CapOnImage: Context-driver Dense-Captioning On Image

Yiqi Gao\textsuperscript{1}; Xinglin Hou\textsuperscript{2}, Yuanmeng Zhang\textsuperscript{2}, Tiezheng Ge\textsuperscript{2}

\textsuperscript{1}School of Computer Science, Northwestern Polytechnical University
\textsuperscript{2}Alibaba Group

1gyqjz@mail.nwpu.edu.cn, 1peng.wang@nwpu.edu.cn
\{xinglin.hxl, zhangyuanmeng.zym, tiezheng.gtz, mengzhu.jyn\}@alibaba-inc.com

Abstract

Existing image captioning systems are dedicated to generating narrative captions for images, which are spatially detached from the image in presentation. However, texts can also be used as decorations on the image to highlight the key points and increase the attractiveness of images. In this work, we introduce a new task called captioning on image (CapOnImage)\textsuperscript{1}, which aims to generate dense captions at different locations of the image based on contextual information. For this new task, we introduce a large-scale benchmark called CapOnImage2M, which contains 2.1 million product images, each with an average of 4.8 spatially localized captions. To fully exploit the surrounding visual context to generate the most suitable caption for each location, we propose a multimodal pre-training model with multi-level pre-training tasks that progressively learn the correspondence between texts and image locations from easy to hard. To avoid generating redundant captions for nearby locations, we further enhance the location embedding with neighbor locations. Compared with other image captioning model variants, our model achieves the best results in both captioning accuracy and diversity aspects.

1 Introduction

Building upon the advances in computer vision and natural language processing areas, the new research direction called vision-and-language has attracted more and more attentions, which pushes to tackle new problems that need to bridge the two areas to advance the concept comprehension and reasoning capabilities.

The image captioning task, as one of the most classic vision-and-language tasks, aims to generate natural language descriptions for images (Vinyals et al., 2015; Anderson et al., 2018; Zhang et al., 2021; Johnson et al., 2016). However, the captions and the images in this task are spatially detached in presentation, without any association between each other (Figure 1(a)). In fact, there are many scenarios where the image and text are tightly associated. For example, the product images on e-commercial website (Figure 1(b)) usually contain descriptive texts, explaining multiple perspectives of the product (e.g., product characteristics, selling points, etc), which makes the image more informative and attractive. On the social media platform, users usually upload daily pictures with descriptive texts as decorations (Figure 1(c)). Therefore, it is significant to explore captioning on the image, which requires to consider not only the visual description, but also the text description placement on the image. Besides, to generate the informative captions in these scenarios, additional textual knowledge is usually needed, such as the product informa-
Therefore, in this work, we introduce a new task called **CapOnImage**, which aims to generate dense captions at different locations of the image based on contextual information. The CapOnImage task involves two steps, where the model needs to first predict a reasonable and aesthetic text layout (Arroyo et al., 2021; Gupta et al., 2020; Jyothi et al., 2019), and then generates a phrase or sentence for each text box. In this work, we simplify this task to generate captions for a provided list of text box locations, thereby removing the requirement for layout prediction. Therefore, the main focus of this work is to generate captions that are most suitable for the corresponding image locations.

The CapOnImage task involves two new challenges: (i) Better understanding of context: the captions at different locations can be greatly diverse. As shown in Figure 1(b), the texts around the product are descriptive captions describing the product features, while those at the bottom introduce the selling points. Therefore, the model needs to fully exploit the visual context around the text box to determine what caption is suitable to generate here. Since our task aims to generate captions on image, location context is vital for our task which is also validated on Table 2. Overall, compared with traditional caption task, the CapOnImage task needs better understanding of context. (ii) Caption redundancy: some texts can be suitable for adjacent locations. Therefore, if the model can only “see” the isolated text box without surrounding ones, it tends to generate the same caption for nearby text boxes because it suits all of them, thus causing the problem of caption redundancy.

To solve the aforementioned challenges, we propose a multi-modal pre-training and fine-tuning framework which contains multi-level pre-training tasks to effectively exploit multi-modal contextual information. **First**, to better exploit context, we design multi-level pre-training tasks to help the model “feel” the context. It explicitly equip the model with the ability to distinguish which captions are appropriate for the current location and image while which are not. Besides, inspired by the learning progression of easy to complex biological vision systems, we further propose a progressive training strategy which learns multi-level pre-training tasks from easy to hard. **Second**, to solve the problem of caption redundancy, we introduce a neighbor-enhanced location encoding module, which utilizes the surrounding text box locations as context, so that our model can “see” the adjacent context. We show the captioning diversity results with different ranges of adjacent text boxes involved in the location encoding module, and demonstrate the importance of such neighbor context.

