

Dual-stream Self-attention Network for Image Captioning

Boyang Wan¹, Wenhui Jiang¹, Yuming Fang¹, Wenying Wen¹, Hantao Liu²

¹Jiangxi University of Finance and Economics, Nanchang, China

²Cardiff University, Cardiff, United Kingdom

Abstract—Self-attention based encoder-decoder models achieve dominant performance in image captioning. However, most existing image captioning models (ICMs) only focus on modeling the relation between spatial tokens, while channel-wise attention is neglected for getting visual representation. Considering that different channels of visual representation usually denote different visual objects, it may lead to poor performance in terms of object and attribute words in the captioning sentences generated by the ICMs. In this paper, we propose a novel dual-stream self-attention module (DSM) to alleviate the above issue. Specifically, we propose a parallel self-attention based module that simultaneously encodes visual information from the spatial and channel dimensions. Besides, to obtain channel-wise visual features effectively and efficiently, we introduce a group self-attention block with linear computational complexity. To validate the effectiveness of our model, we conduct extensive experiments on the standard IC benchmarks including MSCOCO and Flickr30k. Without bells and whistles, the proposed model performs new SOTAs containing 135.4 CIDEr score on MSCOCO and 70.8 CIDEr score on Flickr30k.

Index Terms—Image Captioning, Self-Attention, Spatial Attention, Channel Attention

I. INTRODUCTION

Image captioning (IC) is a fundamental cross-media task that automatically generates a natural language description for an image [39], [40]. The encoder-decoder framework, first proposed for neural machine translation task [35], dominates the IC task. In the early stage, researchers propose image captioning models (ICMs) containing CNN-based encoders to extract visual representation [20], [36], [39] and RNN-based decoders [20], [39] to generate the caption sentences. Next, the attention mechanism is widely introduced in ICMs. As a pioneer, Kelvin *et al.* [43] propose an attention based LSTM to generate image caption by using CNN-based visual features. In [1], Anderson *et al.* present the seminal work “Bottom-Up and Top-Down” to implement the bottom-up and top-down attention as a pretrained object detection model and a two-layer LSTM, respectively.

Recently, the self-attention mechanism [37] is proposed to improve natural language processing (NLP) models and further dominates multiple research areas including computer vision [8], [23] and cross-media [4], [14], [34], [41]. Naturally, self-attention based ICMs become a potential topic, and many studies have been published in [4], [12], [15], [19], [48]. Firstly, Herdade *et al.* [12] integrate the region and geometric features by using a standard Transformer. Then, researchers focus on modifying self-attention architecture for image captioning. In [15], a high-order self-attention module is proposed

This work has been supported by the National Natural Science Foundation of China (No.62161013, No.62132006, No.61901197 and No.62162029), National Natural Science Foundation of Jiangxi Province (No.20212BAB202011), Shenzhen Municipal Science and Technology Innovation Council Under Grant 2021Sszvp051.

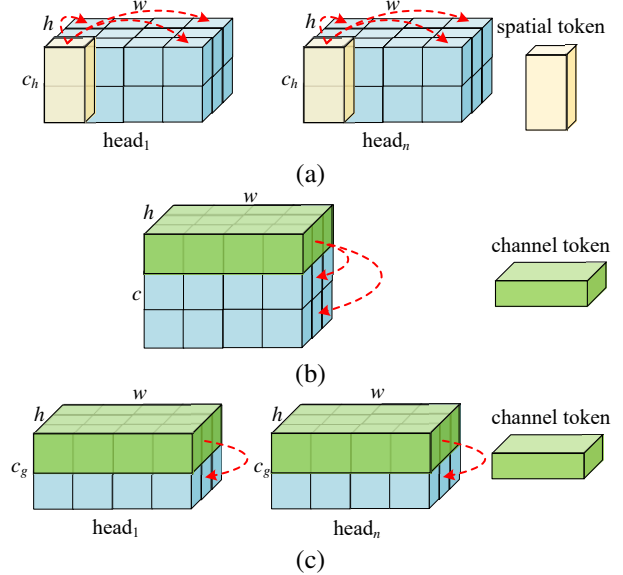


Fig. 1. Comparison of self-attention based modules used in Transformer-based ICMs, where w , h , c , c_h and c_g indicate width, height, channels per head and channels per group. (a) spatial-wise multi-head self-attention modules used in [4], [12], [15], [48]. (b) channel-wise self-attention module in [25]. (c) Our channel-wise group self-attention module.

