on the proposed dataset. We believe that this work will motivate further researches on the understanding of sketch modality and the numerous sketch-based applications in our daily life. The collected dataset is released at <https://github.com/SketchySceneCaption/Dataset>.

1 Introduction

In recent years, sketch has emerged as one important data modality (Eitz et al., 2012; Yu et al., 2015). Compared to a natural image, a sketch only contains sparse and ambiguous visual information. Current works about sketch mainly focus on predicting one class label for a sketch, and such a label provides very limited semantic information (Eitz et al., 2012). Differently, the tasks about natural image are abundant, such as classification (Deng et al., 2009), captioning (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018), and visual question answering (Anderson et al., 2018). What hinders the research about sketch is the lack of sketch datasets. Specifically, natural images are easy to be obtained and a lot of efforts have been put into annotating the images. By contrast, sketch is created by human and the generation of a sketch is time-consuming, which limits the volume of a sketch dataset and the visual details of the sketches in the dataset. Hence, most of current sketch datasets only contain sketches with a single object and the corresponding class label. Drawing inspiration from the task of natural image captioning, it is attractive to expand a sketch to a sketchy scene which contains several objects, and extend the class label to a sentence-level or even a paragraph-level description. In a word, a sketchy scene dataset with multi-level descriptions is in urgent need to promote the research about sketch.

One promising application with above extension is child education (Neshati et al., 2017). Specifically, with the wide popularity of tablet PC, it becomes common for a child to doodle on a touch screen. To interact with a child, a computer agent needs to understand what a child has drawn and give reasonable response to the child. For example, if a sketchy scene drawn by a child cannot match the sentence or paragraph given by the agent, the child is required to draw the sketchy scene again, which helps improve the drawing skill of the child. Another potential application is the assistance for the visually impaired

people. With simple and sparse visual content, a sketchy scene can be easily turned into a concave-convex plate that can be read by a visually impaired person in a touch manner. With the corresponding caption transformed into human voice (Zou et al., 2018a), the visually impaired person can feel what are depicted in the sketchy scene without the help of others. Other potential applications include large-scale sketchy scene retrieval via human language and automatic sketch management on the Web (e.g., to cluster the numerous sketchy scenes with similar topics).

Motivated by the observations above, we extend the task of sketch classification to Sketchy Scene Captioning, a task that aims to generate multi-level (i.e., sentence-level and paragraph-level) descriptions for a sketchy scene, as shown in Figure 1. To the best of our knowledge, our work is the first attempt to generate multi-level and dense descriptions for a sketchy scene. Currently, to achieve the goal of sketchy scene captioning is very challenging for two reasons. **First**, compared to a natural image captioning dataset, only the generation of a sketchy scene is time-consuming, let alone the annotation of the sketchy scene. **Second**, a sketchy scene only contains very sparse visual cues. That is, an object is depicted only with some lines. In addition, the visual cues of a sketchy scene are also ambiguous. That is, the objects in different sketchy scenes have great variations in appearance, making it difficult to distinguish the objects. To overcome the above two challenges, we create a sketchy scene dataset with multi-level descriptions and achieve the goal of sketchy scene captioning using several sequence-learning-based models in the field of image captioning (Vinyals et al., 2015; Xu et al., 2015; Krause et al., 2017). The contributions of this work are three-folds: 1) A new task termed Sketchy Scene Captioning is proposed to generate multi-level descriptions for a sketchy scene. This task can be treated as a new paradigm for comprehensive understanding of the sparse visual cues; 2) A sketchy scene captioning dataset is constructed based on *SketchyScene* dataset (Zou et al., 2018b). Currently, the new dataset contains 1,000 sketchy scenes with both the sentence-level and paragraph-level captions; and 3) Promising experimental results have been achieved on the newly collected dataset, demonstrating the potentials of Sketchy Scene Captioning. We hope this work could help motivate further researches on mining multi-level semantic information from sketchy scenes.