In order to evaluate our model and benchmark progress in the CapOnImage task, we introduce the **CapOnImage2M** dataset. It contains 2.1 million product images crawled from an e-commercial website, and each image contains multiple spatially localized captions describing the associated product. We automatically acquire the text contents and their spatial locations from the image via OCR (Li et al., 2017; Liu et al., 2018), and finally collect 4.8 captions for each image on average. We also crawl the product title and attributes as additional context information for caption generation. With the empirical analysis on CapOnImage2M dataset, we show that the visual context, location information and the additional product information are beneficial for the caption generation, and our model can generate corresponding types of captions at different spatial locations (Figure 3). Furthermore, we demonstrate that our proposed neighbor-enhanced location encoding module and multi-level pre-training tasks significantly improve the captioning accuracy and diversity.

The main contributions of this work are as follows:

- We introduce a new vision-and-language task called CapOnImage, which requires the model to tightly associate image and texts as a whole.
- We analyze the challenges of CapOnImage task and propose a context enhanced model with progressive training strategy, which achieves the best result compared with other image captioning model variants.
- We propose a large-scale multi-modal dataset called CapOnImage2M, with 50 categories images and localized captions to support the CapOnImage research.

2 Related Work

In recent years, significant progress has been made in the image captioning task (Vinyals et al., 2015; Anderson et al., 2018; Huang et al., 2019; Pan et al., 2020; Zhang et al., 2021), which aims to describe the image content in one natural sentence. With
the advances in visual understanding abilities, researchers are not satisfied with generating dull and less informative captions and extend the traditional image captioning task along two directions.

The first direction is called dense captioning (Johnson et al., 2016; Melas-Kyriazi et al., 2018; Krishna et al., 2017a; Wang et al., 2021; Song et al., 2021), which targets to describe detailed visual content with a set of sentences. Johnson et al. (Johnson et al., 2016) propose a fully convolutional localization network to unify the object detection (Sermanet et al., 2014; Girshick et al., 2014; Ren et al., 2015) and image captioning in one framework to predict a set of descriptions across object regions. Krishna et al. (Krishna et al., 2017a) migrate it to the video, which aims to predict sequential event proposals and generate description for each clip. In these works, the dense captions deliver more details of visual content than traditional single sentence. The second direction is called text-aware image captioning (Biten et al., 2019; Sidorov et al., 2020; Yang et al., 2021), where the model generates captions not only according to the image, but also utilizes additional textual information as context. Besides, Sidorov et al. (Sidorov et al., 2020) and Gurari et al. (Gurari et al., 2020) propose to generate image captions with scene texts, which exploit OCR tokens as the textual context.

Although impressive progresses have been made along the two directions, they still remain separate. The proposed CapOnImage task can be considered as a combination of the two directions, where the model needs to first predict the text layout (spatial locations on the image) and then generate caption for each location conditioned on both the image and textual information.

There are several key distinctions between our proposed CapOnImage task and dense captioning: 1) Dense image captioning aims to generate captions for subregions within an image, and there is no length limitation of captions. However, our task is to generate text and affix it to specific regions within an image, and the text length should be controlled according to the region size. 2) The visual content is the only input of dense captioning. While for our task, there are three inputs: visual content, additional textual information, and the specified location to affix. All of them will impact the generated text content. 3) The generated text is a plain description of the visual content for dense captioning. But for our task, the generated text is on the image, making it more informative, together with the visual content.

3 CapOnImage2M Dataset

In this section, we introduce our proposed CapOnImage2M dataset, which is the benchmark for the CapOnImage task. We first present an overview of the dataset collection and statistics, and then compare it with other related image captioning datasets. A more detailed datasheet describing the motivation, composition, and recommended uses of our CapOnImage2M dataset following (Gebru et al., 2018) can be found in the Appendix A.

3.1 Dataset Collection and Statistics

The CapOnImage2M dataset contains 2.1 million product images crawled from a Chinese e-commercial website2, where each image contains both the product and descriptive captions describing the product features, efficacy, brand and so on. (detail information, e.g. word cloud, can be found in Appendix A.) For each image, we employ an OCR toolkit to recognize the texts and their spatial locations on the image, and remove the noise with high perplexities by a pre-trained GPT.

3.2 Comparison with Other Datasets

In Table 1, we compare our CapOnImage2M dataset with other image captioning datasets. The CapOnImage2M dataset contains 2.1 million product images crawled from a Chinese e-commerce website2, where each image contains both the product and descriptive captions describing the product features, efficacy, brand and so on. (detail information, e.g. word cloud, can be found in Appendix A.) For each image, we employ an OCR toolkit to recognize the texts and their spatial locations on the image, and remove the noise with high perplexities by a pre-trained GPT.

