applied sciences

MDPI

Article

Panoptic Segmentation-Based Attention for Image Captioning

Wenjie Cai [1], Zheng Xiong [1], Xianfang Sun [2], Paul L. Rosin [2], Longcun Jin [1,*] and Xinyi Peng [1]

[1] School of Software Engineering, South China University of Technology, Guangzhou 510006, China; sewcai@mail.scut.edu.cn (W.C.); 201730684281@mail.scut.edu.cn (Z.X.); adxypeng@scut.edu.cn (X.P.)
[2] School of Computer Science and Informatics, Cardiff University, Cardiff CF10 3AT, UK; SunX2@cardiff.ac.uk (X.S.); RosinPL@cardiff.ac.uk (P.L.R.)
* Correspondence: lcjin@scut.edu.cn

Received: 12 December 2019; Accepted: 1 January 2020; Published: 4 January 2020

Abstract: Image captioning is the task of generating textual descriptions of images. In order to obtain a better image representation, attention mechanisms have been widely adopted in image captioning. However, in existing models with detection-based attention, the rectangular attention regions are not fine-grained, as they contain irrelevant regions (e.g., background or overlapped regions) around the object, making the model generate inaccurate captions. To address this issue, we propose panoptic segmentation-based attention that performs attention at a mask-level (i.e., the shape of the main part of an instance). Our approach extracts feature vectors from the corresponding segmentation regions, which is more fine-grained than current attention mechanisms. Moreover, in order to process features of different classes independently, we propose a dual-attention module which is generic and can be applied to other frameworks. Experimental results showed that our model could recognize the overlapped objects and understand the scene better. Our approach achieved competitive performance against state-of-the-art methods. We made our code available.

Keywords: image captioning; attention mechanism; panoptic segmentation

1. Introduction

Image captioning, the task of automatically generating natural language descriptions of images, has received increasing attention in computer vision and natural language processing. This task has several important practical applications. For example, it can help people with visual impairments. Therefore, it requires accurate recognition of the objects and a thorough understanding of the images. With the advances in deep neural networks, image captioning models now tend to use the "encoder-decoder" framework. In this framework, a convolutional neural network (CNN) is used to encode images into vectors, and a recurrent neural network (RNN) or one of its variants LSTM [1], is used to generate captions step by step.

The main problem in image captioning is the coarse representation of images. In the vanilla encoder-decoder framework, the encoder simply compresses an entire image into a global representation. This representation is coarse and has two drawbacks. First, it is fixed and thus does not correspond to the dynamic decoding process of caption generation. Second, it does not contain the spatial structures of the image. In order to obtain a fine-grained image representation, visual attention mechanisms [2–6] have been widely adopted. These mechanisms aim to focus on a specific region while generating the corresponding word.

However, the image features in existing attention-based methods are not fine-grained. Xu et al. [2] proposed a top-down attention mechanism that represents the image with the parameters from

the convolutional layer of the CNN, allowing the model to preserve the spatial information and dynamically attend to different regions when generating words.

However, in the top-down attention mechanism, the attention regions correspond to a uniform grid of receptive fields. Since the sizes and shapes of these receptive fields are equal, they are independent of the content of the image. With the advances in object detection, detection-based attention mechanisms [3–5] were proposed to enable the model to attend to the detected salient image regions. Compared with the top-down attention mechanism, the detection-based attention mechanism has a better performance as it can generate variable numbers and sizes of rectangular attention regions. However, it is still not fine-grained enough because there may be other objects or background in the rectangular attention regions.

In this paper, we introduce a novel attention mechanism called panoptic segmentation-based attention, as illustrated in Figure 1. This mechanism comes from the goal of finding more fine-grained attention regions. Naturally, image segmentation, a more fine-grained form of detection, was taken into consideration. We considered it first. Instance segmentation [7,8] is a more challenging task than object detection, as it requires not only detecting all objects in an image but also segmenting the instances of each object class. The results from instance segmentation only contain the main part of the object and do not include the background or overlapped regions. Therefore, we can obtain more fine-grained attention regions with the aid of instance segmentation. However, instance segmentation only segments instances of "things" classes (countable objects with specific shapes; e.g., cars and persons), neglecting "stuff" classes [9] (amorphous background regions; e.g., sky and grass). Losing information of the stuff regions may weaken the model's ability to understand the scene. Recently, Kirillov et al. [10] proposed panoptic segmentation that performs segmentation for all classes (things and stuff) in the image. Inspired by this work, we built our attention mechanism upon this idea.

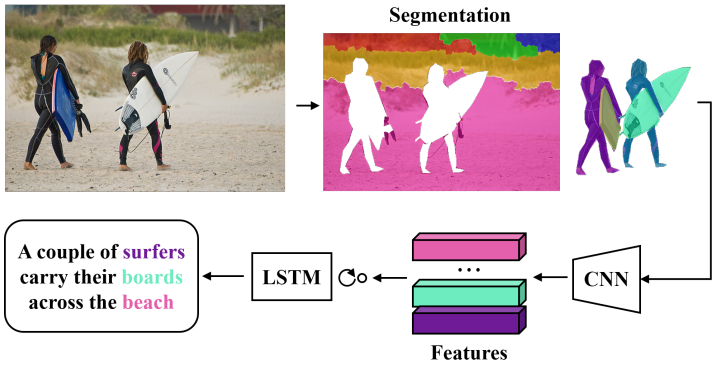

Figure 1. An overview of our proposed method. Given an image, we first generate segmentation regions of the image and use a convolutional neural network to extract the segmentation-region features. For readability, we have applied a color map to each segment. The segmentation-region features are then fed to the LSTM to generate the captions.

Based on the segmentation regions generated from panoptic segmentation, our method extracts image features based on the shape of the segmentation regions and generates captions based on the attention-weighted features. As shown in Figure 2, compared with the detection-based attention mechanisms that contain irrelevant regions around the objects in their attention regions, the attention regions of our approach contain one instance in each region with irrelevant regions masked out, which is more fine-grained and can avoid the negative impact of the background or overlapped regions. Moreover, while the detection-based attention mechanisms only detect things classes and extract scene information in a top-down way [3,6], our approach processes things and stuff classes independently via a dual-attention module. Incorporating the features of stuff regions can provide richer context information and yield better performance.

The main contributions of this paper are:

- Introducing a novel panoptic segmentation-based attention mechanism together with a dual-attention module that can focus on more fine-grained regions at a mask-level when generating captions. To our best knowledge, we are the first to incorporate panoptic segmentation into image captioning.
- We explored and evaluated the impact of combining segmentation features and stuff regions on image captioning. Our study reveals the significance of the fine-grained attention region features and the scene information provided by stuff regions.
- Our proposed method is evaluated on the MSCOCO [11] dataset. Results show that our approach outperforms the baseline detection-based attention approach, improving the CIDEr score from 112.3 to 119.4. Our approach achieves competitive performance against state-of-the-art.