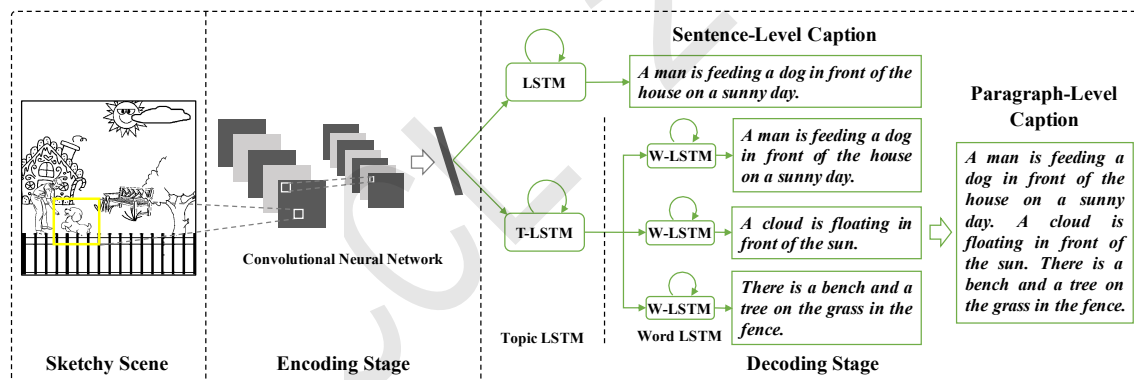


Figure 1: An overview of the proposed Sketchy Scene Captioning framework. The visual attention models are not given for conciseness.

2 Related Works

In this section, we will briefly review the related works about sketch and image captioning. The differences between the prior works and ours will be discussed as well.

Current works about sketch mainly focus on Sketch Classification and Sketch-based Image Retrieval (SBIR). Sketch Classification is a task of recognizing what object is depicted in a sketch. SBIR aims to retrieve a natural image for a given query sketch. In recent years, great progresses have been made in the field of sketch. For example, Yu et al. (2015) proposed Sketch-a-Net, a multi-scale and multi-channel deep neural network, to yield the sketch recognition performance surpassing that of humans on the *TU-Berlin* sketch dataset (Eitz et al., 2012). Sangkloy et al. (2016) proposed the *Sketchy* dataset, which

was the first large-scale collection of sketch-photo pairs for image retrieval. He et al. (2017) proposed a deep visual-sequential fusion mechanism to model the visual and sequential patterns of the strokes of a sketch. Liu et al. (2019) proposed a semantic-aware knowledge preservation method for sketch-based image retrieval. In spite of the above progresses, the related works about sketch classification are limited to assigning a class label to each sketch. In this paper, we go a step further to generate multi-level and dense descriptions for a sketchy scene.

Current methods for image captioning can be mainly divided into three categories, that is, template-based, retrieval-based, and sequence-learning-based. In the template-based method, the salient objects, their attributes, and the relations among objects in an image were first recognized, and a pre-defined template was then filled with the detected information to yield a full sentence (Elliott and Keller, 2013). The retrieval-based method first obtained the visually similar image with the query image, and then used the description of the retrieved image as the description of the query image (Karpathy et al., 2014). However, these two methods could only generate relatively fixed sentences, relying on the given image-caption dataset. In the era of deep learning, sequence learning was adopted to adaptively generate a description for a natural image, where a Convolutional Neural Network (CNN) (He et al., 2016) encoder was used to encode the image into a high-level visual representation, and a Recurrent Neural Network (RNN) (Sutskever et al., 2014) decoder was adopted to “translate” the image representation into a sentence. Typically, Vinyals et al. (2015) first proposed to use Inception (Ioffe and Szegedy, 2015) convolutional neural network as the encoder to convert an image into a fixed-length vector, and then use Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) neural network as the decoder to generate a caption for the image. Xu et al. (2015) introduced two spatial visual attention mechanisms to help the model dynamically focus on the image regions corresponding to the word that was about to be generated. Besides, Krause et al. (2017) designed a hierarchical LSTM model to generate a paragraph-level description for a natural image. Overall, current works on image captioning mainly focus on natural images. Differently, we explore the caption generation problem in the field of a different domain, that is, sketchy scene which only contains sparse and ambiguous visual information.

3 A New Dataset for Sketchy Scene Captioning

To the best of our knowledge, there is no available dataset for sketchy scene captioning. Hence, we need to first construct a sketchy scene dataset with sentence-level and paragraph-level descriptions. Next, we will describe how the dataset is collected in details.

3.1 *SketchyScene* Dataset without Descriptions

In the field of sketch, several sketch datasets, such as *TU-Berlin* (Eitz et al., 2012) and *Sketchy* (Sangkloy et al., 2016), have been proposed for sketch classification or cross-modal retrieval, and the sketches in these datasets are created by humans. However, each sketch in these datasets only contains one object with discrete class labels or together with the stroke orders. As a result, the related researches based on these datasets can only deal with single object, which indicates that to create a sketch dataset with annotations is very challenging. With single object in a sketch, these datasets cannot be used for captioning. To extend the research on sketch, Zou et al. (2018b) propose a brand new dataset called *SketchyScene* recently. The dataset consists of scene sketches where each scene sketch contains multiple objects. Each object in a scene sketch is assigned with one class label out of 45 categories. Because every scene has a corresponding natural or cartoon image for reference, all the sketchy scenes are supposed to be consistent with the real world. Besides, there is also segmentation information for each sketch. Because each scene sketch in *SketchyScene* is constructed by combining the separate instances of several categories, the volume of *SketchyScene* can grow relatively large, which ensures the diversity of the sketchy scenes. Four examples from the *SketchyScene* dataset are shown in Figure 2.

