$\mathcal{S}^2$ Transformer for Image Captioning

Pengpeng Zeng*, **Haonan Zhang***, **Jingkuan Song**†, **Lianli Gao**

School of Computer Science and Engineering and Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Chengdu, China.

{is.pengpengzeng, zchiowal, jingkuan.song}@gmail.com and {lianli.gao}@uestc.edu.cn

Abstract

Transformer-based architectures with grid features represent the state-of-the-art in visual and language reasoning tasks, such as visual question answering and image-text matching. However, directly applying them to image captioning may result in spatial and fine-grained semantic information loss. Their applicability to image captioning is still largely under-explored. Towards this goal, we propose a simple yet effective method, Spatial- and Scale-aware Transformer ($\mathcal{S}^2$ Transformer) for image captioning. Specifically, we firstly propose a Spatial-aware Pseudo-supervised (SP) module, which resorts to feature clustering to help preserve spatial information for grid features. Next, to maintain the model size and produce superior results, we build a simple weighted residual connection, named Scale-wise Reinforcement (SR) module, to simultaneously explore both low- and high-level encoded features with rich semantics. Extensive experiments on the MSCOCO benchmark demonstrate that our method achieves new state-of-art performance without bringing excessive parameters compared with the vanilla transformer. The source code is available at https://github.com/zchoi/S2-Transformer.

1 Introduction

As a fundamental task of visual and language reasoning, image captioning, which automatically generates a natural language description for an image, has attracted extensive attention [Vinyals *et al.*, 2016; Cornia *et al.*, 2019; Wang *et al.*, 2020; Chen *et al.*, 2021]. Originally inspired by neural machine translation [Sutskever *et al.*, 2014], its general paradigm is: firstly encoding an image to extract visual features, and then feeding those features into an encoder-decoder framework to generate descriptions [Xu *et al.*, 2015]. Due to its specific proprieties, such as rich visual information and sophisticated semantics of descriptions, it remain a challenging problem.

For visual feature extracting, two types of features are widely adopted: region and grid features, as shown in Fig. 1a (i) and (ii), respectively. The region features are designed to explore object instances, which strongly correlate with nouns in textual descriptions (*e.g.,* "giraffe", "grass" and "tree"). To detect explicit object boxes and output region features, existing off-the-shelf methods such as Faster-RCNN [Ren *et al.*, 2015] are pre-trained on VG dataset [Krishna *et al.*, 2017], which is computationally expensive and not flexible. Beyond that, the detected regions may lack contextual information (*e.g.,* "stands on" and "in the forest") and fine-grained details (*e.g.,* "eat leaves"). By contrast, the grid features are designed to extract all patch information to cover the whole image. Previous studies [Jiang *et al.*, 2020; Zhang *et al.*, 2021] revise the advantage of grid features and find them to perform better than region features both in terms of performance and time-cost. However, directly operating at grid features in a flatting manner unavoidably disrupts the spatial association between grids. One natural solution is to combine the above two visual features as visual inputs, but it suffers from computation costs and complex fusion procedures.

Furthermore, transformer-based models are applied as the encoder-decoder for high quality image captioning [Li *et al.*, 2019; Pan *et al.*, 2020; Zhang *et al.*, 2021] due to its strong modeling capabilities and excellent performance, shown as Fig. 1b (i). Most of them are focused on modifying the attention block. For example, [Huang *et al.*, 2019] proposes an "attention on attention" module, which extends self-attention mechanisms to determine the relevance between attention results and queries. [Pan *et al.*, 2020] proposes a X-Linear attention block that fully employs bilinear pooling to capitalize on visual information or perform multi-modal reasoning selectively. [Cornia *et al.*, 2020] proposes a $\mathcal{M}^2$ transformer that designs a memory-augmented attention to encode a priori information and a mesh cross attention (MCA) to take advantage of scale-wise features to fully explore rich visual semantics, shown as Fig. 1b (ii). However, $\mathcal{M}^2$ transformer (w/o memory) based on grid feature has suffered a performance degradation and brought a parameters increase compared with the vanilla transformer, where the results are summarized in Fig. 1b (iv). Thus, how to effectively and efficiently incorporate grid features with transformer-based architecture remains to be explored for image captioning.

*Pengpeng Zeng and Haonan Zhang contribute equally to this paper.

†Corresponding author: Jingkuan Song.