Table 1: Comparisons of different datasets. CI, IC and FC refer to Caption on Image, Image Caption and Fashion Caption.

| dataset     | #image | #text  | avg_len | dense | task |
|-------------|--------|--------|---------|-------|------|
| CapOnImage  | 2.1M   | 10.07M | 4.9     | ✓     | CI   |
| Flickr30K   | 30K    | 150K   | 12.3    | -     | IC   |
| MSCOCO      | 123K   | 616K   | 10.4    | -     | IC   |
| VG          | 108K   | 5M     | 5.7     | ✓     | IC   |
| TextCaps    | 28K    | 142K   | 12.4    | -     | IC   |
| FACAD       | 993K   | 130K   | 21.0    | -     | FC   |

2https://taobao.com
4 Model

In this section, we introduce our CapOnImage method based on the pre-training and fine-tuning framework as illustrated in Figure 2. First, we introduce the multi-modal representation of visual, location coordinates, and product information of the given input images. Then, a progressive training strategy with multi-level pre-training tasks is proposed to enhance the correspondence learning between textbox locations and captions for caption generation with a multi-layer transformer.

4.1 Input Representation

The inputs of our model include three parts from different modalities: the visual image, the textbox location coordinate and the textual product information. We independently encode each modality input as a sequence of \( d \)-dimensional feature vectors as follows.

**Image representation.** Given the image, we extract the grid features with standard ResNet-50 (He et al., 2016) backbone, which is further end-to-end fine-tuned with our model. We flatten the \( k \times k \) feature map into a sequence and add spatial position embedding similarly as DETR (Carion et al., 2020). Specifically, for the \( i \)-th grid whose horizontal and vertical indexes are \( x_i \) and \( y_i \), we add learnable spatial embedding and segment embedding which indicates the image modality to the appearance feature \( v_i \) as follows:

\[
\hat{v}_i = v_i + \{\text{Emb}_h(x_i); \text{Emb}_v(y_i)\} + \text{SE}_{v_i},
\]

where \( \text{Emb}_h(\cdot) \) and \( \text{Emb}_v(\cdot) \) are horizontal and vertical embedding layers with the output dimension of \( \frac{d}{2} \), \([;]\) denotes concatenation and SE denotes segment embedding. Finally, we represent the image with a sequence of patch features \( \hat{V} = \{\hat{v}_1, \ldots, \hat{v}_{k\times k}\} \).

**Neighbor-enhanced location representation.** We represent a textbox location with 2D coordinates \( \{(x_{\text{min}}, y_{\text{min}}), (x_{\text{max}}, y_{\text{max}})\} \), where \( (x_{\text{min}}, y_{\text{min}}) \) is the top left corner coordinate and \( (x_{\text{max}}, y_{\text{max}}) \) is the bottom right corner coordinate. We map the real value coordinates into the \( k \times k \) grid and represent them with the same spatial position embeddings as image:

\[
e_{\text{cur}} = [\text{Emb}_h(x_i); \text{Emb}_v(y_i); \text{Emb}_h(x_j); \text{Emb}_v(y_j)],
\]

where \( \{(x_i, y_i), (x_j, y_j)\} \) is the corresponding grid index.

Furthermore, to avoid the problem of caption redundancy for adjacent locations, we enhance the location representation with neighbor locations as context. We define the distance of two text boxes as the distance of their centers, the text boxes whose upper left corner are with smaller value of x-coordinate plus y-coordinate than the current one as the previous text boxes, and those larger than the current one as the next text boxes. Then, we employ the nearest previous textbox location and the nearest next textbox location as the neighbor context, and encode them similarly as \( e_{\text{cur}} \). After encoding, we concatenate them with the current location embedding and add a segment embedding indicating the location modality as follows:

\[
l = [W_1^T \cdot e_{\text{prev}}; e_{\text{next}}]; W_2^T \cdot e_{\text{cur}}] + \text{SE}_t,
\]

where \( W_1 \in \mathbb{R}^{d \times d} \) and \( W_2 \in \mathbb{R}^{d \times d} \) are learned matrices, \( e_{\text{prev}} \) and \( e_{\text{next}} \) are the neighbor location embeddings, SE is the segment embedding.

**Product information representation.** To generate informative product descriptions, we also exploit product information as the textual context, which is the product title and attribute in this work. We concatenate them with a special \(<\text{SEP}>\) token. Given the product information \( X = \{x_1, \cdots, x_K\} \) with \( K \) words, we embed these words via the same word embedding matrix as the target caption words, and add positional and segment embeddings as follows:

\[
w_{i}^{\text{info}} = W_e \cdot x_i + \text{PE}_i + \text{SE}_x,
\]

where \( W_e \) is the word embedding matrix, PE denotes sequence positional embedding as in BERT (Devlin et al., 2019) and SE denotes segment embedding. Finally, we represent the product title with a sequence of \( d \)-dimensional feature vectors as \( W^{\text{info}} = \{w_{k}^{\text{info}}\}^{K}_{k=1} \).