In spite of the advantages of *SketchyScene*, no sentence-level or paragraph-level descriptions are provided in *SketchyScene*. In this work, we choose to construct our sketchy scene captioning dataset based on *SketchyScene* with the following two reasons. **First**, *SketchyScene* provides realistic and diverse sketchy scenes, which makes the dataset suitable for sketchy scene captioning. It can be observed from

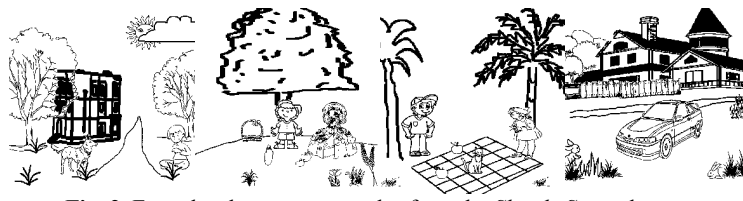


Figure 2: Four sketchy scene examples from the *SketchyScene* dataset.

Figure 2 that the objects of each sketchy scene are quite diverse and the object arrangement in each scene is reasonable, making it meaningful to generate a sentence-level or even a paragraph-level description for the sketchy scene. **Second**, *SketchyScene* provides class labels for the objects in each sketchy scene, offering important hints for the annotators to give more accurate descriptions for each sketchy scene.

3.2 Description Collection for *SketchyScene* Dataset

We conduct the data collection in a manner of crowdsourcing. Hence, we first create a website to ease the annotation job. On the website, several annotation examples, a randomly picked target sketchy scene, and the category labels of the target sketchy scene are presented. To ensure the annotation quality, we only invite a number of graduate students in universities as volunteers who are well trained in English. We realize that some annotation instructions for volunteers are still needed to further improve the annotation quality. After analyzing a few initial annotated captions without instructions and being inspired by the proposed requirements when collecting *MSCOCO* (Chen et al., 2015) for image captioning, we summarize the following rules that the volunteers should obey when annotating a sketchy scene.

- Be faithful to the visual content of the presented sketchy scene. Do not describe anything unrelated to the sketchy scene (e.g., what may happen in the past or future).
- Do not describe what people may say in a sketchy scene.
- Do not give a name to a person or animal.
- Do not use any abbreviation in the descriptions.
- Try to use more specific words when possible. For example, use words such as “girl”, “boy”, “woman”, and “man” instead of “people”.
- For sentence-level annotation, describe the sketchy scene with a brief summary, not necessary to include everything.
- For paragraph-level annotation, describe the sketchy scene as detailed as possible with all the given category labels.

With the settings above, we successfully collect 1,000 annotated scene sketches, where each scene sketch is associated with one sentence-level description and one paragraph-level description. Our website for data collection is still open for more annotations, and a new version of the dataset is expected to be released in the future. To share the idea of sketchy scene captioning and the collected dataset with other researchers timely, we currently use the collected 1,000 samples for exploration in this work.

3.3 Dataset Analysis

In this section, we will take a look at the newly collected dataset. Three representative examples are shown in Figure 3. Column “Tags” shows the category labels corresponding to the objects in a sketchy scene, and these labels act as the guiding words for the volunteers when annotating a sketchy scene. The following two columns show the sentence-level and the paragraph-level descriptions of a sketchy scene, respectively. By analyzing the descriptions, we find that the annotators tend to take the most salient




Sketchy Scene	Tags	Sentence-Level Caption	Paragraph-Level Caption
	cloud, cow, fence, flower, grass	A cow is standing on the grass.	A cow is standing on the ground. Some flowers and grass grow on the ground. There is a fence behind the cow. Two clouds are floating in the sky.
	car, cloud, cow, grass, house, mountain, road, tree	A car is on the road and a cow is eating grass beside the house.	There is a house beside the road. A cow is standing beside a tree and eating grass. A car is on the road and a mountain is behind the house.
	cat, cloud, dog, fence, flower, grass, people, sun, tree	A girl is standing on the ground with a dog and a cat in a sunny day.	A girl is standing in front of a fence. Some flowers and grass grow on the ground. A tree is standing near the girl. A dog and a cat is sitting under the tree. The sun is shining and two clouds are floating in the sky.

Figure 3: Three representative examples of the collected multi-level captions for sketchy scenes.

object as the subject of a sentence-level caption and describe its interactions with other possible objects in a sketchy scene. Differently, more objects and their interactions are described in a paragraph-level caption. Although the volunteers may not follow the instructions strictly, the quality of the captions is still good enough for our research.