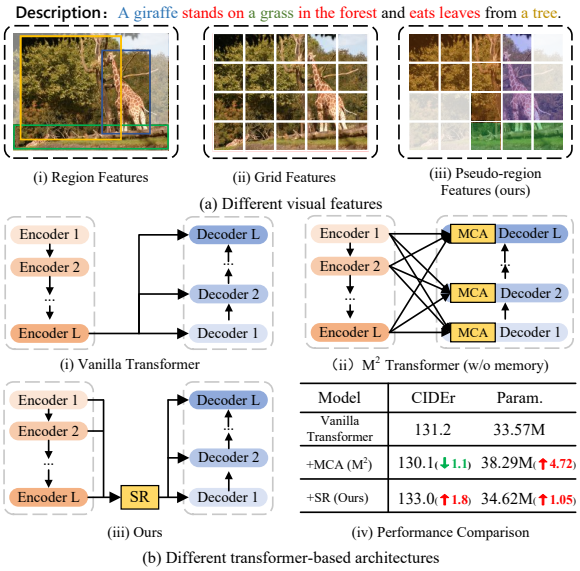

Figure 1: (a) Comparison of different visual features. Based on grid features, our proposed SP module aims to implicitly learn spatial information about grids in a pseudo-supervised manner instead of directly using explicit region features. (b) Comparison of different transformer-based architectures, where all models adopt grid features as visual features. Our SR module simultaneously explores both low- and high-level encoded features to produce superior results while maintaining a relatively small model size.

To address the above problem, this paper proposes a novel Spatial- and Scale-aware Transformer ($\mathcal{S}^2$ Transformer). Specifically, we firstly propose a **Spatial-aware Pseudo-supervised** (SP) module that aims to solve the spatial information loss of grid features caused by the flattening operation. In practice, we utilize a number of learnable semantic clusters to quantize grid features into semantic clusters, which implicitly represent discriminative regions. Furthermore, to maintain the model size and produce superior performance, we propose a simple weighted residual connection, named **Scale-wise Reinforcement** (SR), module to simultaneously explore both low- and high-level encoded features, shown as Fig. 1b (iii). From the Fig. 1b (iv), we can see that compared with vanilla transformer, only adopting our SR can achieve an improvement of 1.8 CIDEr points with a slight parameters increase (*i.e.*, 1.05M), while $\mathcal{M}^2$ with a mesh operation increases parameters (*i.e.*, 4.72M) and decreases the CIDEr by 1.1. To summarize, our contributions are threefold:

- We devise a $\mathcal{S}^2$ Transformer, a simple yet effective method, which extends the vanilla transformer framework to fully exploit gird visual features in terms of spatial and scale perception.
- We propose a SP module, which generates valid pseudo-region features for grid features to capture spatial information based on their clustering information. Moreover, we propose a simple SR module that further takes advantage of both low- and high-level encoded features without excessive increasing model size.
- We comprehensively evaluate our approach ($\mathcal{S}^2$ Transformer) on the MSCOCO benchmark. Experimental results demonstrate that our method performs best while maintaining the small model size.

2 $\mathcal{S}^2$ Transformer

In this section, we present a novel Spatial- and Scale-aware Transformer (**$\mathcal{S}^2$ Transformer**) for image captioning. The overview of the architecture is depicted in Fig. 2.

2.1 Overview

Given an image I, the task of image captioning is to automatically generate a description D about visual contents in images, following the paradigm of an encoder-decoder framework. Technically, $\mathcal{S}^2$ Transformer first applies a feature extraction to obtain gird features $G = \{g_m\}_{m=1}^{M}$ about an image, where M indicates the number of grids. As for the spatial information loss caused by flattening operation when feeding G into an encoder-decoder model, our proposed **Spatial-aware Pseudo-supervised** (SP) module is adopted to implicitly learn possible and discriminative regions to obtain pseudo-region features P:

$$P = \{p_n\}_{n=1}^{N} = SP(G), \tag{1}$$

where N means the number of pseudo regions. Then, we use the same encoder to exploit the visual information of original grid features G and pseudo-region features P simultaneously:

$$\begin{aligned} \bar{G} &= Encoder(G), \\ \bar{P} &= Encoder(P), \end{aligned} \tag{2}$$

where the Encoder is consistent with the vanilla Transformer's encoder without any modifications, which consists of two main components: Multi-head Self-Attention (MSA), and Feed Forward Network (FFN). Note that for the sake of concise expression, positional encoding, residual operation and layer normalization are omitted.