to integrate into an LSTM model. Besides, Cornia *et al.* [4] propose a Transformer based model containing a meshed connection between its encoder and decoder. Compared to other attention based ICMs, the self-attention ICMs obtain better performance on the standard captioning dataset including MSCOCO [3] and Flickr30k [47]. However, most existing self-attention ICMs only build relation among spatial tokens by using the vanilla multi-head self-attention (MSA) module, as shown in Fig. 1-(a). Different from spatial attention, which determines *where to pay attention*, channel attention [13], [29], [42], which adaptively reweights features in the channel dimension and can be seen as an object selection process [11], decides *what to pay attention*. SENet [13] is a pioneer channel attention based model with a squeeze-and-excitation (SE) block, which is proposed to build relation among channel tokens. The SE block improves the performance of SENet on several CV tasks and becomes a popular channel attention block widely applied in CNN models [32], [42]. To address the local informative loss caused by the squeeze operation, which is implemented as the global average pooling, in the SE block, Song *et al.* [29] propose a discrete cosine transform (DCT) channel attention block to build relation among channel tokens with multiple frequency components. Inspired by the success of the self-attention on spatial tokens, directly applying the SA on channel tokens is an intuitive idea and Fu *et*

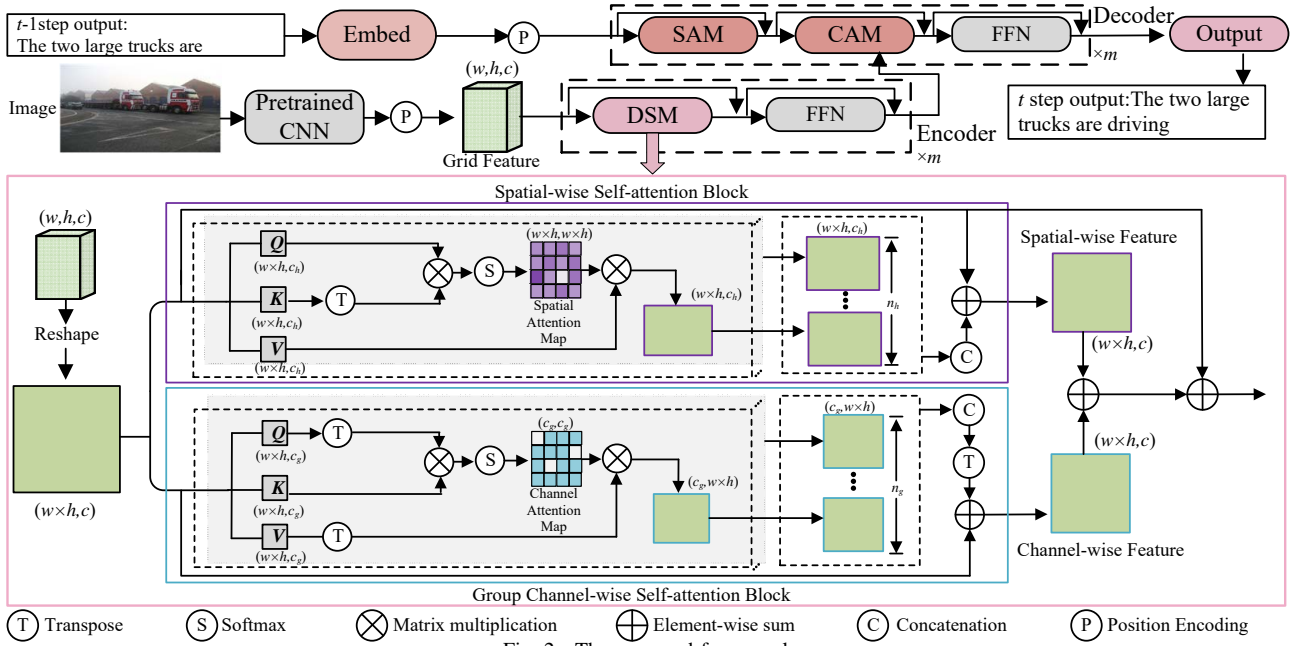


Fig. 2. The proposed framework.