We further conduct a quantitative analysis on the collected captions. For the sentence-level captions, there are 504 different words in total. The lengths of the sentence-level captions are concentrated in 10~20 words, and the distribution of caption length is roughly in line with the Gaussian distribution. Besides, most sentences have only 1~3 relations (e.g., verb and preposition) among objects, which means that the annotators tend to focus on the salient parts of a sketchy scene and ignore other details during the sentence-level annotation. For the paragraph-level captions, there are 681 different words in total. Most of the paragraph-level captions contain 3~6 sentences. The lengths of the paragraph-level captions are concentrated in 25~35 words, and the lengths of all the single sentences in the paragraph-level captions are concentrated in 6~14 words. It can be found that the sizes of the two vocabularies above are relatively small, which are caused by two reasons. **First**, compared to a natural image, a sketchy scene contains much less visual details (e.g., the color of an object). **Second**, there are only 45 object categories in the *SketchyScene* dataset and the annotators are required to use the given category labels when constructing a caption. Due to these two reasons, a sketch dataset cannot become as diverse as a natural image dataset, which is a stubborn problem in current research on sketch.

4 Multi-Level Sketchy Scene Captioning through Sequence Learning

In this work, the popular sequence-learning-based method is adopted for flexible sketchy scene captioning, as shown in Figure 1. Our framework integrates Sketchy Scene Encoder for Deep Visual Features (i.e., encoding a sketchy scene at an abstract level to obtain a discriminative visual representation) and Sketchy Scene Decoder with Spatial Visual Attention (i.e., grasping more visual details of a sketchy scene while generating the description). It is a new attempt to generate multi-level descriptions for a sketchy scene through the sequence learning paradigm.

4.1 Sketchy Scene Encoder for Deep Visual Features

CNN is adopted as image encoder in sketchy scene captioning. Considering that a sketchy scene contains very sparse visual information, the outputs of different CNN layers are chosen as the visual features of a sketchy scene for comparison. The output of the fully-connected layer is a fixed-length vector that is denoted as \mathbf{v}_{fc} . The output of a convolutional layer is a set of spatial feature vectors that are denoted as $\mathbf{v}_{cv} = \{\mathbf{v}_i | \mathbf{v}_i \in \mathbb{R}^{d_v}, 1 \leq i \leq n\}$, where d_v is the feature dimension and n is the region number. These

fine-grained features can be used for visual attention. The global representation of a sketchy scene \mathbf{v}_0 is used to initialize the decoder and can be computed as:

$$\mathbf{v}_0 = \mathbf{v}_{fc} \quad \text{or} \quad \left(\sum_{i=1}^n \mathbf{v}_i \right) / n, \quad \mathbf{v}_i \in \mathbf{v}_{cv}. \quad (1)$$

4.2 Sketchy Scene Decoder with Spatial Visual Attention

Sentence-Level Decoder. The LSTM neural network is exploited as image decoder in sketchy scene captioning. The decoder generates a sentence $S = (s_0, \dots, s_c, s_{c+1})$ conditioned on the input visual features (\mathbf{v}_{fc} or \mathbf{v}_{cv}), where c is the length of the sentence, and s_0 and s_{c+1} denote the starting and ending tokens respectively. Each word in S is denoted as a one-hot vector. An embedding matrix is used to convert each word to a low-dimensional vector as follows:

$$\mathbf{x}_t = \mathbf{W}_e s_t, \quad 0 \leq t \leq c + 1, \quad (2)$$

where $\mathbf{W}_e \in \mathbb{R}^{d_e \times V}$, d_e is the dimension of word embedding, and V is the vocabulary size. The inputs of the decoder at time step t ($1 \leq t \leq c + 1$) include the embedding of the previous word \mathbf{x}_{t-1} and a contextual vector \mathbf{z}_t that is computed through the soft attention (Xu et al., 2015) as follows:

$$e_i^t = f_{att}(\mathbf{v}_i, \mathbf{h}_{t-1}), \quad 1 \leq i \leq n, \quad t \geq 1, \quad (3)$$

$$\alpha_i^t = \exp(e_i^t) / \sum_{k=1}^n \exp(e_k^t), \quad (4)$$

$$\mathbf{z}_t = \sum_{i=1}^n \alpha_i^t \mathbf{v}_i, \quad (5)$$

where \mathbf{h}_{t-1} is the hidden state of the decoder at time step $t - 1$, and f_{att} is the soft visual attention function that is implemented as a fully-connected neural network. It should be noted that \mathbf{z}_t exists only when a sketchy scene is encoded as a set of spatial feature vectors, otherwise the sketchy scene decoder is just a vanilla LSTM. Given the global visual representation \mathbf{v}_0 of a sketchy scene, the initial memory state \mathbf{c}_0 and the initial hidden state \mathbf{h}_0 can be obtained by feeding \mathbf{v}_0 into two separate fully-connected neural networks as follows:

$$\mathbf{c}_0 = f_c(\mathbf{v}_0), \quad \mathbf{h}_0 = f_h(\mathbf{v}_0), \quad (6)$$

where \tanh nonlinearity is adopted. It should be noted that the global visual representation \mathbf{v}_0 is only used to initialize the LSTM decoder. Given the visual features \mathbf{F} (i.e., \mathbf{v}_{fc} or \mathbf{v}_{cv}) of a sketchy scene, the LSTM decoder is learned by minimizing the negative logarithmic probability of the target sentence S as follows:

$$L_s = -\log P(S|\mathbf{F}) = -\sum_{t=1}^{c+1} \log P(s_t | s_0^{t-1}, \mathbf{F}), \quad (7)$$

where the whole conditional logarithmic probability can be decomposed into the multiplication of the logarithmic probability at each time step. The t -th word can be predicted by the output layer as follows:

$$P(s_t | s_0^{t-1}, \mathbf{F}) \propto \exp(\mathbf{E}_0(\mathbf{E}_1 \mathbf{h}_t + \mathbf{E}_2 \mathbf{z}_t)), \quad (8)$$

where $\mathbf{E}_0 \in \mathbb{R}^{V \times d_m}$, $\mathbf{E}_1 \in \mathbb{R}^{d_m \times d_m}$, $\mathbf{E}_2 \in \mathbb{R}^{d_m \times d_v}$, and d_m is the number of the LSTM cell units.

Paragraph-Level Decoder. Considering that the average length of a paragraph is about 30 words, it is difficult for a single LSTM to generate such a long sequence with correct meanings. Thus, a hierarchical LSTM (*H-LSTM*) network (Krause et al., 2017) is exploited to generate a paragraph-level caption for a sketchy scene. The *H-LSTM* network consists of a topic LSTM and a word LSTM. The topic LSTM

takes the visual features v_0 of a sketchy scene as input and generates a sequence of guiding signals $G = (g_0, \dots, g_N, g_{N+1})$, where N is the number of topics, g_0 is the starting signal, $g_t = 1 (1 \leq t \leq N)$ indicates that a new sentence needs to be generated, and $g_{N+1} = 0$ indicates stopping generating the paragraph. Visual attention is also conducted by the topic LSTM. At time step t , the hidden state h_t^T and the contextual vector z_t^T of the topic LSTM are concatenated to obtain the topic vector that is used to guide the caption generation of the word LSTM. The word LSTM works similar to the sentence-level captioning decoder except that the visual features used to initialize it are replaced with the topic vector. The loss function of *HLSTM* can be formulated as:

$$L_p = -\lambda_T \log P(G|v_{cv}) - \lambda_W \sum_{t=1}^N \log P(S_{W_t}|v_{cv}, h_t^T, z_t^T), \quad (9)$$

where λ_T and λ_W are the weighting coefficients of the topic LSTM loss and the word LSTM loss respectively, and S_{W_t} denotes the t -th sentence generated by the word LSTM.

5 Experiment and Analysis

5.1 Dataset and Preprocessing

We conduct the experiments on the newly collected dataset to verify the feasibility of the sketchy scene captioning task. The dataset is divided into training, validation, and testing sets with a ratio of 8:1:1, that is, 800 <sketchy scene, caption> pairs for training, 100 for validation, and 100 for testing. The words that appear at least 5 times in the training captions are kept, and a vocabulary of size 174 for sentence-level captions and another one of size 223 for paragraph-level captions are constructed. Each vocabulary includes a starting token “<start>”, an ending token “<end>”, and an unknown word token “<UNK>” for those words that appear less than 5 times in the training set.

5.2 Model Learning and Inference

In the experiments, the training samples are <sketchy scene, caption> pairs. That is, the captioning models are trained to generate a description for a sketchy scene. ResNet-101 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) is used as the sketchy scene encoder. Because of the domain gap between a natural image and a hand-drawn sketch, the fine-tuning of the encoder is turned on during training. For each sketchy scene, the size of the output feature from the fully-connected layer before softmax operation is 1,000, and the size of the features from the convolutional layer before the last average pooling layer is $14 \times 14 \times 2,048$. The number of LSTM cell units is set to 512. The dimension of word embedding is set to 512. Adam (Kingma and Ba, 2015) is used as the model optimizer. Dropout (Srivastava et al., 2014) and early stopping are exploited to achieve model regularization. The BLEU (Papineni et al., 2002) score on validation set is used for model selection. The initial learning rate is set to 0.0004 with a 0.5 decay ratio. The batch size is set to 40. Beam search is used for inference with a beam size of 5. The weighting coefficients λ_T and λ_W in Eq. (9) are both set to 1.