Different from previous Transformer-based models, which only feed the encoded feature obtained from the top encoder layer to the decoder, our proposed **Scale-wise Reinforcement** (SR) module is to simultaneously explore both low- and high-level encoded features to obtain augmented encoded features V:

$$\begin{aligned} V_G &= SR(\bar{G}_1, \bar{G}_2, ..., \bar{G}_L), \\ V_P &= SR(\bar{P}_1, \bar{P}_2, ..., \bar{P}_L), \end{aligned} \tag{3}$$

where $\bar{G}_L$ (or $\bar{P}_L$) means the output of L-th encoder layers and V_G and V_P represent grid and pseudo-region augmented encoded features, respectively. Finally, we fuse V_G and V_P to obtain the final encoded features V^* and feed it to the decoder:

$$\begin{aligned} V^* &= [V_G; V_P]W_V, \\ D &= Decoder(V^*), \end{aligned} \tag{4}$$

where $[;]$ means the operation of concatenate, W_V is a learnable parameter and the decoder is the same as the vanilla Transformer's decoder. The detail of our two main components (SP and SR) is described in the next subsection.

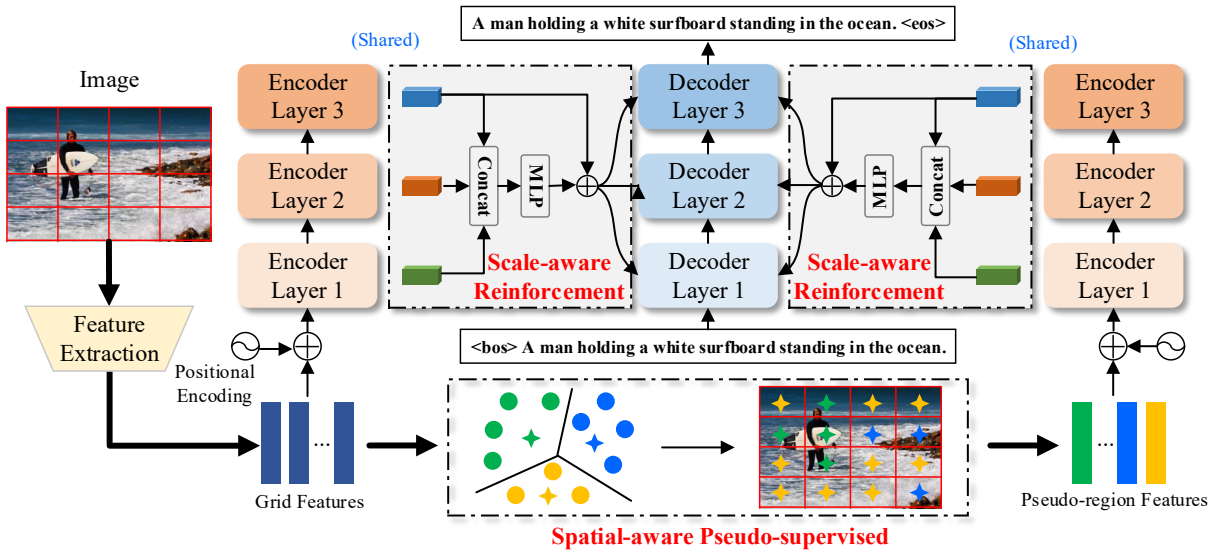

Figure 2: Overview of our proposed $\mathcal{S}^2$ Transformer architecture for image captioning. It consists of five main components: Feature extraction, Encoder, Decoder, Spatial-aware Pseudo-supervised (SP), and Scale-aware Reinforcement (SR), where the encoder and decoder both adopt the same as that of the vanilla Transformer without any modification. SP resorts to feature clustering to help preserve spatial information for grid features while SR simultaneously explores both low- and high-level encoded features. Note that both two encoders and two SRs respectively share parameters.

2.2 Spatial-aware Pseudo-supervised (SP) Module

As discussed above, directly operating at grid features leads to the loss of spatial information of regions. A plain idea is introducing region features to compensate for the deficiency. However, combining gird and region features will inevitably increase the computational complexity of the model. Intuitively, if we implicitly select and aggregate discrete grid features into several sub-spaces to obtain pseudo-region features, this operation would become more flexible. Motivated by this spirit, we propose a SP module to cluster the grid features with multiple centroids without explicit supervision. The purpose of these centroids is to integrate grids features of similar semantic information together to represent possible and discriminative regions.