4.2 Pre-training Tasks and Strategy

After encoding each input modality into the common embedding space, we employ transformer layers on the multi-modal input to fuse the multimodal information. To generate appropriate and diverse descriptions at different textbox locations,
Figure 2: Illustration of our model with four input modalities: image patches, location coordinates, product information, and the predicted text tokens. Two pre-training tasks are employed to optimize the model with a progressive learning strategy from easy to difficult. "=" denotes parameter sharing and "+" denotes addition. We add English translation for product information for better understanding.

we pre-train the model with two pre-training tasks, including Caption Generation (CG) and Caption Matching (CM). We first pre-train the model with both CG and CM tasks, and then fine-tune it only with the CG task for the final caption generation.

**Task #1: Caption Generation (CG).** We generate captions using the same multi-modal transformer layers as decoder following the prefix LM (Raffel et al., 2020; Dong et al., 2019). Each word prediction can attend to all the image features, neighbor-enhanced location and product information embeddings, as well as previous generated words. We adopt the auto-regressive training objective for the CG task, which can be expressed as follows:

$$\mathcal{L}_{CG} = -\frac{1}{T} \sum_{t=1}^{T} \log p(y_t^* | y_{<t}, \tilde{V}, l, W^{\text{info}}, \Theta),$$

where $y_t^*$ denotes the $t$-th word of ground-truth caption for the current textbox location, and $\Theta$ denotes all learnable parameters of the pre-training model. During the inference phase, we first encode the image, location and product information embeddings, and then feed a special start token [SOS] to predict the caption word by word.

**Task #2: Caption Matching (CM).** To help the model learn which captions are appropriate for the current image and location while which are not, we further introduce another pre-training task called Caption Matching. It is similar to the ITM task commonly used in vision-and-language pre-training models (Chen et al., 2020; Li et al., 2020; Lu et al., 2019; Zhuge et al., 2021), which requires the model to predict if the image and caption are semantically aligned. A score $s$ between 0 and 1 is predicted by the hidden output of the [SOS] token. The positive examples of this task are corresponding pairs in the dataset, while the negative examples can be diverse. In this work, we design three levels of negative example construction and progressively learn the task from easy to difficult.

**Level-I: Image caption matching.** The first negative level is to randomly replace the correct caption with descriptions of other images. Therefore, it is not consistent with the current image content. We expect the model can recognize such negative examples according to the visual image and product information, which are the easiest negative cases.

**Level-II: Location caption matching.** The second negative level is to replace the caption with those in other locations of the same image. It is more difficult than the Level-I because the negative caption exactly describes the current image but is not suitable for the current location. For example, the product efficacy descriptions may be inappropriate to appear on the left corner of the product image, while the product brand is more suitable. We expect the model can learn the relationship of texts and textbox locations according to the surrounding visual context.

**Level-III: Neighbor-location caption matching.** Since the captions in neighbor locations are the most confusing samples, we further introduce the third negative level, where we randomly replace the caption with those in neighbor locations, including the nearest previous location and the nearest next location defined in Section 4.1. It can be seen as a
Table 2: We report BLEU (B), METEOR (M), CIDEr and Diversity (D) scores for the captioning on image task on the CapOnImage2M dataset. Since the CapOnImage task is a newly proposed task in this work, we adapt conventional state-of-the-art image captioning models to this task by introducing text location and textual knowledge for comparison. We run our experiments 5 times under different random seed and report the average value.

| Methods              | Validation | Text |
|----------------------|------------|------|
|                      | B@1       | B@4 | M    | CIDEr   | D@1 | D@2 | B@1 | B@4 | M    | CIDEr   | D@1 | D@2 |
| Up-down w/o TAtt     | 15.51     | 9.79 | 7.93 | 108.11 | 48.73 | 41.22 | 13.48 | 8.71 | 7.89 | 99.52 | 46.53 | 39.51 |
| M2 w/o TAtt          | 24.32     | 20.96 | 14.58 | 200.32 | 54.83 | 49.29 | 23.19 | 18.74 | 12.16 | 184.34 | 54.11 | 47.23 |
| RSTNet w/o TAtt      | 25.18     | 20.46 | 14.73 | 196.91 | 56.17 | 50.85 | 22.87 | 19.35 | 11.81 | 180.50 | 55.36 | 46.86 |
| Up-down w/ TAtt      | 20.49     | 13.54 | 11.52 | 181.04 | 65.24 | 56.27 | 18.94 | 12.30 | 10.81 | 166.26 | 65.72 | 56.48 |
| M2 w/ TAtt           | 34.18     | 24.31 | 18.14 | 273.35 | 63.29 | 53.43 | 31.63 | 22.11 | 17.71 | 265.22 | 62.81 | 54.19 |
| RSTNet w/ TAtt       | 33.42     | 23.91 | 17.28 | 267.60 | 64.12 | 54.53 | 30.54 | 20.94 | 17.20 | 259.29 | 63.10 | 55.69 |
| M4C w/o copying      | 36.46     | 27.08 | 20.19 | 296.73 | 65.69 | 55.69 | 35.73 | 26.24 | 19.8 | 287.61 | 64.31 | 55.01 |
| M4C w/ copying       | 35.98     | 28.35 | 20.58 | 299.31 | 65.98 | 55.03 | 36.23 | 27.15 | 20.01 | 288.35 | 64.03 | 55.23 |

where \( s \) refers to the predicted matching score of a training sample and \( r \in [0, 1] \) is the ground-truth label indicating whether it is a negative or positive sample.