5.3 Quantitative Evaluation

We first verify the effectiveness of sentence-level sketchy scene captioning. Because the visual information of a sketchy scene is sparse and ambiguous, we explore how the representation of a sketchy scene affects the model performance by considering two factors: 1) visual feature (“FC” and “CV” denote the output features from the fully-connected layer and the convolutional layer, respectively); and 2) visual attention (“ATT” and “NAT” indicate that visual attention is turned on and off, respectively). It should be noted that visual attention is not conducted for the visual feature from the output of the fully-connected layer because a sketchy scene is simply encoded as a fixed-length vector in this case. The name of a model is denoted as the combination of the two factors above. The BLEU@n ($B@n$, $n=1, 2, 3, 4$), METEOR (M) (Denkowski and Lavie, 2014), ROUGE-L (R) (Lin, 2004), and CIDEr (C) (Vedantam et al., 2015) scores on testing set are reported in Table 1. These scores are computed by the *MSCOCO* captioning evaluation tool¹.

¹<https://github.com/tylin/coco-caption>

Model	$B@1$	$B@2$	$B@3$	$B@4$	M	R	C
<i>FC_NAT</i>	25.6	16.2	11.0	7.8	10.3	31.1	39.0
<i>CV_NAT</i>	29.0	21.1	16.9	14.4	14.2	25.2	58.4
<i>CV_ATT</i>	37.6	25.3	18.0	13.0	14.9	31.9	59.5

Table 1: The experimental results of sentence-level sketchy scene captioning on the new dataset.

Two aspects can be observed from Table 1. **First**, model *CV_NAT* performs better than model *FC_NAT* across all the metrics except R . This indicates that the fine-grained visual representation from the convolutional layer can better characterize the visual content of a sketchy scene compared to the visual representation from the fully-connected layer. The reason behind is that the visual information of a sketchy scene is sparse and the pooling operation before the fully-connected layer causes too much information loss. Differently, the output from the convolutional layer can preserve more discriminative local features of a sketchy scene. With such discriminative details, the captioning model can better recognize the objects in a sketchy scene, which further helps the model infer the correct interactions among the objects in the sketchy scene. **Second**, model *CV_ATT* achieves higher scores than those of model *CV_NAT* across all the metrics except $B@4$, which is mainly due to the precise visual features produced by the attention mechanism. Specifically, the information loss of a sketchy scene still exists when its initial representation is obtained by averaging the spatial visual features. Meanwhile, its visual details are gradually forgotten by the captioning model as the process of caption generation goes on. However, with the help of visual attention, the captioning model can be guided with fine-grained visual details by focusing on the relevant regions when generating words, which helps alleviate the problem of forgetting.

We also conduct the experiments on paragraph-level sketchy scene captioning, and the results are reported in Table 2. It can be observed that the $B@n$, M , and R scores are comparable to the best results of the sentence-level captioning models, while the C score is worse. Considering the definitions of these metrics, our captioning model can generate a relatively fluent paragraph-level description with correct semantics. However, the captioning model may sometimes focus on the wrong key points of the sketchy scene, which are not consistent with those identified by a person to a certain degree.

Model	$B@1$	$B@2$	$B@3$	$B@4$	M	R	C
<i>H_LSTM</i>	43.6	28.4	19.8	14.3	17.0	33.4	30.8

Table 2: The experimental results of paragraph-level sketchy scene captioning on the new dataset.

5.4 Qualitative Evaluation

We first have an analysis on the sentence-level captioning results. As shown in Figure 4, one generated sentence-level caption with the corresponding attention map sequence from model *CV_ATT* is given, where the attention maps highlight the regions that the captioning model learns to focus on at different time steps. It can be seen that what the caption describes matches the salient visual content of the sketchy

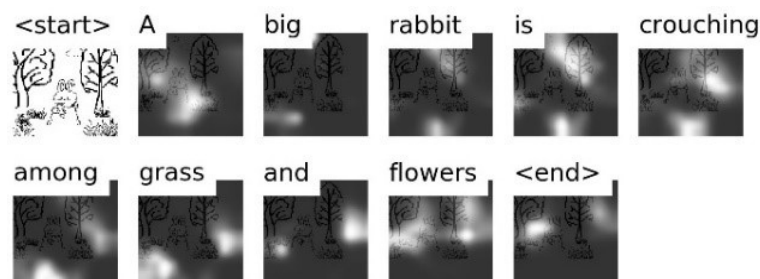


Figure 4: A generated sentence-level caption with the corresponding attention map sequence.

