Transformer Based Image Captioning with Spatial Relation Representation

Zhen Wang, Yiqun Duan, Jingya Wang
zwan4121@uni.sydney.edu.au
UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia

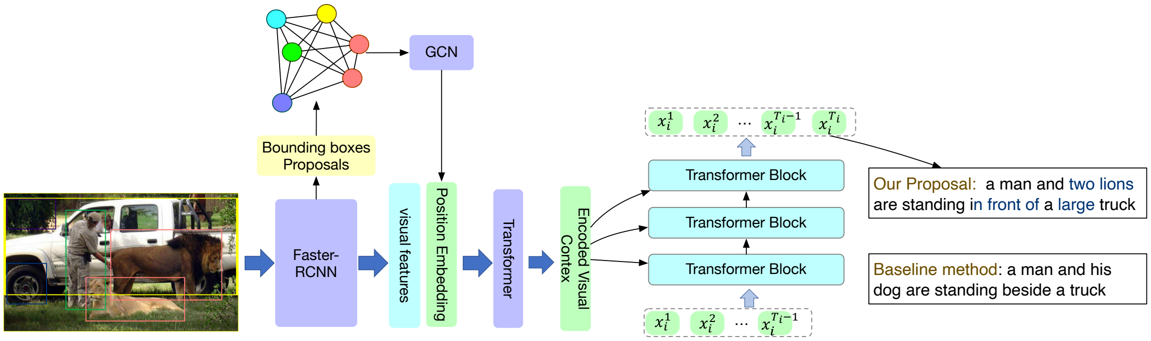

GCN
Bounding boxes Proposals
Faster-RCNN
visual features
Position Embedding
Transformer
Encoded Visual Contex
Transformer Block
Transformer Block
Transformer Block
x_i^1 x_i^2 ··· $x_i^{T_i-1}$ $x_i^{T_i}$
x_i^1 x_i^2 ··· $x_i^{T_i-1}$
Our Proposal: a man and two lions are standing in front of a large truck
Baseline method: a man and his dog are standing beside a truck

Figure 1: Overall structure of our proposed transformer-based image captioning with spatial representation. The figure also gives direct comparison of generated caption between our proposal and baseline methods.

ABSTRACT

Image captioning aims to generate a language description of a given image. This problem can be solved by learning semantic attributes of visual objects and generates captions by language model based on learned semantic attributes. However, the position relationship between extracted visual features are rarely given detailed consideration by existing methods. In this work, we propose a transformer-based encoder-decoder structure as shown in Fig. 1 for image captioning. Our proposed encoder firstly extracts both regional and global visual features then incorporates spatial information aligned to each visual feature, where bidirectional multi-head attention is used to integrates visual and spatial feature into better semantic visual context. A single directional transformer decoder is designed to decode captions based on well-designed visual context. To make a better representation of spatial information and correlation between extracted visual features, we propose and compare three subtle approaches to build position embedding, which contains spatial information for visual context. Experimental results illustrate that our proposed model achieves competitive performance on the COCO image captioning dataset compared to SOTA methods. Moreover, we give detailed ablation studies of how the position embedding will affect the image captioning tasks as well.

KEYWORDS

Image Caption, Deep Learning, Vision and Language

ACM Reference Format:
Zhen Wang, Yiqun Duan, Jingya Wang. 2020. Transformer Based Image Captioning with Spatial Relation Representation. In *Proceedings of ACM MM'20: ACM MultiMedia (ACM MM'20).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As an intersection area between vision and language, image captioning simultaneously understands visual objects of an image and generates natural language description, which is useful in multiple real-life scenarios including image-based chatbots, auto-generated social media descriptions, and image-based automatic storytelling. image captioning systems [3, 6, 34, 38, 42] mostly performs an encoder-decoder formation where the encoder extracts visual feature of an image as visual context and then put it in the decoder to generate the caption description. Here, existing methods could be both using a static, global pooled feature [34] and using a serious of spatial visual features [2]. Bottom-Up and Top-Down [2] proved that the image captioning model could reach state-of-the-art performance by using region visual features proposed by detection model (Faster-RCNN [25]).