Formally, in SP, we first design N learnable clusters as $C = \{c_1, ..., c_N\}$. Following [Arandjelovic *et al.*, 2016], we calculate the similarity between grid features and clusters by dot-product. Given each grid feature g_m, it can be mapped to the n-th cluster in the following manner:

$$r_{m,n} = \frac{exp(g_m c_n^T + b_n)}{\sum_{k=1}^{N} exp(g_m c_k^T + b_k)}, \tag{5}$$

where $b_{\{n,k\}}$ is a trainable parameter. The feature representation of each center p_n is obtained by a weighted integration of all grid features:

$$p_n = Norm(\sum_{m=1}^{M} r_{m,n}(g_m - \tilde{c}_n)), \tag{6}$$

where "$Norm$" means ℓ_2-normalization operation and $\tilde{c}_n$ is a learnable parameter which has the same size as c_n. Thus, we define the final features P as pseudo-region features.

2.3 Scale-aware Reinforcement (SR) Module

Recently, transformer-based captioning models have been proved helpful for image captioning. However, existing models neglect the low-level semantic information from the bottom of the encoder layer during the decoding process. Although [Cornia *et al.*, 2020] has provided a solution with a complex meshed cross-attention, we further propose a novel and simple SR module to address the above limitations by incorporating all features from each encoding layer into the top features.

For simplicity, we take gird features as an example. Specifically, given the output features $(\bar{G}_1, \bar{G}_2, ..., \bar{G}_L)$ of each encoder layer, we first concatenate them all together:

$$\tilde{G} = [\bar{G}_1, ..., \bar{G}_L]. \tag{7}$$

Then, to integrate both low- and high-level visual information, we employ a Multi-Layer Perception (MLP) which can weigh the contribution of features of each layer:

$$G^{'} = (\tilde{G}W_1^T)W_2^T, \tag{8}$$

where W_1 and W_2 are trainable projection matrices.

Since the output of the top encoder layer contains more important visual information, to prevent the insertion of additional noise perturbations, we add features $G^{'}$ to $\bar{G}_L$ to obtain the final grid augmented encoded features V_G:

$$V_G = G_L + \lambda G^{'}, \tag{9}$$

where the λ is an adjustable weighting factor. In a same way, we obtain the pseudo-region augmented encoded features V_P.

2.4 Training

Generally, the training of captioning model is split into two stages [Rennie *et al.*, 2017; Zhang *et al.*, 2021]. In the first

SP	SR	B@1	B@4	M	R	C	S	Param. ↓
✗	✗	80.9	38.9	29.0	58.5	131.2	22.7	33.57M
✓	✗	**81.3**	39.6	29.4	59.0	133.2	22.7	33.59M (↑0.02)
✗	✓	81.0	39.5	29.4	58.9	133.0	22.8	34.62M (↑1.05)
✓	✓	81.1	**39.6**	**29.6**	**59.1**	**133.5**	**23.2**	34.64M (↑1.07)

Table 1: Ablation studies of the proposed Spatial-aware Pseudo-supervised (SP) module and Scale-aware Reinforcement (SR) module.

Model	B@1	B@4	M	R	C	S	FLOPs
G	80.9	38.9	29.0	58.5	131.2	22.7	0.92G
R	80.0	38.8	28.7	58.5	130.2	22.3	0.76G
P	80.3	38.2	28.5	57.9	127.6	22.5	0.35G
$G+R$	81.0	39.0	29.2	58.7	131.5	22.7	1.35G
$G+P$	**81.3**	**39.6**	**29.4**	**59.0**	**133.2**	**22.7**	0.96G

Table 2: Ablation studies of different visual features. All models both adopt vanilla transformer without SR. G, R and P represent gird features, region features and our pseudo-region features, respectively.

stage, we utilize cross-entropy loss to optimize our model:

$$L_{CE} = -\sum_{t=1}^{T} log(p_\theta(w_t^*|w_{1:t-1}^*)), \tag{10}$$

where T is the length of word sequence and $w_{1:t-1}^*$ is the ground truth tokens in the description D.