scene quite well. Specifically, “*a big rabbit*” can be generated correctly when the model focuses on the object “*rabbit*”. Meanwhile, the action “*crouching*” and the surroundings “*grass and flowers*” can be correctly recognized as well when the model focuses on the two sides of the sketchy scene. In addition, the model does not generate the description about the weather, such as “*on a cloudy day*”, which may be due to the reason that the “*cloud*” is too small to be salient enough. It is worth noting that the model does not generate the description about the “*trees*” which occupy a large area of the sketchy scene. As mentioned before, a sentence-level caption only describes the most salient parts of a sketchy scene. In the example, the “*rabbit*” has been treated as the salient object and the word “*rabbit*” may co-occur with “*grass*” and “*flowers*” more frequently in the dataset, and thus the “*trees*” are not treated as the salient objects by the captioning model.

Another three representative examples of the sentence-level captioning are given in Figure 5. Generally, the salient objects in the selected sketchy scenes can be well identified except the objects “*woman*” and “*house*” in the first one, and this bad result may be caused by the imprecise visual representation of the sketchy scene. That is, the “*woman*” is occluded by the “*tree*” in front of her, making the captioning model fail to recognize the “*woman*” correctly. At the same time, the mistaken object “*chicken*” usually co-occurs with the object “*fence*”, and the “*house*” is then ignored by the model. It can be observed that the actions of the salient objects, that is, “*standing*”, “*driving*”, and “*playing*”, can be generated properly. The reason is that, the recognized objects co-occur with the actions frequently in the dataset, which helps the model generate the correct actions for the salient objects. In the first example, the mistaken object “*chicken*” is usually followed by the action “*standing*”, which makes the action still correct compared to the ground truth caption. Besides, the descriptions about the weather (i.e., “*on a sunny day*”) in all the examples are generated correctly. The reason may be that, the “*sun*” is usually located on the top of a sketchy scene in the dataset, and thus the visual representation of the “*sun*” can be discriminative enough for the model to recognize the weather.

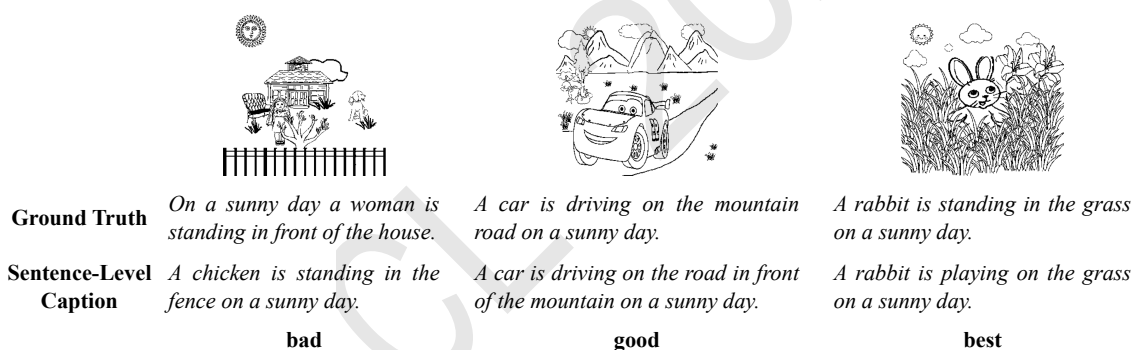


Figure 5: Three representative examples of sentence-level sketchy scene captioning.

In the following, we will have an analysis on two representative examples of paragraph-level sketchy scene captioning, as shown in Figure 6. It can be seen that both paragraphs are quite meaningful and describe many visual details of the sketchy scenes correctly. This indicates that the topic LSTM can give relatively correct guiding signals to the word LSTM and can stop the captioning process properly. Surprisingly, even the number of “*cars*” can be recognized correctly in the first example. These two examples show that it is promising to use a hierarchical LSTM model for paragraph-level sketchy scene captioning. There exist some problems in the results as well. For example, “*eight chicken*” is missed in the first example, and “*school bus*” and “*girls and boys*” are missed in the second example. Hence, how to learn a more discriminative sketchy scene representation and generate correct descriptions for the relatively small objects remains to be explored.

Because no prior work about sketchy scene captioning exists, we cannot compare our results with other methods. However, the qualitative experimental results show that the generated multi-level captions for a sketchy scene by our models are quite meaningful, which proves that the proposed Sketchy Scene Captioning task is feasible and deserves further exploration.