al. [9] propose a channel-wise self-attention block (CSB) that is demonstrated in Fig. 1-(b) and achieves an improvement on the semantic segmentation. Besides, some works [6], [32] demonstrate that combining spatial and channel attention is a potential way to get an effective visual representation. To the best of our knowledge, SDATR [25] is the only ICM that uses both spatial and channel attention simultaneously and reveals that applying the CSB achieves slight improvement for image captioning and is experimentally expensive. Since the CSB is a single-head self-attention block, it has quadratically computation complexity for token numbers and the number of channel tokens usually exceeds one thousand. Building relation among one thousand channel tokens with about tens dimension features of each channel token suffers from the information loss caused by the low-rank attention map of the channel tokens.

In this paper, we build a novel ICM including a self-attention based encoding module named dual-stream self-attention module (DSM). The DSM generates discriminative visual features by parallel building relations among spatial and channel tokens. Besides, to obtain a channel-wise visual feature effectively and efficiently, we introduce a group channel-wise self-attention block (GCSB) with linear computation complexity. As illustrated in Fig. 1-(c), we divide the visual features into multiple groups uniformly along the channel dimension and then apply SA to the visual features of each group.

To validate the effectiveness of our model, we conduct extensive experiments on the MSCOCO and Flickr30k datasets. Without bells and whistles, the proposed model performs new SOTAs on both MSCOCO and Flickr30k datasets. In summary, we would like to make the following technical contributions: 1.) Proposing a novel ICM that first uses a pure self-attention based module to get visual representation encoding spatial- and channel-wise information simultaneously; 2.) Obtaining effective channel-wise features requires only linear computation complexity; 3.) Achieving new SOTAs on mainstream image captioning datasets.

II. METHODOLOGY

In this section, we first revisit the self-attention mechanism in Transformers. Then, we introduce the proposed dual-stream self-attention model, which encodes both channel and spatial representation in parallel. Finally, we present the training method used in this paper.

A. Preliminaries

In this paper, the visual or language feature is denoted as $\mathbf{F} \in \mathbb{R}^{N \times C}$, where N and C indicate the token number and feature dimension of tokens, respectively. The standard multiple head self-attention MSA is defined as:

$$\text{MSA}(\mathbf{F}) = \text{Concat}(\{\text{SA}_i(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)\}_{i=0}^{N_h})\mathbf{W}_o \quad (1)$$

$$\text{SA}_i(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{Softmax} \left[\frac{\mathbf{Q}_i(\mathbf{K}_i)^T}{\sqrt{C_h}} \right] \mathbf{V}_i \quad (2)$$

where $\mathbf{Q}_i = \mathbf{X}_i \mathbf{W}_i^Q$, $\mathbf{K}_i = \mathbf{X}_i \mathbf{W}_i^K$, and $\mathbf{V}_i = \mathbf{X}_i \mathbf{W}_i^V$ are $\mathbb{R}^{N \times C_h}$ dimensional visual or language features for i -th head SA_i , \mathbf{X}_i denotes the i -th head of the input feature and $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ denote the linear projection layer of the i -th head for $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, respectively. $C = C_h * N_h$ and $\{\mathbf{X}_i = \mathbf{F}[:, i:C_h] | i = 1, 2, \dots, N_h\}$. The computational complexity of MSA is approximated as $O(N^2 C)$.

B. Dual stream self-attention model

As shown in Fig. 2, our model is an encoder-decoder based ICM. First, an image is fed into a pretrained CNN [17] to extract the visual grid feature denoted as a $w \times h \times c$ dimension tensor. Then we put the grid feature with an absolute position encoding into an encoder including m encoding layers. Each encoding layer contains a dual self-attention module (DSM) and a standard feed-forward layer. In addition, the DSM includes two parallel blocks named spatial-wise self-attention block (SSB) and the GCSB. Finally, the decoder of our model fuses the output of the encoder and language embedding to generate a caption of the input image word-by-word.