In the second stage, we adopt the strategy of reinforcement learning, which exploits the CIDEr score as reward $r(\cdot)$ with self-critical sequence training [Rennie *et al.*, 2017]:

$$L_{RL} = -E_{w1:T\ p\theta}[r(w_{1:T})]. \tag{11}$$

In addition, we employ the gradient expression in [Cornia *et al.*, 2020], which computes the reward baseline of the reward by the mean operation of rewards, rather than greedy decoding. A sample's gradient expression is defined as:

$$\begin{cases} b = \frac{1}{k}(\sum_{i}^{k} r(w_i)), \\ \nabla_\theta L_{RL} \approx -\frac{1}{k}\sum_{i=1}^{k}((r(w_{1:T}^i) - b)\nabla_\theta \log p_\theta(w_{1:T}^i)), \end{cases} \tag{12}$$

where k is the number of sampled sequences, $w_{1:T}^i$ denotes the i-th sampled sequence, and b represents the average reward earned by the sampled sequences.

3 Experiments

3.1 Experimental Settings

Dataset and Metric. We conduct experiments to verify the effectiveness of our proposed $\mathcal{S}^2$ Transformer on commonly-used image captioning dataset, *i.e.*, MS-COCO. It consists of 123,287 images, each associated with five different descriptions. In offline testing, we follow the setting in [Karpathy and Fei-Fei, 2015], where 113,287 images, 5,000 images, and 5,000 images are used as train, validation, and test set, respectively. The online evaluation is done on the COCO online test server, where ground-truth annotations of 40,775 images are not publicly provided. We measure the captioning performance using the standard evaluation metrics, including BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], ROUGR [Lin, 2004], CIDEr [Vedantam *et al.*, 2015], and SPICE [Anderson *et al.*, 2016].

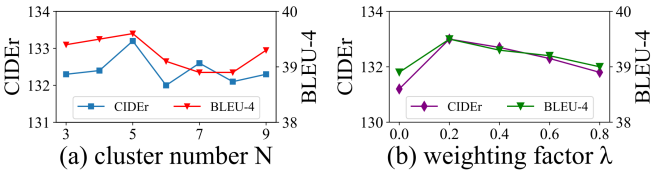

Figure 3: Ablation studies of cluster number N in SP and weighting factor λ in SR. Note that (a) and (b) only use SP and SR, respectively.

Model	B@1	B@4	M	R	C	S	Param.↓
Transformer	80.9	38.9	29.0	58.5	131.2	22.7	33.57M
AoA Transformer	80.8	39.1	29.1	59.1	130.3	22.7	87.37M (↑53.80)
M^2 Transformer	80.8	38.9	29.1	58.5	131.8	22.7	38.66M (↑5.09)
X-Transformer	81.0	**39.7**	29.1	59.0	130.2	22.8	56.94M (↑23.37)
RSTNet	81.1	39.3	29.4	58.8	133.3	23.0	156.31M (↑122.74)
Ours	**81.1**	39.6	**29.6**	**59.1**	**133.5**	**23.2**	34.64M (↑1.07)

Table 3: Comparing with the state of the art on ResNext101 grid features.

Implementation Details. Following [Zhang *et al.*, 2021], we adopt the same pre-trained Faster-RCNN [Ren *et al.*, 2015] provided by [Jiang *et al.*, 2020] to extract grid features, where the gird shape is 7×7 and the dimension of each grid is 2,048. In practice, our encoder and decoder both have 3 layers, where each layer uses 8 self-attention heads and the inner dimension of FFN is 2,048. The number of cluster centers N is 5 and the hyper-parameter $\lambda = 0.2$ in Eq. 9.

We employ Adam optimizer to train all models and set batch size as 50. For cross-entropy (CE) training, we set the minimum epoch as 15. If CIDEr drops in 5 consecutive epochs, we will choose the model with the best CIDEr score for self-critical sequence training. Specifically, we use an epoch decay schedule to adjust the learning rate for CE by

Model	B@1	B@4	M	R	C	S
SCST	-	34.2	26.7	55.7	114.0	-
Up-Down	79.8	36.3	27.7	56.9	120.1	21.4
RFNet	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM	80.5	38.2	28.5	58.3	127.6	22.0
SGAE	80.8	38.4	28.4	58.6	127.8	22.1
ORT	80.5	38.6	28.7	58.4	128.3	22.6
AoANet	80.2	38.9	29.2	58.8	129.8	22.4
M^2 Transformer	80.8	39.1	29.2	58.6	131.2	22.6
TCIC	80.9	39.7	29.2	58.6	132.9	22.4
X-Transformer	80.9	**39.7**	29.5	59.1	132.8	23.4
RSTNet	81.1	39.3	29.4	58.8	133.3	23.0
Ours	**81.1**	39.6	**29.6**	**59.1**	**133.5**	23.2

Table 4: Performance comparision with the state-of-the-art on the MS-COCO "Karpathy" test split.