5 Experiments

Evaluation metrics. For the accuracy measurement, we evaluate the generated captions against the ground-truth with standard metrics used in the image captioning task, including BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015). For the diversity measurement, we concatenate the dense captions within an image as a paragraph and compute the ratio of unique \( n \)-grams, called Div@n (Shetty et al., 2017). For the fitness measurement, we show the relationship of generated caption length to the aspect ratio of text box, and...
Table 3: Captioning results with different pre-training tasks and strategies. *fixed* denotes training the multi-level CM task with a fixed proportion, while *progressive* denotes varying the proportion according to the degree of difficulty and training steps.

| Row | Pretrain tasks | Pretrain strategy | Validation | Test |
|-----|----------------|-------------------|------------|------|
|     | Level-I | Level-II | Level-III | fixed | progressive | B@1 | B@4 | M | CIDEr | B@1 | B@4 | M | CIDEr |
| 1   | -      | -      | -      | -      | -      | 37.69 | 27.98 | 21.91 | 313.64 | 37.02 | 27.23 | 21.14 | 305.50 |
| 2   | ✓      | -      | -      | ✓      | -      | 38.87 | 29.23 | 22.32 | 325.74 | 38.19 | 28.48 | 21.69 | 316.46 |
| 3   | ✓      | ✓      | -      | ✓      | -      | 40.22 | 30.36 | 22.88 | 339.72 | 39.36 | 29.54 | 22.03 | 329.77 |
| 4   | ✓      | ✓      | ✓      | ✓      | -      | 40.55 | 30.71 | 23.34 | 347.30 | 39.40 | 29.57 | 22.51 | 335.49 |
| 5   | ✓      | ✓      | ✓      | ✓      | -      | 41.77 | 32.20 | 24.52 | 357.03 | 40.95 | 31.45 | 23.49 | 345.41 |

the type distribution of generated captions.

**Implementation details.** We initialize the ResNet-50 backbone pre-trained on ImageNet, and fine-tune it with our model in an end-to-end manner. Our model has $L = 6$ transformer layers with the hidden dimension of $d = 1024$ and attention head $A = 8$. In the pre-training stage, we sample the batch of CG and CM tasks with a proportion of 3:1 for 200K steps. We adopt a warming-up strategy for the first 4K steps. For text processing, we tokenize Chinese captions into characters and build a vocabulary with 6263 tokens. We implement our method using pytorch (Paszke et al., 2019). We manually search hyper-parameter.

### 5.2 Comparison with Baseline Models

Since the CapOnImage task is a newly proposed task in this work, we adapt conventional state-of-the-art image captioning models (Up-down (Anderson et al., 2018), RSTNet (Zhang et al., 2021), M2 (Cornia et al., 2020), M4C-Captioner (Sidorov et al., 2020)) to this task as the baselines for comparison. The details of compared baseline methods are expanded on the supplementary materials.

**Variants of our model.** We also compare with different variants of our model. Since all the words to be generated are already in the vocabulary, the copy mechanism bring no significant improvement (Table 2), so we remove it and use M4C-Captioner w/o copying as our baseline model. The no-info model adopts the same architecture as the baseline model except that the product information input is removed. Similarly, the no-image and the no-locations model share the same baseline model architecture but with the image and locations input removed respectively. Our context model is the baseline model enhanced with neighbor location contexts, which is still trained only with the CG task. The full model is our complete model with progressive pre-training by both CG and CM tasks on the same dataset.

Table 2 reports the captioning on image results of different models on the CapOnImage2M validation and test sets. It is shown that the conventional image captioning model without any adaptation perform poorly on the CapOnImage task. This is because these models lack the textual context that can provide rich information for caption generation. Enhancing the Up-down, $M^2$, and RSTNet with textual attentions on the additional product information, the captioning results are significantly improved. However, they are still inferior to the adapted text-aware image captioning model, which has a good ability of multi-modal fusion with the cross transformer encoder. Therefore, it stands for a strong baseline for our model. Compared with the baseline model, our context model enhances the text location embedding with neighbor location contexts, which brings significant improvements on both accuracy and diversity metrics. It demonstrates the importance of location relationship modeling especially for reducing caption redundancy. Although good results have been achieved, the model is only trained with caption generation objective against the ground-truth, which is not sufficient to help the model learn complex correspondence between texts and image locations. Therefore, when pre-training the context model with both CG and CM tasks in a progressive manner, our full model achieves the state-of-the-art results. Nevertheless, there is still a gap with the human annotations on the captioning diversity metrics.