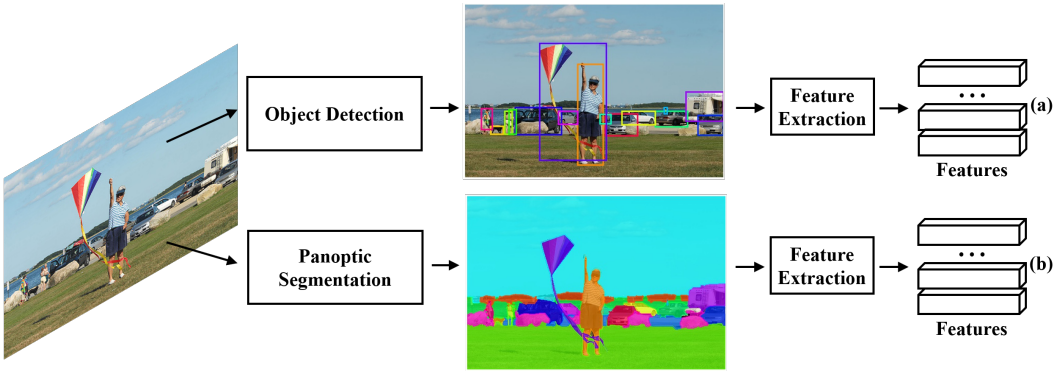

Figure 2. Comparison of the feature extraction procedures between a detection-based attention model and our panoptic segmentation-based attention model: (**a**) The feature extraction procedure of the detection-based attention model. The attention regions are the rectangular regions annotated with colored edges. (**b**) The feature extraction procedure of our panoptic segmentation-based attention model. The attention regions are the fine-grained regions annotated in the colored map. With panoptic segmentation, our method can not only generate fine-grained attention regions, which avoids the negative impact caused by irrelevant regions, but also performs attention to stuff class regions.

2. Related Work

Image Captioning. Image captioning models with the encoder-decoder framework have been widely studied. In recent works [2–4,12–16], attention mechanisms have been introduced to the encoder-decoder framework to obtain a better image representation. Xu et al. [2] first proposed an attention mechanism in image captioning, where weighted CNN features from a convolutional layer are fed to the encoder-decoder framework. Lu et al. [15] developed an adaptive attention mechanism to decide whether to attend to the image or caption at each time step. Leveraging object detection, subsequent works perform attention in different ways. You et al. [13] trained a set of visual concept detectors and performed attention over the detected concepts. Jin et al. [3] generated salient regions of an image by selective search [17] and feed these regions to an attention-based decoder. Similarly to Jin et al. [3], Pedersoli et al. [4] generated object detection proposals and applied a spatial transformer network [18] to them to obtain more accurate regions. Recent detection-based attention methods [6,19,20] use Faster R-CNN [21] to generate detection regions, which significantly increases the quality of the generated captions. However, their attention regions are not fine-grained as they contain other objects or background within their rectangular area.

As object detection models only detect things classes, scene information needs to be provided to the captioning model. With the aid of Latent Dirichlet allocation [22], Fu et al. [23] generated topic vectors from the corpus of captions to represent the scene-specific contexts. Anderson et al. [6] trained an object detector from the Visual Genome [24] dataset which provides annotations of richer categories (e.g., tree, water). The features of the detected regions in an image are then averaged to be regarded as

scene information. Differently from the above methods, our method not only has more fine-grained attention regions but also incorporates stuff regions to obtain richer context information.

Two works that incorporate segmentation into their attention mechanisms are similar to ours. Liu et al. [25] proposed the mask pooling module in video captioning to pool the features according to the shape of the masks. This module is similar to our feature extraction process but it only considers the things classes, and therefore lacks scene information from stuff classes. Zhang et al. [26] proposed the FCN-LSTM network, which incorporates the segmentation information generated from a semantic segmentation model FCN [27] into attention. However, their attention regions are the same as [2], which are not fine-grained, and the segmentation information is merely used to guide the attention. Moreover, their models are incapable of distinguishing instances of the same class in an image, as they use semantic segmentation. As opposed to their methods, our method incorporates panoptic segmentation to distinguish not only things and stuff classes but also their instances with fine-grained attention regions. Our work also places emphasis on demonstrating the advantage of fine-grained segmentation regions over detection regions.

There are other works that focused on other issues in image captioning. Rennie et al. [28] used reinforcement learning to directly optimize the evaluation metric. Dai et al. [29] studied the impact of an RNN with 2D hidden states. Some recent works [30,31] explored the use of graph convolutional networks (GCN) to encode images to improve a visual relationship. Other works attempt to increase the diversity of the captions [32,33].

Instance and Semantic Segmentation. Instance segmentation [7,8,34,35] and semantic segmentation [27,36–38] are two similar tasks but usually employ different approaches. The aim of instance segmentation is to detect and segment each object instance in an image. Most of the instance segmentation approaches [7,8,34] modify the object detection networks to output a ranked list of segments instead of bounding boxes. Hence, instance segmentation can distinguish individual object instances but only for things classes. Semantic segmentation aims to assign a class label to each pixel in an image. It is capable of segmenting stuff and things classes but does not distinguish the individual instances.

Due to the inherent difference mentioned above, although both semantic and instance segmentation techniques aim to segment an image, they had not been unified hitherto. Recently, Kirillov et al. [10] proposed "panoptic segmentation", which unifies the above tasks and requires jointly segmenting things and stuff at the instance level. Our method was developed upon panoptic segmentation and performs attention over the segmentation regions.

3. Captioning Models

In this section, we first describe the generic encoder-decoder image captioning framework (Section 3.1). Then, we describe the up-down attention model in Section 3.2. Our panoptic segmentation-based attention mechanism is based on the up-down attention model with features from segmentation regions. We also proposed a baseline detection-based attention model with features from detection regions as a comparison. Then, we introduce the dual-attention module we proposed for panoptic segmentation features in Section 3.3.

3.1. Encoder-Decoder Framework

First, we briefly introduce the encoder-decoder framework [39]. This framework takes an image $\mathbf{I}$ as input and generates a sequence of words $\mathbf{w} = \{w_0, ..., w_t\}$.

In this framework, captions are generated by LSTM. At a high level, the hidden state of the LSTM is modeled as:

$$h_t = \text{LSTM}(x_t, h_{t-1}), \tag{1}$$

where x_t is the input vector and h_{t-1} is the previous hidden state. For notational convenience, we do not show the propagation of the memory cell.