Model	B@1		B@2		B@3		B@4		METEOR		ROUGE-L		CIDEr-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST	78.1	93.7	61.9	86.0	47.9	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.8	37.2	58.2	37.1	122.9	125.1
GCN-LSTM	80.8	95.9	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
ETA	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
AoANet	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
M^2 Transformer	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer (ResNet-101)	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
X-Transformer (SENet-154)	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
RSTNet (ResNext101)	81.7	96.2	66.5	90.9	51.8	82.7	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4
RSTNet (ResNext152)	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
Ours (ResNext101)	**81.9**	**96.4**	66.7	**91.3**	52.1	**83.1**	40.0	**73.1**	29.5	**39.2**	59.2	74.7	**131.5**	**134.5**
Ours (ResNext152)	**82.2**	**96.5**	**67.0**	**91.4**	**52.4**	**83.3**	40.1	**73.5**	**29.6**	**39.3**	**59.5**	**75.0**	**132.6**	**135.0**

Table 5: Leaderboard of the published state-of-the-art image captioning models on the MS-COCO online testing server.

following [Zhang *et al.*, 2021]:

$$lr = \begin{cases} n/4 \times \text{1e-4}, & n \leq 3, \\ \text{1e-4}, & 3 < n \leq 10, \\ 0.2 \times \text{1e-4}, & 10 < n \leq 12, \\ 0.2 \times 0.2 \times \text{1e-4}, & otherwise, \end{cases} \tag{13}$$

where n denotes the number of current epoch. For self-critical sequence training, the learning rate is fixed at $5\times$1e-7.

3.2 Ablation Studies

The core of our proposed $\mathcal{S}^2$ Transformer is to generate the high-quality visual descriptions by introducing a spatial-aware pseudo-supervised (SP) module and a scale-wise reinforcement (SR) module into a vanilla transformer model. In this section, we conduct comprehensive ablation studies to prove the effectiveness of our method.

Effect of SP and SR. Tab. 1 gives the results of four control experiments to investigate the impact of our proposed SP and SR modules: i) baseline: adapting vanilla transformer model without any modifications, ii) baseline+SP: integrating SP into baseline, iii) baseline+SR: integrating SR into baseline, and iv) baseline+SP+SR: integrating both SP and SR into baseline. Obviously, the performances are enhanced by individually adding SP and SR to the baseline, particularly improving 2.0 and 1.8 points on CIDEr, respectively. Moreover, the combination of the two components achieves further improvement. Also, we report the parameters of each model for measuring its complexity. SP and SR slightly increase parameters by 0.02M and 1.05M compared with the baseline. To sum up, our proposed components achieve huge improvements with a small computational cost, indicating our methods' effectiveness.

Effect of Pseudo-region feature. In Tab. 2, we execute several experiments to examine the effect of different visual features, including grid features (G), region features (R) and our pseudo-region features (P). All models both adopt a vanilla transformer. Using only a single feature, our P performs worst, which indicates that only using pseudo-region features may lose some important visual information. Combining two features (*i.e.*, $G+P$ and $G+R$) can bring performance gains. Meanwhile, $G+P$ obtains more significant improvement than $G+R$, thus indicating the practicality of our pseudo-region features. Besides, in terms of FLOPs, $G+R$ brings an excessive increase of 0.43G while G+P brings a slight increase of 0.04G. The results demonstrate that the highly abstract pseudo-region features are sufficient and complementary for grid features instead of directly using explicit region features.

Effect of N in SP. To determine how many pseudo regions the model needs to learn, we set the range of cluster number N from 3 to 9 as shown in Fig. 3a. Note that our SP is serving for high-level semantic information extraction. From the figure, we can observe that if N is too large, it may be difficult for the model to find discriminative pseudo regions, which harms the performance of the model. On the contrary, if N is too small, much weak semantics will be discarded in large quantities, resulting in poor results. Our approach achieves the best results with N = 5 clusters.