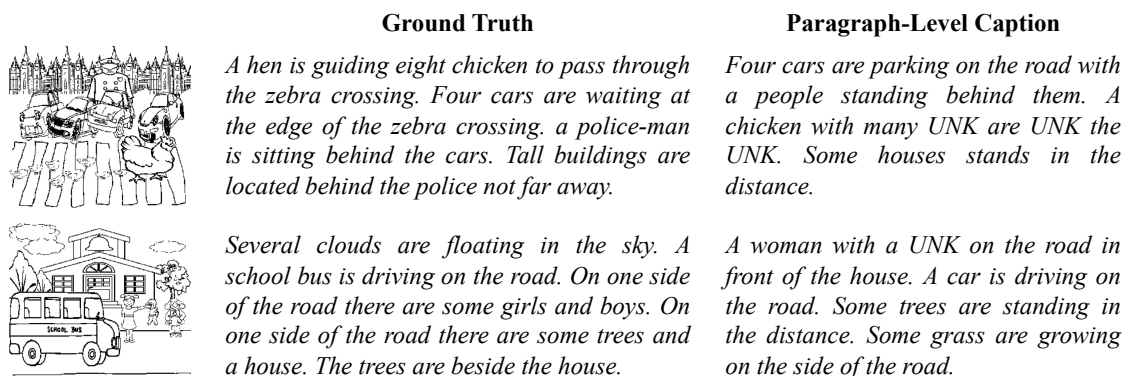


Figure 6: Two representative examples of paragraph-level sketchy scene captioning.

6 Conclusion and Future Work

In this paper, a new task termed Sketchy Scene Captioning is proposed. This task aims to generate multi-level descriptions for a sketchy scene through the sequence learning paradigm. To achieve the goal, a new dataset consisting of 1,000 sketchy scenes with the corresponding sentence-level and paragraph-level captions is created. The experimental results show that our captioning models can recognize the main objects in a sketchy scene and the interactions among these objects. This proves that it is feasible to generate multi-level captions for a sketchy scene. In the future, we plan to increase the volume of the dataset and explore how to learn a better representation of a sketchy scene for caption generation. We hope this work can inspire further researches on the better understanding of sketchy scenes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61976057 and No. 61572140) and Shanghai Municipal RD Foundation (No. 20511101203, No. 20511102702, No. 20511101403, No.19DZ2205700, and No. 2021SHZDZX0103). Yuejie Zhang was the corresponding author.

References

- Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *In Proceedings of the Annual Conference on Neural Information Processing Systems*, Pages 1889–1897.
- Changqing Zou, Haoran Mo, Ruofei Du, Xing Wu, Chengying Gao, and Hongbo Fu. 2018. LUCSS: Language-Based User-Customized Colourization of Scene Sketches. *arXiv preprint arXiv:1808.10544*.
- Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. 2018. SketchyScene: Richly-Annotated Scene Sketches. *In Proceedings of the European Conference on Computer Vision*, Pages 438–454.
- Chin-Yew Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. *In Proceedings of the Workshop on Text Summarization Branches Out (Post-Conference Workshop of ACL 2004)*, Pages 74–81.
- Desmond Elliott and Frank Keller. 2013. Image Description Using Visual Dependency Representations. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Pages 1292–1302.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *In Proceedings of the International Conference on Learning Representations*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *In Proceedings of the Annual Conference on Neural Information Processing Systems*, Pages 3104–3112.

- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A Large-Scale Hierarchical Image Database. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, Pages 248–255.
- Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, Pages 3337–3345.
- Jun-Yan He, Xiao Wu, Yu-Gang Jiang, Bo Zhao, and Qiang Peng. 2017. Sketch Recognition with Deep Visual-Sequential Fusion Model. *In Proceedings of the ACM on Multimedia Conference*, Pages 448–456.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, Pages 770–778.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *In Proceedings of the International Conference on Machine Learning*, Pages 2048–2057.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Pages 311–318.
- Mahmood Neshati, Zohreh Fallahnejad, and Hamid Beigy. 2017. On Dynamicity of Expert Finding in Community Question Answering. *Information Processing and Management*, 53(5):1026–1042.
- Mathias Eitz, James Hays, and Marc Alexa. 2012. How Do Humans Sketch Objects?. *ACM Transactions on Graphics*, 31(4):44:1–44:10.
- Michael J. Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *In Proceedings of the Workshop on Statistical Machine Translation (WMT@ACL 2014)*, Pages 376–380.
- Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, Pages 3156–3164.
- Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM Transactions on Graphics*, 35(4):119:1–119:12.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, Pages 6077–6086.
- Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2015. Sketch-A-Net that Beats Humans. *In Proceedings of the British Machine Vision Conference*, Pages 7.1–7.12.
- Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L. Yuille. 2019. Semantic-Aware Knowledge Preservation for Zero-Shot Sketch-Based Image Retrieval. *In Proceedings of the International Conference on Computer Vision*, Pages 3661–3670.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, Pages 4566–4575.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.
- Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *In Proceedings of the International Conference on Machine Learning*, Pages 448–456.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*.