At each time step t, the probability distribution of the output word is given by:

$$p_\theta(w_t|w_{0:t-1}, \mathbf{I}) = \text{softmax}(W_h h_t). \tag{2}$$

Here we omit the bias term. $W_h \in \mathbb{R}^{\Sigma \times d}$ where Σ is the size of the vocabulary and d is the dimension of the hidden state. θ denotes the parameters of the model.

Given the target ground truth sentence $\mathbf{w}^* = \{w_0^*, ..., w_t^*\}$, the encoder-decoder framework is trained to maximize the probability of $\mathbf{w}^*$. By applying the chain rule to model the joint probability over $w_0^*, ..., w_t^*$, the objective is to minimize the sum of the negative log likelihood:

$$L(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t^*|w_{0:t-1}^*, \mathbf{I})), \tag{3}$$

where T is the total length of the caption.

In the vanilla encoder-decoder framework, the image is only input once, at $t = 0$, to inform the LSTM about the image contents. The input x_t is the previous generated word, given by:

$$x_t = \begin{cases} W_c \text{CNN}(\mathbf{I}) & \text{if } t = 0 \\ E w_{t-1} & \text{if } t \geq 1, \end{cases} \tag{4}$$

where $W_c \in \mathbb{R}^{d \times D}$, $E \in \mathbb{R}^{d \times \Sigma}$, D is the dimension of the image features and E is the word embedding matrix. The beginning of the sentence w_0 and end of the sentence w_t are marked with a BOS token and an EOS token, respectively.

3.2. Up-Down Attention Model

We adopt the framework of the up-down attention model [6] with our segmentation-region/detection-region features. This model is composed of an attention LSTM which generates the attention weights and a language LSTM which generates words. Their hidden states are denoted by h_t^1 and h_t^2, respectively.

Given k image regions, the features of these regions $\mathbf{v}$ are given by:

$$\mathbf{v} = \{v_1, ..., v_k\}, v_i \in \mathbb{R}^D. \tag{5}$$

The input to the attention LSTM is the concatenation of the mean-pooled image feature, the previous hidden state of the language LSTM, and the previous generated word:

$$x_t^1 = [\bar{I}, h_{t-1}^2, E w_{t-1}] \tag{6}$$

where the mean-pooled image feature $\bar{I}$ provides the attention LSTM with a global content of the image. Note that, differently from [6] where $\bar{I} = \frac{1}{k}\sum_i^k v_i$ is the average feature of the detected regions, in this paper, $\bar{I}$ is the average feature of the uniform grid of the image regions. In the baseline detection-based attention model, the image features $\mathbf{v}$ are the features of detection regions. In order to obtain fine-grained representation of the images, the image features $\mathbf{v}$ in our method are the features of segmentation regions. The definitions of detection/segmentation regions are illustrated in Sections 4.1 and 4.2, and the image features $\mathbf{v}$ are described in Section 5.

The input to the language LSTM is the concatenation of the attention weighted image feature and the previous hidden state of the attention LSTM:

$$x_t^2 = [I_t, h_t^1] \tag{7}$$

where the attention weighted image feature I_t is the weighted sum of v_i:

$$a_t^i = W_a^T \tanh(W_{av} v_i + W_{ah} h_t^1) \tag{8}$$

$$\boldsymbol{\alpha}_t = \text{softmax}\,(\boldsymbol{a_t}) \tag{9}$$

$$I_t = \sum_{i=1}^{k} \alpha_t^i v_i \tag{10}$$

where α_t^i is the normalized attention weights, $W_a \in \mathbb{R}^A$, $W_{av} \in \mathbb{R}^{A \times D}$, and $W_{ah} \in \mathbb{R}^{A \times d}$; A is the dimension of the attention layer. For the sake of simplicity, we denote $\{a_t^1, ..., a_t^k\}$ by $\boldsymbol{a_t}$ and denote $\{\alpha_t^1, ..., \alpha_t^k\}$ by $\boldsymbol{\alpha}_t$.

Then, the hidden state of the language LSTM is used to generate the distribution of the next word following (2) and h_t is replaced with h_t^2 following [6]. The other parts of the model remain the same with the definition in Section 3.1.

Reinforcement training [28] is also introduced to directly optimize the CIDEr [40] metric. For a sampled sentence $\mathbf{w}$, a reward function $r(\mathbf{w})$ denoting the CIDEr score of $\mathbf{w}$ is used to measure the quality the sentence. With this reward, the probability of the sampled captions with a higher CIDEr score is increased by reinforcement training following [28]. Therefore, we can directly optimize the CIDEr score by reinforcement learning.

3.3. Dual-Attention Module for Panoptic Segmentation Features

In panoptic segmentation, the segmentation regions include things and stuff classes. As they convey different kinds of information, they have to be processed separately.

In this section, we describe the Dual-Attention Module we propose for panoptic segmentation features. Our framework and the Dual-Attention Module is shown in Figure 3. In this case, the image features $\mathbf{v}$ in (5) include the features of things classes $\mathbf{v}^t$ and the features of stuff classes $\mathbf{v}^s$; i.e., $\mathbf{v} = [\mathbf{v}^t, \mathbf{v}^s]$, where $\mathbf{v}^t$ and $\mathbf{v}^s$ are given by:

$$\mathbf{v}^t = \{v_1^t, ..., v_{k^t}^t\}, v_i^t \in \mathbb{R}^D \tag{11}$$

$$\mathbf{v}^s = \{v_1^s, ..., v_{k^s}^s\}, v_i^s \in \mathbb{R}^D, \tag{12}$$

where k^t is the number of things regions and k^s is the number of stuff regions. The attention process is further given by:

$$a_t^{ti} = {W_a^t}^T \tanh(W_{av}^t v_{k^t}^t + W_{ah}^t h_t^1) \tag{13}$$

$$a_t^{si} = {W_a^s}^T \tanh(W_{av}^s v_{k^s}^s + W_{ah}^s h_t^1) \tag{14}$$

$$\boldsymbol{\alpha}_t^t = \text{softmax}\,(\boldsymbol{a_t^t}) \tag{15}$$

$$\boldsymbol{\alpha}_t^s = \text{softmax}\,(\boldsymbol{a_t^s}) \tag{16}$$

$$I_t^t = \sum_{i=1}^{k^t} \alpha_t^{ti} v_i^t \tag{17}$$

$$I_t^s = \sum_{i=1}^{k^s} \alpha_t^{si} v_i^s \tag{18}$$

$$I_t = [I_t^t, I_t^s]. \tag{19}$$

The final image feature I_t is the concatenation of the attention weighted features I_t^s and I_t^t. Other parts of the model remain the same.