However, the previous methods normally flatten visual feature map to embedding directly, which simultaneously lead to the loss of spatial information. How to properly represent the spatial information between these proposed visual features have not been well investigated. In this paper, we address the problem of how to represent proper spatial relations between regional features to facilitate image captioning tasks. There only exists a few papers

ACM MM'20, Oct, 2020, Seatle

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

that consider spatial relations representation for vision and language tasks. VL-BERT [29] propose to combine image regions and languages as the input of BERT to make interactions between vision and language. However, the position embedding is simply fixed with the same value for all visual regions. Although it has reached remarkable performance on VQA tasks, VL-BERT is not that suitable for caption, which needs a decoding procedure. [36] propose to divide images to blocks and concatenate regional image features with attention embedding associated with these blocks for stacked GRU [4] layers. However, the spatial relations between these visual features have not been well explored.

In this paper, we propose an all transformer-based image captioning model with position embedding as shown in Fig. 1. Transformer [31] structures have reached remarkable performance [5] on extracting relations between features. In that case, we propose to treat series of visual features as individual tokens and use a bidirectional transformer to extract better semantic features with spatial correlation between these features into a visual context as described in section 3.2.3. Then a single directional transformer is applied to generate captions based on extracted visual context as described in section 3.3. Different from NLP tasks in which word tokens have no shape, flatten all visual features into visual embedding lead to loss of spatial and shape information. In that case, we propose three approaches to construct position embedding to represent spatial relations between these visual embedding in section 3.4, where the construction methods including directly projecting coordinates to embeddings and utilizing graph convolutional network to integrates relative spatial information between visual features. The experiment results in section 4 suggest that the proposed transformer-based image captioning model (TPE-Cap) could reach competitive performance on benchmark datasets. Moreover, we conduct ablation study in section 4.4 to compare properties of proposed position embedding approaches. Experiment results suggest that all of our proposed methods could reach competitive performance as transformer has provide stronger encoder ability. This paper has three aspects of contribution as listed below:

- We propose a transformer-based image captioning model with subtle position embedding to facilitate the better performance of the image captioning model.
- We explore the properties of different between spatial information representation methods by conducting a detailed ablation study.
- Experimental results prove that our proposed image captioning model could achieve competitive performance on image captioning tasks.

2 RELATED WORK

Most modern works in image captioning presents an encoder-decoder [22, 34, 38] formation. Here, encoders extract static global pooled [34] or regional pooled [2] visual context from the input image. The decoders normally consist of basic structures for sequence modeling [11, 31] which are in charge of receiving visual context from encoders and generate caption. Attention mechanism between encoder and decoder [34, 35] has been proved of vital importance to generate high quality sentences. Recent trends including unsupervised image captioning [7], object or style controlled image captioning [8, 41], and using reinforcement learning to improve caption diversity [26].

However, these traditional methods normally flatten visual feature map to embedding directly, which simultaneously lead to the loss of spatial information. Only a few paper address the spatial relation representation problem for vision and language tasks, where [37] firstly propose to add position attention on traditional LSTM model to improve the image-text matching task. Along with the risen of transformer [32] based models such as BERT [5] and GPT-2 [24] for language modeling, Visual-BERT [16] and VL-BERT [30] firstly bring bidirectional transformer structure to vision and language area. However, both of them are not suitable for image captioning as they use bidirectional transformer structures. Moreover, both of two papers have not conducted a detailed study about how to represent the position of visual embedding properly. ORT [10] firstly introduces transformer structures into image captioning. However, ORT [10] simply add coordinates of bounding boxes into transformer block through the simple fully connected layer and has paid much attention to integrating relative spatial information between visual features. This paper firstly proposes a brand new transformer-based image captioning model and further proposes then analyzes various spatial representations for the captioning model.

3 MODEL DETAILS

In this section, we firstly give a task definition in section 3.1. Then we respectively provide details of image encoder and caption decoder in section 3.2 and section 3.3. Especially, we propose three approaches of constructing position embedding to represent spatial information in section 3.4.