Spatial-wise Self-attention Block: Previous self-attention based ICMs usually determine tokens with grid or region features \mathbf{F} and encode the information along spatial dimensions. SSB also computes self-attention within spatial-wise tokens, as shown in Fig. 2. and can be represented as follow:

$$\text{SSB}(\mathbf{F}) = \text{LN}(\mathbf{F} + \text{D}(\text{MSA}(\mathbf{F}))) \quad (3)$$

where $\text{LN}(\cdot)$, $\text{D}(\cdot)$ represent Layer Normalization and Dropout function, respectively. The computational complexity of SSB is approximated as $O((h \times w)^2 C)$ and is affordable due to the number of spatial tokens for the visual grid feature being fixed and small. The h and w are set as 7 in this paper. Considering the relative position of visual objects is crucial for IC, we introduce a learning based relative position representation [31] into SSB.

Group Channel-wise Self-attention Block: Instead of only applying the self-attention mechanisms on spatial dimensions, we introduce it to the channel dimension of the grid features, *i.e.*, each channel of all grids is set to a token for channel self-attention encoding.

To reduce the computational complexity and improve the capacity of self-attention based channel attention block, we introduce the GCSB in our model. The GCSB divides channel tokens into multiple groups along the channel dimension and defines each group as a head in the multi-head self-attention model. The number of groups and the number of channels in each group are denoted as N_g and C_g . The GCSB can be defined as follow:

$$\text{GCSB}(\mathbf{F}) = \text{LN}(\mathbf{F} + \text{D}(\text{MSAC}(\mathbf{F}))) \quad (4)$$

where $\text{LN}(\cdot)$, $\text{D}(\cdot)$ and $\text{MSAC}(\cdot)$ represent Layer Normalization, Dropout function and a channel-wise multi-head self-attention model, respectively. The channel-wise multi-head self-attention model can be formulated as follow:

$$\text{MSAC}(\mathbf{F}) = \text{Concat}(\{\text{SA}_i^{\text{group}}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)^T\}_{i=0}^{N_g}) \mathbf{W}_o \quad (5)$$

$$\text{SA}_i^{\text{group}}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{Softmax} \left[\frac{\mathbf{Q}_i^T \mathbf{K}_i}{\sqrt{C_g}} \right] \mathbf{V}_i^T \quad (6)$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{N \times C_g}$ are queries, keys, and values of group channel self-attention. Compared with CSB, the computational complexity of GCSB reduce from $O((h \times w)C^2)$ to $O((h \times w)(\frac{C^2}{N_g}))$. In general, each $\text{SA}_i^{\text{group}}$ get a local visual representation with partial channel tokens, and the linear project layer \mathbf{W}_o fuses all N_g local visual representation into a global channel-wise feature of images.

Fusion Block: In this paper, we fuse the spatial-wise feature generated from the SSB and the channel-wise feature generated from the GCSB by using residual and addition operation. The output of DSM can be get as follow:

$$\text{DSM}(\mathbf{F}) = \text{LN}(\mathbf{F} + \text{D}(\text{SSB}(\mathbf{F}) + \text{D}(\text{GCSB}(\mathbf{F})))) \quad (7)$$

C. Optimization

In this paper, We use a standard two-stage optimization paradigm including cross-entropy (XE) and self-critical sequence training (SCST) [30] optimization stages. Firstly and in the XE stage, our model is optimized by using a cross-entropy objection function that can be formulated as follow:

$$L_{XE}(\theta) = - \sum_{t=1}^T \log(p_\theta(y_t^* | y_{1:t-1}^*)) \quad (8)$$

where $y_{1:T}^*$ is the groundtruth sentence, and θ denotes the parameters of our model. Then, we adopt SCST to optimize our model according to the CIDEr metric [38]:

$$L_R(\theta) = - \mathbf{E}_{y_{1:T} \sim p_\theta} [r(y_{1:T})] \quad (9)$$

where $r(\cdot)$ denotes the CIDEr score of generated captioning sentences. The gradient of L_R can be approximated as follows:

$$\nabla_\theta L_R(\theta) \approx - (r(y_{1:T}^s) - r(\hat{y}_{1:T}^s)) \nabla_\theta \log p_\theta(y_{1:T}^s) \quad (10)$$

where $y_{1:T}^s$ represents a sampled caption and $r(\hat{y}_{1:T}^s)$ defines the CIDEr score of sentences obtained by using greedy decode.