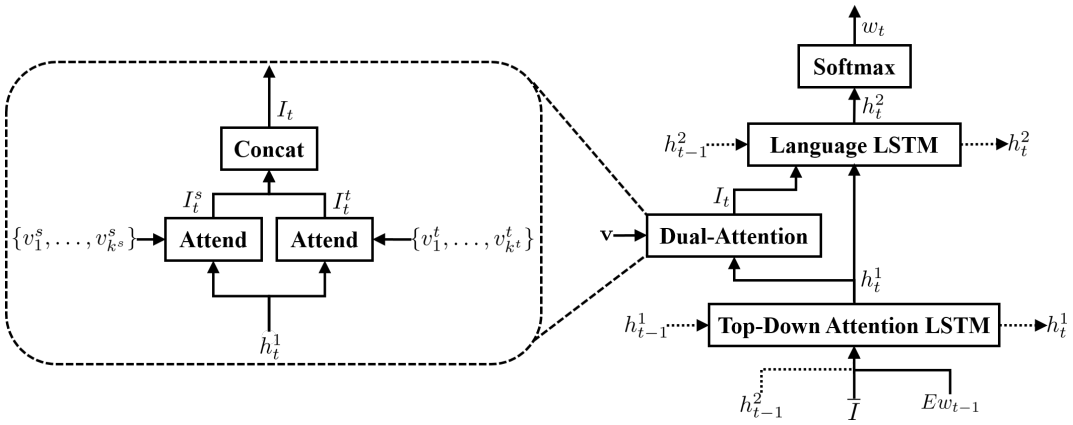

Figure 3. Overview of our framework and the Dual-Attention Module used for handling the panoptic segmentation features. The features of the things and stuff classes are fed into the pathway to perform attention individually. The attended features are then concatenated and fed into the LSTM to generate captions.

4. Attention Regions

In this section, we describe how to obtain the detection and segmentation regions for later feature extraction.

4.1. Detection Regions

We first describe how to obtain the detection regions for the detection-based attention model. Given an image, the output of an object detection model is a set of bounding boxes (i.e., the rectangular boxes that contain the instances):

$$\mathbf{b} = \text{Det}(\mathbf{I}), \tag{20}$$

where $\mathbf{b} = \{b_1, ..., b_L\}; b_i = (x_{min}, y_{min}, x_{max}, y_{max})$ contains the coordinates of the bounding box; L is the number of the instances in the image. Here we do not show the output category prediction. As current object detection models only detect things classes, $\mathbf{b}$ does not contain regions that belong to stuff classes.

4.2. Segmentation Regions

Next, we describe how to obtain the segmentation regions for the segmentation-based attention model. Given an image, a segmentation model generates a set of binary masks for each instance in the image:

$$\mathbf{m} = \text{Seg}(\mathbf{I}) \tag{21}$$

where $\mathbf{m} = \{m_1, ..., m_L\}, m_i \in \{0, 1\}^{H \times W}$ is the mask indicating which pixels belong to the instance. H and W are the height and width of the image, respectively. Here we also omit the output category prediction. Note that in panoptic segmentation $\mathbf{m}$ contains things and stuff classes while in instance segmentation $\mathbf{m}$ only contains things classes.

5. Feature Extraction

In this section, we introduce how we extract the features $\mathbf{v}$ of the segmentation/detection regions that are used in the image captioning model.

For segmentation-region features, we extract the image features by convolutional feature masking [41], denoted as the CFM approach. As shown in Figure 4, given an image $\mathbf{I}$, we first

obtain the masks **m** of the image as in (21). Meanwhile, the feature maps CNN(**I**) of the image are extracted from the convolutional layer of a pre-trained CNN. The masks are resized to match the size of the feature maps. The final segmentation-region features are given by:

$$v_i = \text{CNN}(\mathbf{I}) \odot \text{resize}(m_i) \tag{22}$$

where we perform an element-wise product $\odot$ to every channel of the output feature maps.

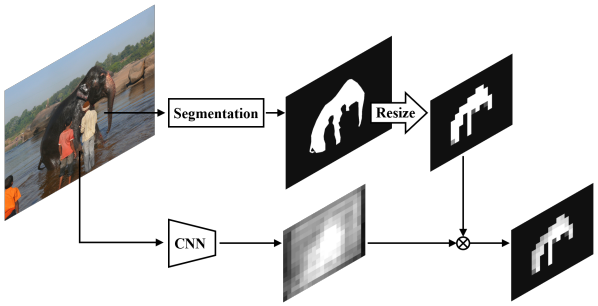

Figure 4. An overview of the convolutional feature masking (CFM) approach for feature extraction. The image is fed to the segmentation model to obtain masks and fed to the CNN to obtain feature maps, respectively. The masks are resized to have the same size as the feature maps. The feature maps are then multiplied by the resized mask to obtain the segmentation-region features.

For detection-region features, we scale the coordinates to match the sizes of the feature maps. The final detection-region features are given by:

$$v_i = \text{crop}(\text{CNN}(\mathbf{I}), \text{resize}(b_i)) \tag{23}$$

where the crop() operation crops the output feature maps based on the resized coordinates. Similarly, this operation is performed to every channel of the output feature maps. Note that unlike [41], in order to obtain richer information, the pixel values of the resized masks are obtained by averaging without thresholding.

6. Experimental Results

6.1. Dataset

Extensive experiments were performed to evaluate our proposed method. All the results were based on the MSCOCO dataset. For validation of offline testing, the "Karpathy" split [42] that has been widely used in prior work was adopted. The training, validation, and test sets respectively, contained 113,287, 5000, and 5000 images, along with five captions per image. We truncated captions longer than 16 words and removed the words that appeared less than five times, resulting in 9587 words.

The COCO-Stuff [9] dataset contains 80 things classes, 91 stuff classes, and one unlabeled class. The stuff classes are organized in a hierarchical way and belong to 15 parent categories. We omitted the unlabeled class. Since the stuff regions are often scattered, and we did not need a fine classification of the sub-classes stuff in image captioning; we used a compact representation for stuff. The regions of the sub-classes that belong to the same parent category were merged by adding the masks of these sub-classes together, resulting in 15 parent stuff categories.

6.2. Implementation Details

In our experiments, in the up-down attention model, the dimension of the hidden states in the language LSTM and the attention LSTM and word embedding were set to 1000. The hidden state A of the attention layer was 512. We used the Adam [43] optimizer with initial learning rate of 5×10^{-4}. The weight-decay and momentum were 1×10^{-4} and 0.9, respectively. We set the batch size to 100

and trained the models for up to 50 epochs. In order to further boost performance, we trained the models with reinforcement learning for another 50 epochs.

We used a pre-trained ResNet-101 [44] as our CNN model to extract image features. The image feature v_i was the mean output of the last convolutional layer of ResNet-101, and thus had a dimension of 2048.