3.1 Task Definition

Normally, the generalized image captioning performs a encoder-decoder [22, 34, 38] style, where image encoder extracts visual features from images and decoder generates caption according to the extracted visual features. Given training corpus $C = \{\mathbf{I}, \mathbf{x}_i\}$, where $\mathbf{I}$ denotes image with certain shape, and $\mathbf{x}_i = \mathbf{x}_i^1, \ldots, \mathbf{x}_i^{L_i}$ denotes the caption sequence with length L_i, the model is trained to learn generative probability $P_\theta(\mathbf{x}|\mathbf{I})$.

3.2 Transformer-Based Image Encoder

Attention has been proved as an efficient approach to improve the image captioning by paying different attention according to different token generating [34, 35]. However, the previous methods mostly use flattened visual features and deal with the vision and language intersection by the LSTM model, which limits the representation ability. The flatten operation may lose spatial correlation between visual features, which may simultaneously depress the performance of the caption model. Intuitively, we propose to employ an attention-based framework, transformer [31], to design the encoder-decoder structure of image captioning model, while with a good representation of both the visual feature and the spatial relations between visual features. Especially, the position embedding, which is our emphasis in this paper is separately introduced in section 1. The overall structure of our proposed model is shown in Fig. 2. The image encoder structure consists of visual feature

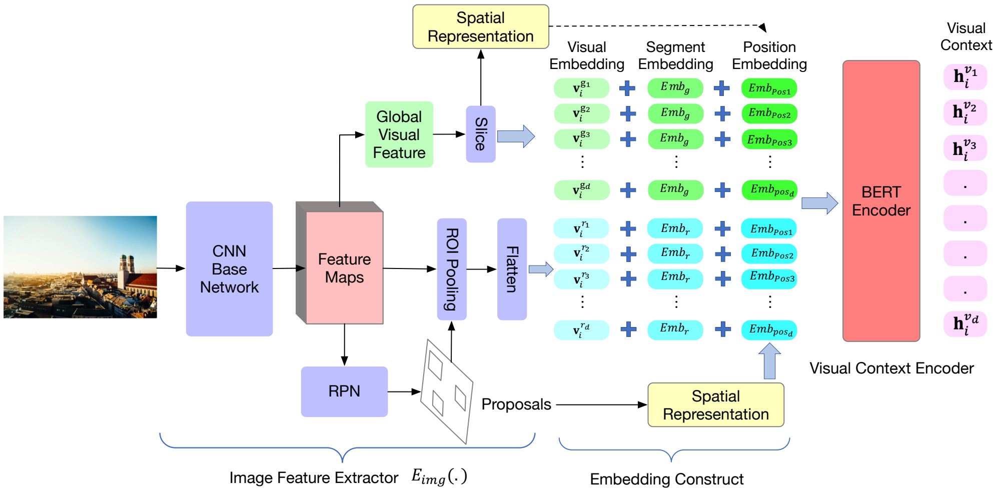

Spatial Representation
Visual Embedding
Segment Embedding
Position Embedding
Visual Context
Global Visual Feature
Slice
$\mathbf{v}_i^{g_1}$ + Emb_g + Emb_{Pos1}
$\mathbf{v}_i^{g_2}$ + Emb_g + Emb_{Pos2}
$\mathbf{v}_i^{g_3}$ + Emb_g + Emb_{Pos3}
$\mathbf{v}_i^{g_d}$ + Emb_g + Emb_{pos_d}
$\mathbf{v}_i^{r_1}$ + Emb_r + Emb_{Pos1}
$\mathbf{v}_i^{r_2}$ + Emb_r + Emb_{Pos2}
$\mathbf{v}_i^{r_3}$ + Emb_r + Emb_{Pos3}
$\mathbf{v}_i^{r_d}$ + Emb_r + Emb_{pos_d}
BERT Encoder
$\mathbf{h}_i^{v_1}$
$\mathbf{h}_i^{v_2}$
$\mathbf{h}_i^{v_3}$
$\mathbf{h}_i^{v_d}$
CNN Base Network
Feature Maps
ROI Pooling
Flatten
RPN
Proposals
Spatial Representation
Visual Context Encoder
Image Feature Extractor $E_{img}(\cdot)$
Embedding Construct

Figure 2: Image Encoder Structure. Green bars denote sliced global pooled visual embeddings while blue bars denote regional visual embeddings.

extractor, embedding construct and BERT context encoder, which will be described in detail next.