III. EXPERIMENTS

A. Experiments Setup

Datasets: In this paper, all experiments are conducted on a challenging dataset MSCOCO captioning [3] and Flickr30k [47]. MSCOCO captioning dataset contains 123287 images with five groundtruth sentences. All competing models are evaluated using offline splits, also named Karpathy splits, containing 113287 training images, 5000 validation images and 5000 testing images. Flickr30k dataset contains 31,783 images with five groundtruth sentences and includes about 29000 training images, 1000 validation images and 1000 testing images.

Evaluation Metric: Similar to previous works [27] [4], we use standard coco caption metrics including BLEU [28], METEOR [5], ROUGE-L [22] and CIDEr [38] to evaluate the competing ICMs.

Implementation Details: In this paper, we set the dimensionality C of each layer to 512 and the numbers of heads N_h and groups N_g to 8, respectively. So the channels per head C_h and channels per group C_g are 64 for our model. We use Dropout [33] with a probability of 0.1 in each attention and feed-forward layer. For a fair comparison, we adopt a learning rate setting from [25]. Specifically, the learning rate is set as 1e-4, 2e-5, 4e-6, 5e-6, 5e-7, 2.5e-7 and 5e-8 at epoch 1, 4, 10, 20, 25, 40, 50 and 55, respectively. The XE and SCST training stages require 25 and 35 epochs, respectively. We adopt the Adam optimizer [21] with a batch size of 50 and a beam size of 5 in all experiments.

B. Ablation Studies

TABLE I
COMPARISON OF ABLATION STUDY ON KARPATHY TEST SPLIT, WHERE B@N, M, R, C, AND S ARE SHORT FOR BLEU@N, METEOR, ROUGE-L, CIDEr, AND SPICE SCORE.

	B@1	B@4	M	R	C
GCSB	81.0	38.9	29.1	58.9	132.5
SSB	81.0	39.1	29.3	58.9	132.6
GCSB→SSB	81.2	39.1	29.3	58.9	133.7
SSB→GCSB	81.3	39.5	29.4	59.0	133.8
SSB+GCSB	81.5	40.0	29.6	59.2	135.4

The comparison results using different module architectures in Tab. I illustrate the boost brought by the proposed GCSB and the parallel architecture in our model. It should be noted that “GCSB→SSB” means that GCSB and SSB are linked by using the series connection, while “SSB+GCSB” means they are used in parallel. As shown in Tab. I, the performance of the GCSB is slightly lower than the SSB ones. It confirms that spatial-wise feature is nonsubstitutable for IC. Two series modules achieve similar performance and both outperform single block based modules. It verifies that combining spatial-wise and channel-wise attention is an effective method to

improve ICMs. Besides, the SSB+GCSB obtains the best performance in comparison indicating the parallel architecture better than series connection for combining spatial-wise and channel-wise features.

To further fairly demonstrate the effectiveness of the proposed channel-wise attention block GCSB, we compare our GCSB with off-the-shelf channel-wise attention blocks including SENet [13], CBAM [42], DCT [29], CSB [25] and CAB [25]. As shown in Table II, our GCSB outperforms other channel-wise attention blocks on all metrics. Therefore, it confirms the effectiveness of the proposed GCSB.

TABLE II
COMPARISON OF DIFFERENT CHANNEL ATTENTION BLOCKS ON KARPATY TEST SPLIT.