As there is no current available model to jointly perform both elements of the panoptic segmentation, we used Mask R-CNN [7] to perform instance segmentation for things classes and DeepLab [36] to perform semantic segmentation for stuff classes. Their outputs were merged to represent the result of panoptic segmentation. The minimum detection confidence of Mask R-CNN was 0.6 and the non-maximum suppression threshold was 0.5. The code of our method can be accessed at https://github.com/jamiechoi1995/PanoSegAtt.

We denote the model that uses panoptic segmentation features by PanopticSegAtt, and denote the baseline model that uses detection features by DetectionAtt. To evaluate the impact of stuff regions, we also propose InstanceSegAtt, a model that only uses instance segmentation features (i.e., without stuff regions) as another baseline.

6.3. Evaluation

In this subsection, we first compare our results with state-of-the-art models on MSCOCO dataset. To demonstrate the effect of the segmentation-region features, we conducted qualitative and quantitative analyses of the difference between InstanceSegAtt and DetectionAtt. Moreover, to demonstrate the effect of the features of stuff regions, we also conducted qualitative and quantitative analyses of the difference between PanopticSegAtt and InstanceSegAtt. We report results using the COCO captioning evaluation tool [11], which reports the following metrics: BLEU [45], METEOR [46], ROUGE-L [47], and CIDEr [40]. Table 1 shows the overall results on the MSCOCO dataset.

Compared with the method of Zhang et al. [26] which is most similar to our method, our PanopticSegAtt model surpasses their method in all metrics by a large margin. We consider that it is because of the more fine-grained attention regions and the combination of panoptic segmentation in our method. We also compared our PanopticSegAtt with PanopticSegAtt (w/o Dual-Attend). The full model improved the CIDEr from 118.2 to 119.4, which shows that with the dual-attention module, our model can generate more accurate captions.

We then compared our method with the typical attention methods [2–4,13–15]. For example, SCA-CNN [14] uses spatial and channel-wise attention in the CNN. Lu et al. [15] adaptively attends to the image and caption during decoding. Our PanopticSegAtt model significantly outperforms these methods in all metrics, which demonstrates the power of our panoptic segmentation-based attention mechanism. We also compared our method with state-of-the-art methods [48–52]. Our PanopticSegAtt model outperforms these methods in most of the metrics, especially on the CIDEr metric, which is considered to be the metric most aligned with human judgments. Note that Stack-Cap [53] has higher scores, as the model of this method has three LSTMs to perform coarse-to-fine decoding, which is more complex than our method. Since Anderson [6] used the extra Visual Genome [24] dataset to train the object detector, their attention regions are much richer than ours. Thus, we did not compare with these two methods directly.

Figure 5 shows the statistical results of the CIDEr scores of the captions in the Karpathy test split for DetectionAtt, InstanceSegAtt, and PanopticSegAtt. Captions within the interval with a score from 0 to 1 are not accurate enough to describe the images. Among them, the number of DetectionAtt is the highest, InstanceSegAtt is second, and PanopticSegAtt is the smallest. In the interval with a score from 1 to 3, captions can accurately describe the images. Among them, the number of PanopticSegAtt is the highest, InstanceSegAtt is second, and DetectionAtt is the smallest. In the interval where the score is greater than 3, the number of captions of the three methods is almost the same. This is because the images in this interval are simple so that all the models work well for them. The above results indicate the advantages of using the segmentation features and the stuff regions in our method.

Table 1. The results obtained on the MSCOCO Karpathy test split [42]. † indicates ensemble models. Higher is better in all columns. Scores were multiplied by a factor of 100.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
Hard-Attention [2]	71.8	50.4	35.7	25.0	-	23.0	-
VIS-SAS [48]	72.5	52.6	38.1	28.1	55.4	23.7	82.1
Jin et al. [3]	69.7	51.9	38.1	28.2	50.9	23.5	83.8
ATT-FCN † [13]	70.9	53.7	40.2	30.4	-	24.3	-
Zhang et al. [26]	71.2	51.4	36.8	26.5	-	24.7	88.2
Areas of Attention [4]	-	-	-	30.7	-	24.5	93.8
SCA-CNN [14]	71.9	54.8	41.0	31.1	53.1	25.0	95.2
Aneja et al. [49]	72.2	55.3	41.8	31.6	53.1	25.0	95.2
Fu et al. † [23]	72.4	55.5	41.8	31.3	53.2	24.8	95.5
Lu et al. [54]	-	-	-	33.1	53.9	25.8	99.3
Chen et al. [50]	74.0	57.6	44.0	33.5	54.6	26.1	103.4
Jiang et al. [51]	74.3	57.9	44.2	33.6	54.8	26.1	103.9
Adaptive [15]	74.2	58.0	43.9	33.2	-	26.6	108.5
Att2all [28]	-	-	-	34.2	55.7	26.7	114.0
Dognin et al. [52]	-	-	-	-	-	26.9	116.1
Stack-Cap (C2F) [53]	**78.6**	**62.5**	**47.9**	**36.1**	**56.9**	**27.4**	**120.4**
DetectionAtt	77.1	60.4	45.4	33.7	55.6	26.3	112.3
InstanceSegAtt	77.9	61.4	46.7	34.9	56.3	27.0	117.3
PanopticSegAtt (w/o Dual-Attend)	**78.2**	**61.8**	46.9	34.9	56.4	26.9	118.2
PanopticSegAtt	78.1	61.7	**47.1**	**35.3**	**56.6**	**27.3**	**119.4**

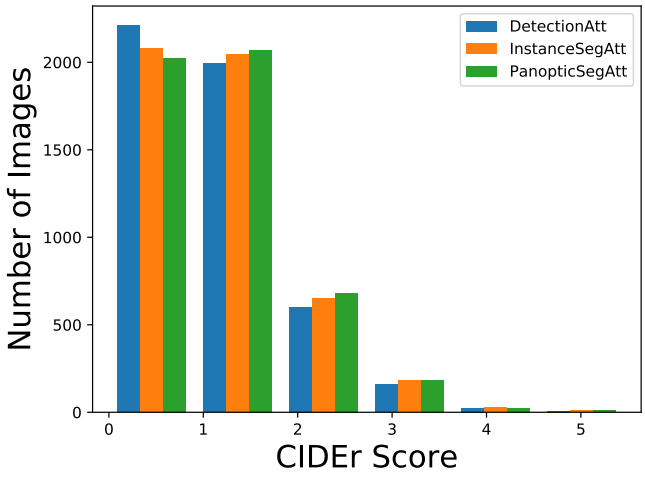

Figure 5. Histogram of CIDEr scores of DetectionAtt, InstanceSegAtt, and PanopticSegAtt.