Effect of λ in SR. To choose the best weighting factor λ in Eq. 9, we conduct a series of experiments by setting the different values of λ. The results are shown in Fig. 3b. We find that the performance drastically drops with the increase of λ and the best results are obtained when $\lambda = 0.2$. It reveals that too larger λ introduces more redundant noise for the decoder. Thus, we set $\lambda = 0.2$ in the final model.

Fair comparison with strong transformer-based baselines. For a fair comparison, we report experimental results utilizing the same ResNext101 grid feature as the visual input shown in Tab. 3. All models are based on improved versions of the vanilla transformer. Specifically, our method achieves state-of-the-art performance on most metrics except B@4, which demonstrates superior performance without the interference of diverse features. Moreover, compared to the SOTA method RSTNet, which increases the Transformer parameters by 122.74M, our model brings only slight growth of 1.07M on parameters. It further demonstrates that our method can effectively and efficiently incorporate grid features with transformer-based architecture.

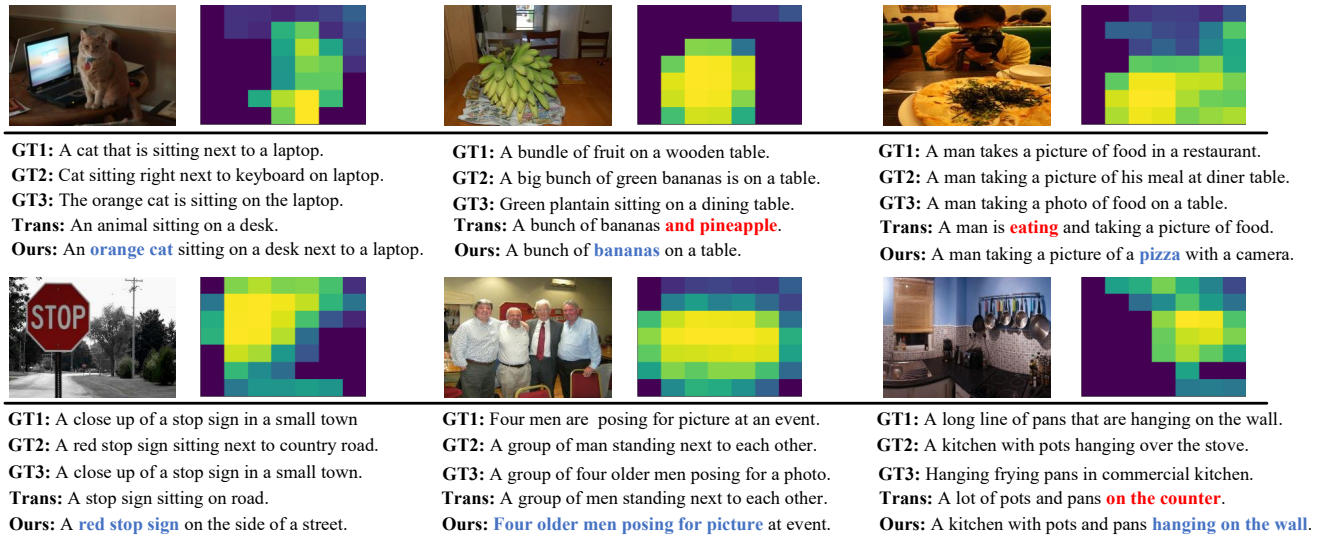

Figure 4: Visualization of the proposed $\mathcal{S}^2$ Transformer. Each example consists of a raw image, a learned map of cluster indices by SP, the ground-truth descriptions, and the generated description by the transformer and ours. The size of these learned maps is 7×7.

3.3 Quantitative Analysis

Compared Methods. In this section, we compare our proposed $\mathcal{S}^2$ Transformer with the state-of-art methods both on offline and online evaluation, including SCST [Rennie *et al.*, 2017], Up-Down [Anderson *et al.*, 2018], RFNet [Jiang *et al.*, 2018], GCN-LSTM [Yao *et al.*, 2018], SGAE [Yang *et al.*, 2019], ORT [Herdade *et al.*, 2019], AoANet [Huang *et al.*, 2019], $\mathcal{M}^2$Transformer [Cornia *et al.*, 2020], X-Transformer [Pan *et al.*, 2020], TCIC [Fan *et al.*, 2021] and RSTNet [Zhang *et al.*, 2021].