	B@1	B@4	M	R	C
SE [13]	81.0	39.3	29.1	59.0	133.5
CBAM [42]	81.3	39.5	29.5	59.1	133.8
DCT [29]	81.1	39.5	29.4	58.9	134.1
CSB [25]	81.1	39.1	29.2	58.7	133.0
CAB [25]	81.3	39.7	29.5	59.1	134.5
Ours	81.5	40.0	29.6	59.2	135.4

C. Comparison with the State-of-the-arts

TABLE III
COMPARISONS OF THE STATE-OF-THE-ART METHODS ON THE KARPATY TEST SPLIT.

	B@1	B@4	M	R	C
SCST [30]	-	34.2	26.7	55.7	114.0
Up-Down [1]	79.8	36.3	27.7	56.9	120.1
RFNet [18]	79.1	36.5	27.7	57.3	121.9
GCN-LSTM [45]	80.5	38.2	28.5	58.3	127.6
SGAE [44]	80.8	38.4	28.4	58.6	127.8
AoANet [15]	80.2	38.9	29.2	58.8	129.8
ORT [12]	80.5	38.6	28.7	58.4	128.3
Transformer [37]	80.7	38.6	29.1	58.5	130.1
\mathcal{M}^2 Transformer [4]	80.8	39.1	29.2	58.6	131.2
X-Transformer [27]	80.9	39.7	29.5	59.1	132.8
RSTNet [48]	81.1	39.3	29.4	58.8	133.3
SDATR [25]	81.3	39.7	29.5	59.1	134.5
Ours	81.5	40.0	29.6	59.2	135.4

Tab. III shows the comparison of our method against existing SOTA ICMs including SCST [30], Up-Down [1], RFNet [18], GCN-LSTM [45], SGAE [44], AoANet [15], ORT [12], Transformer [37], \mathcal{M}^2 Transformer [4], X-Transformer [27], RSTNet [48] and SDATR [25] on the MSCOCO Karpathy test split. As shown in Tab. III, our model outperforms the competing SOTA ICMs by a significant gap. Noticeably, compared with the SDATR, which also is dual-stream ICM and performs the highest scores on IC, our model obtained an absolute gain of 0.8% in terms of CIDEr and 0.3% in terms of Bleu@4 on the MSCOCO Karpathy test split.

To further validate the effectiveness of our method, we conduct the comparison experiment on the Flickr30k dataset. Tab. IV summarizes the image captioning performance on the Flickr30k dataset. Compared with the existing SOTAs on Flickr30k, our model boosts the performance on all metrics. Specifically, our model provides a relative gain of more than 5.98% in terms of CIDEr and 6.12% in terms of Bleu@4 on the test set of Flickr30k. We believe that the consistent improvement shown on multiple image captioning datasets proves the effectiveness of our model.

D. Visualization

We visualize some caption sentences generated by SDATR and our model in Fig. 3. Compared to SDATR, our model

TABLE IV
COMPARISONS OF THE STATE-OF-THE-ART METHODS ON THE FLICKR30K TEST SPLIT.

	B@1	B@4	M	R	C
Deep VS [20]	57.3	15.7	15.3	-	24.7
Google NIC [39]	66.4	18.3	-	-	-
m-RNN [26]	60.0	19.0	-	-	-
Soft-Attention [43]	66.7	19.1	18.5	-	-
Hard-Attention [43]	66.9	19.9	18.5	-	-
emb-gLSTM [16]	64.6	20.6	17.9	-	-
ATT [46]	64.7	23.0	18.9	-	-
Log Bilinear [7]	60.0	17.1	16.9	-	-
Adaptive [24]	67.6	25.1	20.4	-	53.1
SEM [2]	73.1	29.0	22.0	-	66.8
DA [10]	73.8	29.4	23.0	-	66.6
SDATR [25]	71.2	27.4	21.9	48.8	64.0
Ours	74.2	31.2	23.3	51.2	70.8