We also evaluated our model on the online COCO test server in Table 2. Our PanopticSegAtt model achieves comparable scores compared to the state-of-the-art models.

Table 2. The results obtained on the online MSCOCO test server. † indicates ensemble models. Higher is better in all columns. Scores are multiplied by a factor of 100.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
sgLSTM [55]	67.9	49.9	36.5	26.9	49.9	23.5	82.1
Aneja [49]	71.5	54.5	40.8	30.4	52.5	24.6	91.0
Adaptive [15]	74.8	58.4	44.4	33.6	55.2	26.4	104.2
Att2all † [28]	-	-	-	35.2	56.3	27.0	114.7
Stack-Cap (C2F) [53]	77.8	61.6	46.8	34.9	56.2	27.0	114.8
PanopticSegAtt	**79.0**	**63.0**	**48.3**	**36.5**	**57.3**	**27.7**	**117.8**

6.3.1. Segmentation-Region Features versus Detection-Region Features

We used Mask R-CNN [7] to perform instance segmentation and object detection, which resulted in equal numbers of attention regions in both tasks. Thus, the difference between segmentation regions and detection regions is that the segmentation regions are more fine-grained and better-matching to the shape of the instances. We compared InstanceSegAtt with DetectionAtt to demonstrate the impact of segmentation-region features.

As shown in Table 1, comparing against DetectionAtt verifies the effectiveness of using segmentation-region features. Our InstanceSegAtt model improves the CIDEr score from 112.3 to 117.3 compared with the DetectionAtt model. The performance gap between InstanceSegAtt and DetectionAtt demonstrates that using features from more fine-grained regions is beneficial.

The training curves in terms of CIDEr metric are shown in Figure 6. During 50 epochs' training, the InstanceSegAtt model consistently surpasses the DetectionAtt model. This result indicates that, since segmentation regions do not include irrelevant regions that have negative impact on captioning models during training, using segmentation-region features leads to better convergence and performance.

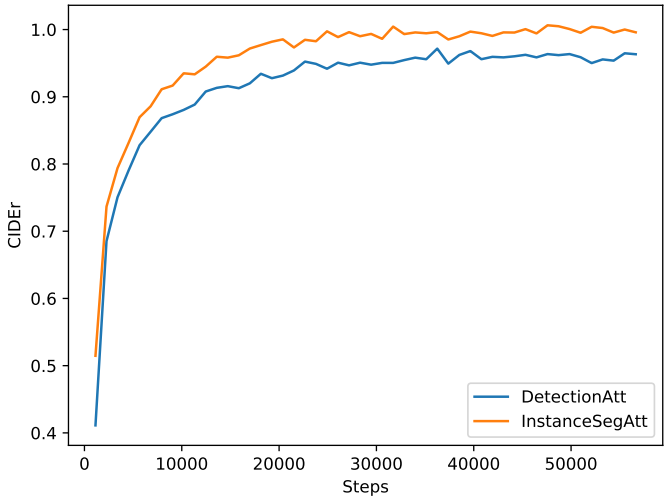

Figure 6. Comparison of CIDEr scores on the validation set for InstanceSegAtt and DetectionAtt during 50 epochs' training.

Dense Annotation Split. In order to better evaluate the effect of using features of fine-grained segmentation regions, we present a new split from the COCO dataset called dense annotation split. This is based on the intuition that the segmentation-based attention model ought to distinguish instances even if they are overlapped, while the detection-based attention model may be confused by the features of the rectangular regions which include irrelevant regions. The dense annotation split consists of images in which some of the instances are highly overlapped. We generated this split by selecting images for which the IoU (intersection-over-union) between any of their two instances was over 0.5, resulting in 12,967, 570, and 608 images in the training, validation, and testing sets of Karpathy split [42], respectively. We then evaluated the performance of InstanceSegAtt and DetectionAtt on this split.

As shown in Figure 7, compared with the DetectionAtt model that uses detection-region features, InstanceSegAtt is better at handling images with overlapped objects, as the segmentation-region features do not overlap with each other. For example, in the first column of Figure 7, the detection region of the person and snowboard are highly overlapped. InstanceSegAtt correctly generates the word "snowboard", while DetectionAtt cannot. Similarly, in the second column of Figure 7, the dense detection regions make DetectionAtt hard to generate "a woman" like InstanceSegAtt. In the third column of Figure 7, the detection regions of the two giraffes contain each other, which could confuse the DetectionAtt model. In the last column of Figure 7, the rectangle detected region of the boy in the second row contains the region of the man; therefore, DetectionAtt cannot correctly recognize their relationship.

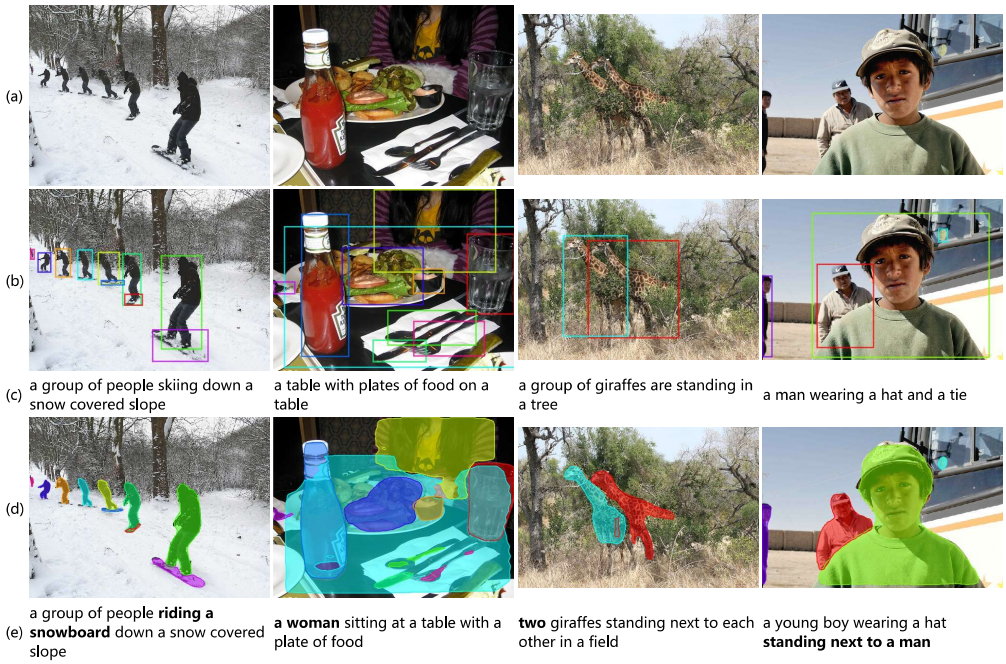

Figure 7. Examples of captions generated and the attention regions of DetectionAtt and InstanceSegAtt. (**a**): Original image; (**b**) detection regions generated by the object detection model; (**c**) captions generated by DetectionAtt; (**d**) instance segmentation regions generated by the instance segmentation model; (**e**) captions generated by InstanceSegAtt. Images are selected from the dense annotation split. Bold text indicates where InstanceSegAtt has included more detail in the captions compared to DetectionAtt. Results of PanopticSegAtt are not shown, as the differences between the captions of PanopticSegAtt and InstanceSegAtt are not obvious in these images.