To further explore the contribution of each input modalities, we also report the captioning results with some input removed (no-locations, no-info, and no-image). It shows that the location information is more important than the visual image and the textual information for the CapOnImage
task. However, these three models are severely inferior to the baseline model with multi-modal input, which shows the necessity of multi-modal fusion for this new task.

### 5.3 Ablative Analysis

#### Parameters determination.

In Figure 4, we conduct ablation studies to investigate the suitable parameters for our model. We use our context model and operate our experiment on test set. The parameter that need to be determined are number of layers of transformer, hidden dimension of transformer and grid feature size of resnet backbone. In Figure 4(a), we study the impact of these three parameters for caption performance (BLEU@4 and CIDEr). We choose 6 layer transformer with hidden dimension 1024 and 8×8 grid size resnet for the intuition of Accuracy-Efficiency Trade-Offs.

#### Pre-training tasks and strategy.

In Table 3, we ablate the proposed multi-level pre-training tasks and progressive training strategy. It shows that pre-training with only Level-I of the CM task (row 2) can significantly improve the non-pretrained model with only CG objective (row 1). It demonstrates the importance of multi-modal alignment to the CapOnImage task. Upgrading the CM task with more difficult negatives in Level-II helps the model better learn the relationship of captions and text locations and thus yields better results (row 3). Further incorporating negatives in Level-III bring additional gains (row4), which confirms the importance of context information in CapOnImage task. However, since three levels of negative samples are built from easy to difficult, we seek to boost the learning process of CM task in an adaptive fashion: the ratio of pre-training tasks need to be adapted to the training status and vary in a progressive manner (as opposed to a fixed proportion of 30%:40%:30%). Therefore, we propose a progressive training strategy with the proportion of easy task decreased and hard task increased in the training process. It boosts the results stably (row 5).

### 5.4 Caption length to textbox aspect ratio.

Given a textbox location, the generated caption should exactly fit in with it for visual aesthetic. The text length and font size are the influencing factors. Since the short side of the textbox determines the font size, the aspect ratio (long side length / short side length) can reflect the most suitable text length.

### 5.5 Caption type to textbox location.

Besides the caption length, the types of captions in CapOnImage2M dataset are diverse at different image locations. To explore whether the type of our generated captions is suitable to the given locations, we visualize the caption type distribution on the image. Since the caption type annotations are not available, we automatically group the ground-truth captions into 4 categories by k-means based on their sentence-level BERT (Devlin et al., 2019) embeddings. We then display the same type of captions using the same color on an image. As shown in Figure 3(a), the captions with the same type are located together, which shows that the caption type is very related to its location. We assign our generated captions to the 4 clusters and visualize them in the same way in Figure 3(b). It looks very similar to the ground-truth type distribution map, which shows that our model effectively learns the relationship of text location and text type. The meaning of each color are illustrated in Figure 3(c).

### 5.6 Human Evaluation

In addition to the objective evaluation, we also conduct human evaluation on 400 randomly sampled images from the test set. We render the generated dense captions from baseline model and our full model on the image via opencv. We instruct 5 workers to choose which one is better or they are not distinguishable based on relevance, diversity and informativeness respectively and do the majority voting. To avoid the prior bias, we anonymize the model names and shuffle the predictions randomly. Table 4 shows the human evaluation results. Our full model significantly outperforms

|                      | Base wins (%) | Full wins (%) | Delta  |
|----------------------|--------------|---------------|--------|
| relevance            | 31.2         | 68.8          | +37.6  |
| diversity            | 34.5         | 65.5          | +31.0  |
| informativeness      | 32.8         | 67.2          | +34.4  |
the baseline model especially on all three aspects, which demonstrates the effectiveness of the proposed neighbor-enhanced location embedding and multi-level progressive pre-training.

5.7 Qualitative Results

Figure 5 visualizes some results of our full model and baseline model. The baseline model is shown to generate repetitive captions due to the lack of global layout awareness. For example, for adjacent locations, the baseline model repeats the concept of “mild”, while our model generates more informative caption of “sensitive skin friendly”. Furthermore, our model is also shown to better exploit the visual context to generate more suitable captions. In the second example, our model generates the text “suitable for large area makeup” for the down-right region where a hand appears, while the baseline model fails to distinguish it with the up-right region and generates similar descriptions about the “oblique slop brush”. More visualization results can be found in the supplementary material.