	SDATR: three dogs and a cat laying on a couch Ours: two dogs and a cat laying on a couch GT_1: two dogs and a cat take a nap on a couch GT_2: two dogs sleep on the couch near the grey cat GT_3: dogs and cat sleeping on big comfortable couch
	SDATR: a motorcycle parked on the side of a street Ours: a blue motorcycle parked in front of a truck GT_1: a blue motorcycle is parked on the street next to a truck GT_2: a blue motorcycle parked in front of a white truck GT_3: a blue motorcycle that is parked on the street
	SDATR: a small bird perched on a branch of a window Ours: a small bird sitting on a bird feeder GT_1: a small bird is perched on an empty bird feeder GT_2: a picture of a bird on a rustic looking feeder GT_3: a small bird perched on the edge of a bird feeder

Fig. 3. Visualization of 3 corresponding ground truth sentences and captions generated by SDATR and our model. The most significant difference between the two competing captions is highlighted in colors.

generates captions containing more accurate objects, which is demonstrated in the bottom row image of Fig. 3, and more detailed attributes of visual objects. For example, as shown in the second-row images of Fig. 3, the phrase “blue motorcycle” provide more accurate details of objects in generated captions. Quantifiers have a significant impact on the quality of caption sentences. Our model achieves better performance on quantifier generation, as shown in the first-row image of Fig. 3.

IV. CONCLUSION

In this paper, we propose a novel pure self-attention model for image captioning. Specifically, we design the DSM for simultaneously utilizing the spatial- and channel-wise visual representation and a linear computation complexity self-attention model named GCSB for getting the channel-wise feature efficiently. By encoding the visual presentation along with both spatial and channel dimensions, the DSM provides a better visual representation to generate more detailed captions for images. While the GCSB gets channel-wise features efficiently due to its linear computation complexity. Experiments on a challenging dataset demonstrate the effectiveness of our model for image captioning. In the future, we will investigate the interaction between spatial and channel attention to obtain a more robust visual representation for IC.

REFERENCES

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.
- [2] Wenjie Cai and Qiong Liu. Image captioning with semantic-enhanced features and extremely hard negative examples. *Neurocomputing*, 413:31–40, 2020.
- [3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [4] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10578–10587, 2020.
- [5] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 376–380, 2014.
- [6] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. *arXiv preprint arXiv:2204.03645*, 2022.
- [7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, pages 1254–1259, 2021.
- [9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019.
- [10] Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. Deliberate attention networks for image captioning. In *AAAI*, volume 33, pages 8320–8327, 2019.
- [11] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, pages 1–38, 2022.
- [12] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NeurIPS*, pages 11137–11147, 2019.
- [13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.
- [14] Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv e-prints*, pages arXiv–2102, 2021.
- [15] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, pages 4634–4643, 2019.
- [16] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *ICCV*, pages 2407–2415, 2015.
- [17] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, pages 10267–10276, 2020.
- [18] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, pages 499–515, 2018.
- [19] Wenhui Jiang, Minwei Zhu, Yuming Fang, Guangming Shi, Xiaowei Zhao, and Yang Liu. Visual cluster grounding for image captioning. *IEEE Transactions on Image Processing*, 2022.
- [20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.
- [21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.
- [23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.
- [24] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 375–383, 2017.
- [25] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Knowing what it is: Semantic-enhanced dual attention transformer. *IEEE Transactions on Multimedia*, 2022.
- [26] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015.
- [27] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, pages 10971–10980, 2020.
- [28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.
- [29] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *ICCV*, pages 783–792, 2021.
- [30] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024, 2017.
- [31] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL*, pages 464–468, 2018.
- [32] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global-local, spatial-channel attention for image retrieval. In *WACV*, pages 2754–2763, 2022.
- [33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [34] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022.
- [35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *NeurIPS*, 27, 2014.
- [36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.
- [38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.
- [39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, 2016.
- [40] Boyang Wan, Wenhui Jiang, Yu-Ming Fang, Minwei Zhu, Qin Li, and Yang Liu. Revisiting image captioning via maximum discrepancy competition. *Pattern Recognition*, 122:108358, 2022.
- [41] Boyang Wan, Wenhui Jiang, and Yuming Fang. Informative attention supervision for grounded video description. In *ICASSP*, pages 1955–1959, 2022.
- [42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.
- [43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.
- [44] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694, 2019.
- [45] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699, 2018.
- [46] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.
- [47] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [48] 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*, pages 15465–15474, 2021.