3.2.1 Visual Feature Extractor. Given an image $\mathbf{I}$, the encoder uses deep convolutional neural networks (CNN) to extract semantic features from it. Here, existing methods could be roughly categorized into two aspects, 1) using a static, global pooled representation of the image [34] or 2) using a series of spatial visual features [2]. Bottom-Up and Top-Down Attention [2] has proved that, by proposing image regions (based on Faster-RCNN [25]), each with an associated feature vector and incorporates these features through attention, it could reach sate-of-the-art performance on image captioning task [2].

Similarly, our proposed method uses the Faster-RCNN [25] with base-net ResNet101 [9] as [2] did. However, instead of only extracting regional features proposed by cropping bounding boxes [2] from Faster-RCNN, we also add global pooled visual features directly from the base-net. Since we use a transformer to encode visual context features, we propose to extract as many visual features as possible, not only the proposed objects but also background visual features, and utilize the strong ability of the transformer to acquire a better visual context feature. For regional feature, after performed non-maximum suppression, the output of Faster-RCNN is a set of regions and features. Given region i, regional pooled features $\{\mathbf{v}_i^{r_i}\}_{i=1}^{d}$ is defined as the ROI (region of interest)-pooled convolutional feature from this region (eg. with shape $(16, 16, d)$). The global visual feature $\mathbf{v}_i^g$ is calculated by the base-net in the first stage (eg. with shape $(64, 64, d)$). It should be noted that normally, global pooled features form base-net (eg. ResNet-101) is larger than regional pooled features. In that case, we slice global features to adopt the shape as the same to each of the ROI pooled regional features. All of these enhanced features will be input to the following embedding construction.

3.2.2 Embedding Construct. In order to acquire a better semantic representation of visual features, we propose to treat a set of visual features as a visual sequence with spatial relations and use transformer to learn a better visual context representation. In that case, instead of simply flatten visual feature maps for single soft attention layer to transfer visual features to semantic features [2, 34], we propose to use bidirectional transformer layer to integrate both the visual features and the relations between these features to build a better visual context for caption decoder.

We propose to use similar concepts from transformer [5] to represent these relations with embedding as it shows in Fig. 2, where **visual embedding** contains original visual features, **segmentation embedding** provides auxiliary information to distinguish global features $\mathbf{v}_i^g$ and regional features $\mathbf{v}_i^r$, and **position embedding** denotes the spatial information of corresponding visual feature. In order to adopt a set of visual features into forms of transformers [5], we directly flatten each visual feature as the embedding dimension (eg. with length 1024) and concatenate these features as tokens. Here the spatial correlations between these flatten visual features is represented by Position Embedding introduced in section 3.4. Since we directly concatenate regional visual features and global visual features, segmentation embedding is applied to distinguish two different kinds of features as it shows in Fig. 2. It should be noted that segmentation embedding is directly acquired by word2vec from the token word 'Seg-global' and 'Seg-regional'.

3.2.3 BERT Context Encoder. Transformer-based models [5, 24] have achieved remarkable performance on various sequence modeling tasks. Considering that we propose to utilize all extracted visual features with attention without decoding at this stage, we select the BERT [5] structure with two layers of transformer blocks. The constructed embedding is fed to BERT as shown in Fig. 2 by assign input of query projection, key projection and value projection, $(\mathbf{q})$, $(\mathbf{k})$, $(\mathbf{v})$ as the sum of three embeddings mentioned above. The mathematical process of final image decoder $E_{img}(\cdot)$ could be described as it in Eq. 1.

$$
\begin{aligned}
\mathbf{h}_i^v &= E_{img}(\mathbf{I}) = BERT(\mathbf{q}, \mathbf{k}, \mathbf{v}) \\
\mathbf{q} = \mathbf{k} = \mathbf{v} &= ([\mathbf{v}_i^g, \mathbf{v}_i^r] + Emb_{pos} + [Emb_g, Emb_r]),
\end{aligned} \tag{1}
$$

where global and regional visual feature $\mathbf{v}_i^g$, $\mathbf{v}_i^r$ are respectively extracted by Faster-RCNN ($FR(\cdot)$) described above.