6 Conclusion

In this work, we propose a new vision-and-language task called CapOnImage, which aims to generate dense captions at different locations on an image with visual and textual context. We propose a multi-modal pre-training and fine-tuning model with multi-level pre-training tasks from easy to difficult for the correspondence learning between image location and text, and enhance the current location embedding with neighboring locations to reduce captioning redundancy. Experimental results show that our model can generate controllable length and type of captions at different image locations. In the future work, we will explore to generate dense captions with self-predicted text layout and combine the layout generation with caption generation in one joint framework to benefit from each other.
Limitations

The definition of our proposed task, i.e., generating text on image locations based on visual and textual context, can be found in many scenarios, such as billboard photos, posters, social platform images, etc. In this paper, we only report performance on our collected e-commercial dataset for the convenience of validating our key idea and our proposed task does not rely on any priors of specific inputs, so it can be expanded to a wide range of scenarios. In the future, we plan to collect more types of datasets, which can help us to apply our approach to more scenarios. Also, our dataset only contains caption annotations in Chinese.

Acknowledgement

This work was supported by National Key R&D Program of China (No. 2020AAA0106900), the National Natural Science Foundation of China (No. U19B2037, No. 61876152), Shaanxi Provincial Key R&D Program (No. 2021KWZ-03), Natural Science Basic Research Program of Shaanxi (No. 2021JCW-03) and Alibaba Group through Alibaba Innovative Research Program. We thank Yuqing Song, Wei Suo, Mengyang Sun, and Peng Wu for their helpful discussion.

References

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Diego Martín Arroyo, Janis Postels, and Federico Tombari. 2021. Variational transformer networks for layout generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13642–13652.

Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furui Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Annual Conference on Neural Information Processing Systems*, pages 13042–13054.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.

Kamal Gupta, Alessandro Achille, Justin Lazarow, Larry Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2020. Layout generation and completion with self-attention. *CoRR*, abs/2006.14615.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on attention for image captioning. In *IEEE International Conference on Computer Vision*, pages 4633–4642.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.
Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. 2019. Layoutvae: Stochastic scene layout generation from a label set. In IEEE International Conference on Computer Vision, pages 9894–9903.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In IEEE International Conference on Computer Vision, pages 706–715.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li-Fei Fei. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis., 123:32–73.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, pages 11336–11344.

Hui Li, Peng Wang, and Chunhua Shen. 2017. Towards end-to-end text spotting with convolutional recurrent neural networks. In IEEE International Conference on Computer Vision, pages 5248–5256.

Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. 2018. FOTS: fast oriented text spotting with a unified network. In IEEE Conference on Computer Vision and Pattern Recognition, pages 5676–5685.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Annual Conference on Neural Information Processing Systems, pages 13–23.

Luke Melas-Kyriazi, Alexander M. Rush, and George Han. 2018. Training for diversity in image paragraph captioning. In Conference on Empirical Methods in Natural Language Processing, pages 757–761.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In IEEE Conference on Computer Vision and Pattern Recognition, pages 10968–10977.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.
A Appendix

In the appendix, we first conduct further analysis of the choices of contextual locations and the diversity of our generated captions. Then claim the motivation and the challenge for the novel CapOnImage task. At last, we take hierarchical presentation of the CaptionOnImage2M from different perspectives.

A.1 Choice of contextual locations.

Table A5: Captioning results on the test set with different contextual locations. The number in () means how many locations used as the context to enhance the current location embedding.

| Methods                  | B@1    | B@4    | CIDEr  | D@1   | D@2   |
|--------------------------|--------|--------|--------|-------|-------|
| w/o context (0)          | 35.73  | 26.24  | 287.61 | 69.31 | 59.01 |
| w/ two random (2)        | 33.98  | 24.79  | 283.73 | 67.43 | 57.82 |
| w/ top-1 nearest (2)     | 37.02  | 27.23  | 305.50 | 70.98 | 60.90 |
| w/ top-2 nearest (4)     | 37.88  | 27.10  | 307.61 | 70.12 | 60.34 |

In Table A5, we take a further study on the neighbor-enhanced location embedding module with different contextual locations. With the nearest neighbor (previous and next) locations used as the context as described in Section 4.1, our model significantly improves the accuracy and diversity metrics. To figure out where the benefit comes from, we compare with the model using the same amount of randomly selected locations as context. Experimental results show that the randomly selected locations cannot improve the results and may even bring noise, which demonstrates the effectiveness of our model in encoding neighboring layout information to generate more appropriate and diverse captions. When further expanding the contextual range from the top-1 nearest to the top-2 nearest (top-2 previous and top-2 next), the model achieves slightly better result on the accuracy metric. To balance the efficiency and quality, we finally use the top-1 nearest locations as the context in our model.

A.2 Diversity from the textual input.

In Table A6, we calculate the Bleu score between the input product information and our generated captions. Results show that there is only a small percentage of copy text in our generated captions, demonstrating that our model is not just simply “copying text from the input product information”, but generating diverse captions conditioned on the multi-modality input.