Table 3 shows the evaluation results on the dense annotation test split. The performance gap between InstanceSegAtt and DetectionAtt is larger than their gap in Table 1, which demonstrates that the segmentation-based attention model has the advantage in handling densely annotated images. We consider that the features of overlapped regions make it hard for DetectionAtt to distinguish the individual instances in overlapped regions. Thus, DetectionAtt has lower scores compared with InstanceSegAtt. It is also observable that the performance of PanoSegAtt is better than that of InstanceSegAtt. This suggests that the contextual information from stuff region features is of benefit for the model to recognize the partially occluded objects. Such a conclusion is widely acknowledged in object detection [56–58].

Table 3. The results obtained on the MSCOCO dense annotation test split. Higher is better in all columns. Scores are multiplied by a factor of 100.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
DetectionAtt	76.9	61.0	45.5	35.7	57.0	26.6	112.2
InstanceSegAtt	78.4	62.6	48.3	37.0	57.5	27.2	117.8
PanopticSegAtt	**79.2**	**63.8**	**49.5**	**38.1**	**58.2**	**27.6**	**120.9**

The above results demonstrate that, with fine-grained attention regions, the model can not only avoid the negative impact from irrelevant regions but also benefit from the context information and is more capable of distinguishing instances in images with overlapped objects.

6.3.2. With Stuff versus without Stuff

Stuff regions play an important role in image captioning, as they provide the context information (scene, location, etc.) to the model. As shown in Table 1, when comparing PanopticSegAtt with InstanceSegAtt, PanopticSegAtt further improves the CIDEr score by 2.1, which shows that using features of stuff regions also enhances performance. To qualitatively demonstrate the superiority of using features of stuff regions, we show the example captions generated by InstanceSegAtt and PanopticSegAtt in Figure 8. Compared with the InstanceSegAtt model that does not have the features from stuff regions, PanopticSegAtt can generate captions with richer scene information. For example, in the first column of Figure 8, the segmentation regions generated by panoptic segmentation contain the brick wall region in the image and provide the feature of brick wall to the captioning model. Thus, the caption generated by PanopticSegAtt contains the background information "brick wall". Similarly, in the second column of Figure 8, the purple area provides the context information of the image, so the PanopticSegAtt can generate the phrase "with plants". In the third column of Figure 8, while InstanceSegAtt does not generate the scene of the photo, PanopticSegAtt correctly generates the scene phrase "in a field". In the fourth column of Figure 8, the purple area provides the location of the train to the PanopticSegAtt model which generates a more accurate scene word "mountain" while InstanceSegAtt generates the scene word "field".

The above results demonstrate that, with the aid of stuff regions, our panoptic segmentation-based attention method can generate captions with richer context information.

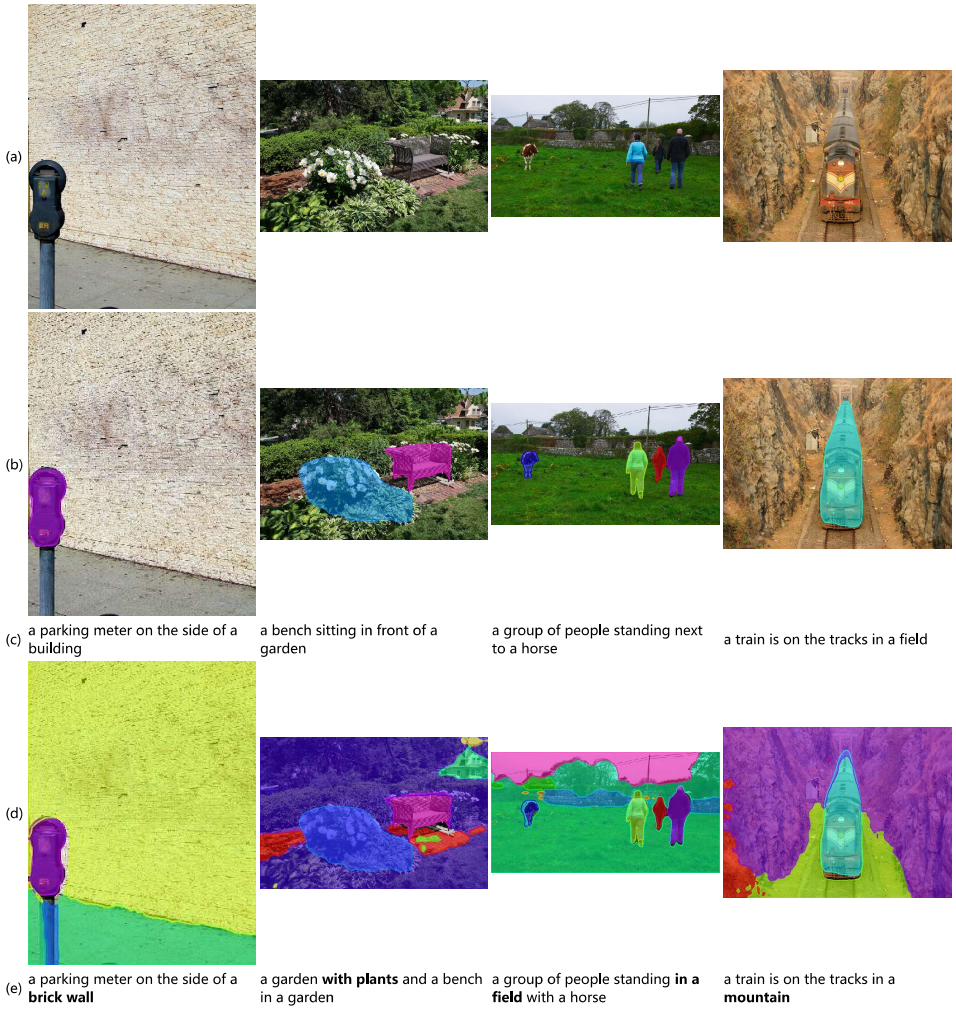

Figure 8. Examples of captions and the attention regions of InstanceSegAtt and PanopticSegAtt. (**a**) Original image; (**b**) instance segmentation regions generated by the instance segmentation model; (**c**) captions generated by InstanceSegAtt; (**d**) segmentation regions generated by the panoptic segmentation model; (**e**) captions generated by PanopticSegAtt. Images were selected from the Karpathy test split. Bold text indicates where PanopticSegAtt has included more detail in the captions compared to InstanceSegAtt.