Offline Evaluation. In Tab. 4, we show the image captioning results of our method and compare it to the aforementioned competitors on the offline test split. Overall, our method outperforms all compared methods in terms of B@1, M, R, C, and S. Specifically, compared with the best counterpart RSTNet using extra knowledge from a pre-trained language model, our method yields better gains on all metrics, demonstrating the superiority of our approach.

Online Evaluation. To further verify the benefit of our $\mathcal{S}^2$ Transformer, we estimate it on the online COCO test server. Following the compared methods, we integrate the results of four models with different initialization for testing. The comparison results are summarized in Tab. 5. It is clear that our $\mathcal{S}^2$ Transformer outperforms state-of-the-art models on most metrics. Particularly, with respect to the best competitor RSTNet (ResNext152), our method $\mathcal{S}^2$ Transformer with ResNext152 achieves improvements of 0.7 and 1.0 CIDEr points on 5 reference captions (c5) and 40 reference captions (c40), respectively.

3.4 Visualization

Fig. 4 provides some qualitative results to show the pseudo regions learned via the proposed SP in heat maps and the high-quality descriptions generated by our proposed model. In the heat map, different colors represent different index values, which indicate different pseudo-regions. As we can see, SP focuses on specific visual regions in foregrounds but also reserves discriminative background information, confirming the usefulness of exploiting pseudo regions to retain the spatial information. Besides, our model can generate more accurate and diverse descriptions compared to basic transformer model. More visualizations are included in the supplementary material.

4 Conclusion

In the paper, we study how to effectively and efficiently incorporate grid features with transformer-based architecture for image captioning. To achieve this target, we propose a $\mathcal{S}^2$ Transformer—a simple yet effective approach that implicitly learns pseudo regions through a series of learnable clusters in a SP module and simultaneously explores both low- and high-level encoded features in a SR module. Noticeably, pseudo regions can effectively capture spatial information lost by the flattening operation of gird features. Extensive experiments on the MSCOCO benchmark and visualization analysis confirm the effectiveness and interpretability of our method. Besides, our approach does not bring excessive parameters compared with the vanilla transformer.

Broader Impact. Our paper focuses on learning image captioning tasks, which has broader application in real-world scenarios such as human-machine interaction and visual-impaired assistance. Our method provides positive impacts, including 1) implicitly learning discriminative region features instead of using explicit region features, which reduces the increase in parameters and computation, and 2) providing a simple task-specific transformer-based model, which generates more high-quality descriptions. However, it is still challenging to deploy existing models into real-world scenarios because of their susceptibility to attacks, which remains our responsibility to grow awareness of these potential dangers.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62020106008, No. 62122018, No. 61772116, No. 61872064), Sichuan Science and Technology Program (Grant No.2019JDTD0005).

References

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[Arandjelovic *et al.*, 2016] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005.

[Chen *et al.*, 2021] Wenqing Chen, Jidong Tian, Caoyun Fan, Hao He, and Yaohui Jin. Dependent multi-task learning with causal intervention for image captioning. In *IJCAI*, 2021.

[Cornia *et al.*, 2019] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, 2019.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.

[Fan *et al.*, 2021] Zhihao Fan, Zhongyu Wei, Siyuan Wang, Ruize Wang, Zejun Li, Haijun Shan, and Xuanjing Huang. TCIC: theme concepts learning cross language and vision for image captioning. In *IJCAI*, 2021.

[Herdade *et al.*, 2019] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NeurIPS*, 2019.

[Huang *et al.*, 2019] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[Jiang *et al.*, 2018] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, 2018.

[Jiang *et al.*, 2020] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, 2020.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[Krishna *et al.*, 2017] 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. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[Li *et al.*, 2019] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *ICCV*, 2019.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2004.

[Pan *et al.*, 2020] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[Rennie *et al.*, 2017] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NeurIPS*, 2014.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[Vinyals *et al.*, 2016] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *TPAMI*, 2016.

[Wang *et al.*, 2020] Ziwei Wang, Zi Huang, and Yadan Luo. Human consensus-oriented image captioning. In *IJCAI*, 2020.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[Yang *et al.*, 2019] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

[Yao *et al.*, 2018] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[Zhang *et al.*, 2021] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *CVPR*, 2021.