Table A6: Bleu score between the input product information and our generated captions.

| #  | Bleu1 | Bleu2 | Bleu3 | Bleu4 |
|----|-------|-------|-------|-------|
| Test| 0.032 | 0.021 | 0.013 | 0.007 |

A.3 Motivation

For what purpose was the dataset created? The dataset was created to support the research on the captioning on image (CapOnImage) task, which aims to generate informative captions at different appropriate locations in the given image. CapOnImage is a valuable task for both vision-and-language research and industrial applications. We show the pipeline of our task on Figure A6.

A.4 Composition

What do the instances that comprise the dataset represent? Each instance in the CapOnImage2M dataset contains 50 categories product image, a sentence of product title, several product attributes, and multiple spatially localized captions with bounding box coordinates, describing the product from multiple aspects. We show the word cloud of product categories on Figure A7. Figure A8 shows some examples of the CapOnImage2M dataset.

How many instances are there in total? The dataset consists of 2.1M images and 10.07M texts in total. Each image contains an average of 4.8 spatially localized captions.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? CapOnImage2M is a new independent dataset. The instances in CapOnImage2M dataset are crawled from an e-commercial website. New product images with texts will continue to emerge. Therefore, the cur-
Figure A6: Pipeline of our CapOnImage task. We focus on generating captions for a provided list of text box locations in this work.

Figure A7: Word Cloud of 50 product categories. We show the Word Cloud according to the instances number of each category.

rent version of the dataset does not contain all possible instances.

What data does each instance consist of? Raw data (e.g., unprocessed text or images) or features? Each instance contains a product image with short side as 256 pixel in PNG format, the processed product title and captions. The original image with high resolution can be downloaded by the provided image URL.

Is any information missing from individual instances? No. Everything is included in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes. The dataset is split into 2.06M images for training, 20K for validation, and 20K for testing.

Are there any errors, sources of noise, or redundancies in the dataset? We have removed redundant or similar images whose captions are overlapped over a threshold. Therefore, the instance redundancy in the dataset is low. The captions are automatically recognized by an OCR model, therefore, there are noise in the captions of training set. To ensure the evaluation accuracy, we further manually clean the OCR errors for the validation and testing sets.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is self-contained, except that the original high resolution images are linked to a public website. Nevertheless, the self-contained images are enough to reproduce the results, and the high resolution images are used for better visualization.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? No. All data was collected from a publicly available e-commercial website.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No. The dataset only consists of product images.

Does the dataset relate to people? Most of the images only contain products, and only a few images contain public product spokesmen.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No. The dataset does not contain confidential information since all information was crawled from a public e-commercial website.
Figure A8: Examples of the CapOnImage2M dataset. We add the English translation for better understanding.
Figure A9: Qualitative results of our full model on the CapOnImage2M test set. We add the English translation for better understanding.
A.5 Collection Process

What mechanisms or procedures were used to collect the data? The raw images and product title sentences in this dataset were automatically crawled from Taobao website 3. The spatially localized captions are extracted from the image by an OCR model, and further manually cleaned for the validation and testing sets.

A.6 Preprocessing

Was any preprocessing/cleaning/labeling of the data done? The following steps were taken to process the data: (1) Crawling raw images and product titles. We first crawl the raw product images and titles from the e-commercial website, and then resize the images with short side as 256. (2) Detecting texts on the image. For each image, we employ an OCR toolkit to automatically detect the texts on the image as well as their bounding box coordinates. (3) Removing redundant instances. We remove redundant instances whose images contain similar captions that exceed the overlap threshold. (4) Removing discount information. We remove redundant instances that have high correlation with discount information. (5) Cleaning instances. We remove the texts longer than 10 or shorter than 2 characters, and remove the instances with only one caption on the image. Then, we input the automatically recognized captions into a pre-trained GPT model, and remove the captions with high generation perplexities. (6) Manual labeling. We further manually clean the captions with OCR errors in the validation and testing sets for accurate evaluation.

3https://taobao.com

Is the software used to preprocess/clean/label the instances available? Yes. All software used to process the data is open source and has been mentioned above.

A.7 Uses

Has the dataset been used for any tasks already? No, the dataset is newly collected from scratch in this work to support the proposed new task.

What (other) tasks could the dataset be used for? The dataset was created to support the CapOn-Image task. In addition, since each product image is accompanied by a product title sentence, it can be directly used for the Fashion Captioning (Yang et al., 2020) task. It may support a wider range of vision-and-language tasks as well.

A.8 Distribution

We show the distribution of caption length and number of captions per image in Figure A10.

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes. The dataset will be released publicly.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset can be downloaded from Google Drive and Baidu Disk as a gzipped tar file.

When will the dataset be distributed? The dataset will be released upon the publication of this work.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? There will be no license. Users only need to fill in an agreement form regarding the dataset not to be used for commercial purposes and citation suggestions etc.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No. There are no fees or restrictions.

A.9 Maintenance

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes. The dataset will be updated for fair com-
parison with future works if there is any kind of changes.