7. Conclusions

In this paper, we present a novel panoptic segmentation-based attention mechanism for image captioning, which provides more fine-grained regions for attention with the aid of panoptic segmentation. Our method achieves competitive performance against state-of-the-art methods. Qualitative and quantitative evaluation results show that our approach has better scene and instance recognition of an image compared with the detection-based attention method, which demonstrates the superiority of using features of fine-grained segmentation regions in image captioning. Our research provides a novel perspective for academics and practices on how to improve the performance of image captioning. The above results indicate that extracting fine-grained image features is a prospective research topic for future work.

We plan to further utilize the available model in the panoptic segmentation task to generate segmentation regions for things and stuff in a more elegant way. Investigating different ways to incorporate stuff-region features into the captioning model is also a future research direction.

Author Contributions: Conceptualization, W.C.; methodology, W.C.; software, W.C. and Z.X.; validation, W.C.; formal analysis, W.C.; investigation, W.C.; resources, W.C.; data curation, W.C.; writing—original draft preparation, W.C.; writing—review and editing, W.C., Z.X. X.S., P.L.R., L.J., and X.P.; visualization, W.C.; supervision, L.J.; project administration, L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Pearl River S and T Nova Program of Guangzhou: 201710010020, Fundamental Research Funds for the Central Universities: 2019MS086, and National Science Foundation of China: 61300135.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
2. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
3. Jin, J.; Fu, K.; Cui, R.; Sha, F.; Zhang, C. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv* **2015**, arXiv:1506.06272.
4. Pedersoli, M.; Lucas, T.; Schmid, C.; Verbeek, J. Areas of attention for image captioning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1242–1250.
5. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
6. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; Volume 3, p. 6.
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
8. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3150–3158.
9. Caesar, H.; Uijlings, J.; Ferrari, V. COCO-Stuff: Thing and stuff classes in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1209–1218.
10. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P. Panoptic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9404–9413.
11. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
12. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W.W.; Salakhutdinov, R.R. Review networks for caption generation. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2361–2369.
13. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4651–4659.
14. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6298–6306.
15. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 6, p. 2.

16. Wang, W.; Chen, Z.; Hu, H. Hierarchical attention network for image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8957–8964.
17. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
18. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems—NIPS'15, Montreal, QC, Canada, 7–12 December 2015; Volume 2; pp. 2017–2025.
19. Yang, Z.; Zhang, Y.J.; ur Rehman, S.; Huang, Y. Image Captioning with Object Detection and Localization. In Proceedings of the International Conference on Image and Graphics, Shanghai, China, 13–15 September 2017; pp. 109–118.
20. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Neural Baby Talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7219–7228.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
22. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
23. Fu, K.; Jin, J.; Cui, R.; Sha, F.; Zhang, C. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2321–2334. [CrossRef] [PubMed]
24. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
25. Liu, A.; Qiu, Y.; Wong, Y.; Su, Y.; Kankanhalli, M. A Fine-Grained Spatial-Temporal Attention Model for Video Captioning. *IEEE Access* **2018**, *6*, 68463–68471, doi:10.1109/ACCESS.2018.2879642. [CrossRef]
26. Zhang, Z.; Wu, Q.; Wang, Y.; Chen, F. Fine-Grained and Semantic-Guided Visual Attention for Image Captioning. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1709–1717.
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
28. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
29. Dai, B.; Ye, D.; Lin, D. Rethinking the form of latent states in image captioning. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 282–298.
30. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10685–10694.
31. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring visual relationship for image captioning. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September, 2018; pp. 684–699.
32. Liu, X.; Li, H.; Shao, J.; Chen, D.; Wang, X. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 338–354.
33. Dai, B.; Lin, D. Contrastive learning for image captioning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 898–907.
34. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. *arXiv* **2016**, arXiv:1611.07709.
35. Liu, S.; Jia, J.; Fidler, S.; Urtasun, R. SGN: Sequential grouping networks for instance segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
36. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]

37. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
38. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
39. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
40. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
41. Dai, J.; He, K.; Sun, J. Convolutional feature masking for joint object and stuff segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3992–4000.
42. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
46. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MA, USA, 26–27 June 2014; pp. 376–380.
47. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Association for Computational Linguistics Workshop, Barcelona, Spain, 21–26 July 2004. Available online: https://www.aclweb.org/anthology/W04-1013.pdf (accessed on 20 December 2019).
48. Zhou, L.; Zhang, Y.; Jiang, Y.G.; Zhang, T.; Fan, W. Re-Caption: Saliency-Enhanced Image Captioning Through Two-Phase Learning. *IEEE Trans. Image Process.* **2019**, *29*, 694–709. [CrossRef] [PubMed]
49. Aneja, J.; Deshpande, A.; Schwing, A.G. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5561–5570.
50. Chen, X.; Ma, L.; Jiang, W.; Yao, J.; Liu, W. Regularizing RNNs for Caption Generation by Reconstructing the Past With the Present. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
51. Jiang, W.; Ma, L.; Chen, X.; Zhang, H.; Liu, W. Learning to Guide Decoding for Image Captioning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
52. Dognin, P.; Melnyk, I.; Mroueh, Y.; Ross, J.; Sercu, T. Adversarial Semantic Alignment for Improved Image Captions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10463–10471.
53. Gu, J.; Cai, J.; Wang, G.; Chen, T. Stack-Captioning: Coarse-to-Fine Learning for Image Captioning. In Proceedings of theThirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
54. Lu, S.; Hu, R.; Liu, J.; Guo, L.; Zheng, F. Structure Preserving Convolutional Attention for Image Captioning. *Appl. Sci.* **2019**, *9*, 2888. [CrossRef]
55. Xian, Y.; Tian, Y. Self-Guiding Multimodal LSTM-when we do not have a perfect training dataset for image captioning. *IEEE Trans. Image Process.* **2019**, *28*, 5241–5252. [CrossRef] [PubMed]
56. Borji, A.; Iranmanesh, S.M. Empirical Upper-bound in Object Detection and More. *arXiv* **2019**, arXiv:1911.12451.

57. Divvala, S.K.; Hoiem, D.; Hays, J.H.; Efros, A.A.; Hebert, M. An empirical study of context in object detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1271–1278.
58. Heitz, G.; Koller, D. Learning spatial context: Using stuff to find things. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 30–43.