3.3 Caption Decoder

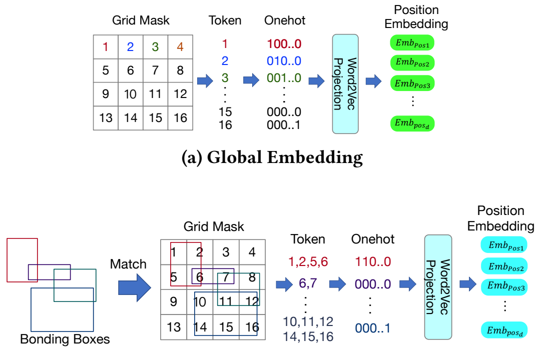
We design our caption decoder in a neural response generation formation, where we take the encoded visual feature as context, and sample word by word using single directional transformer blocks (GPT-2) [24] based on extracted context. In that case, the caption decoder is expected to learn probability $P(x^{T_i} | \{\mathbf{h}^{v_j}\}_{j=1}^{d}, x^1, x^2, ..., x^{T_i - 1})$ of sampling token x^{T_i} , where $\{\mathbf{h}^{v_j}\}_{j=1}^{d}$ denotes visual context defined in section 3.2. The model structure of the caption decoder is the same to standard transformer [31] decoder. The single directional transformer decoder receives well represented visual context through encoder-decoder attention as defined in Eq. 2.

$$
\begin{aligned}
\mathbf{h}^{dec_j} &= MH(\mathbf{h}^v, \mathbf{h}^{dec_{j-1}}, \mathbf{h}^{dec_{j-1}}) \\
\mathbf{h}^{dec_0} &= MH(\mathbf{k}, \mathbf{q}, \mathbf{v}), \mathbf{k} = \mathbf{v} = [x^1, x^2, \ldots, x^{T_i - 1}], \quad \mathbf{q} = \mathbf{h}^v
\end{aligned} \tag{2}
$$

where $MH(\cdot)$ denotes multi-head attention block introduced in transformer [31]. Besides first decoding block, query $\mathbf{q}$ of each decoding transformer is the visual context $\mathbf{h}^v$. Then the current token x^{T_i} is sampled through equation:

$$
x^{T_i} \sim p^{T_i} = softmax(\mathbf{W}\mathbf{h}^{dec_j}), \tag{3}
$$

where $\mathbf{W}$ denotes softmax parameters and x^{T_i} is sampled according to probability distribution p^{T_i}.



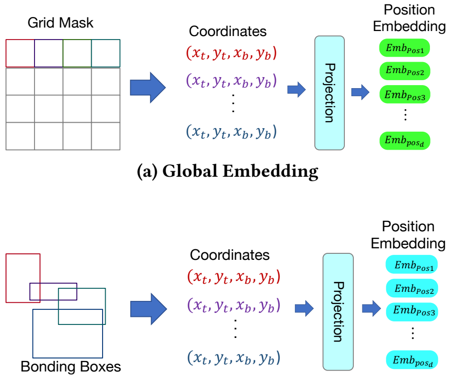
(a) Global Embedding

(b) Regional Embedding

Figure 3: Learned Position Embedding with Grid Mask. Here, sub-figure (a) and (b) respectively denotes schematic diagram of generation process for global and regional features.

3.4 Spatial Representation

Since we have proposed to treat visual features as a sequence of embeddings and use position embedding respectively assigned to each visual embedding to represent the spatial relations between these features, how to acquire a better position embedding becomes a crucial task. In this section, we propose three different position embedding approaches and compare the differences between these three approaches as following.



(a) Global Embedding

(b) Regional Embedding

Figure 4: Learned Position Embedding with Bounding Boxes. Here, sub-figure (a) and (b) respectively denotes schematic diagram of generation process for global and regional features. Especially, for global position embedding, we directly use the coordinates of grid masks as bounding boxes.

3.4.1 *Learned Position Embedding from Grid Mask.*

Position embedding in BERT for language modeling could be learned from discrete position tokens. Naturally, it is feasible to learn a proper embedding by assign position tokens to each of the visual embeddings and use the embedding layer to transfer these discrete tokens to high dimensional embeddings. In that case, we firstly propose to manually set a grid mask on an image, and assign each grid anchor as token id to visual features as it shows in Fig. 3.

Since there is a strong spatial relation between these manually selected tokens, if there is enough training data, it's feasible to learn a good spatial representation. Similar to word to vector projection in NLP area, we feed the one-hot formation of token id to a linear transformation to acquire position embedding. For global visual feature, we directly assign visual tokens because it is sliced from the global feature map according to the grid mask. For regional visual features, we match the bounding box proposed by Faster-RCNN with the proposed mask grid. It should be noted that, since the bounding box of regional features might simultaneously appear in more than one position token id, the one-hot representation may suggest existing of more than one token id.

3.4.2 *Learned Position Embedding from Bounding Boxes.*

The position embedding learned from the grid mask defined above relies on a manually selected grid mask. However, since we use Faster-RCNN as our basic feature extractors, we naturally could get bounding boxes of each regional feature directly. Compared to discrete token ids, the continuous bounding boxes are naturally containing spatial information. We propose to use a linear layer with a designed normalization to project bounding box coordinates to high dimensional embedding (same to visual embedding size) as it shows in Fig. 4.

The mathematical process of calculating position embedding for given visual embedding could be defined as:

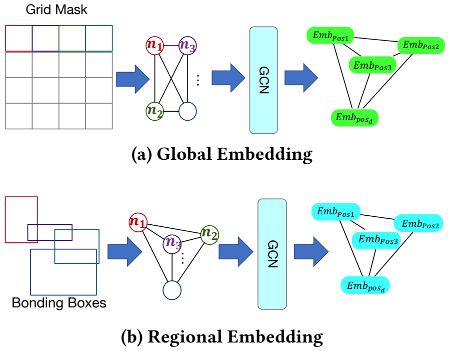
$$Emb_{pos_i} = tanh(\mathbf{W}[x_t, y_t, x_b, y_b]), \tag{4}$$

where x_t, y_t and x_b, y_b are respectively the coordinates of top left and bottom right points of the bounding boxes. Here, $tanh(\cdot)$ denotes Tangent activation function, where the function $tanh(x) = \frac{2}{1+e^{2x}} - 1$ limits the output range within range $(-1, 1)$. Projected embedding with both negative and positive is easier to learn through training procedure [28]. For global features that do not have corresponding bounding boxes, we directly use the coordinates of the slicing mask as it shows in Fig. 4.

3.4.3 Position Embedding from Spatial Graph. The two methods proposed above majorly concentrate on simple construct projection between input visual feature location (coordinates) information and the position embedding. However, merely learn position embedding from separated coordinates has not taken the full utilization of spatial information implied in these visual features. We propose to construct the position embedding while considering the relationship between objects. Specifically, we use graph convolutional networks (GCN) [17] to calculate the position embedding based on information of all objects. The main approach is as shown in Fig. 5, where each proposed bounding boxes by Faster-RCNN is treated as a node with spatial feature $\mathbf{n}_i$ and the GCN integrates all relative spatial informations from each node and calculate the position embedding. We construct the spatial feature $\mathbf{n}_i$ as vector $[x_t, y_t, x_b, y_b]$, which consists coordinates information indicates the shape and relative position of each node. The mathematical process of one single GCN layer could be defined as :

$$\begin{aligned} Emb_{posi} &= \sigma(\sum_{j \in N_i} \frac{1}{C_{ij}} \mathbf{n}_j \mathbf{W} + \mathbf{n}_i \mathbf{W} + b^i) \\ \frac{1}{C_{ij}} &= \frac{1}{\sqrt{deg(\mathbf{n}_i)}\sqrt{deg(\mathbf{n}_j)}}, \end{aligned} \tag{5}$$

where $\frac{1}{C_{ij}}$ is the normalization term in degree wise, N_i denotes the set adjacent nodes of $\mathbf{n}_i$, $\mathbf{W}$ denotes the trainable weights of the proposed GCN layer. By introducing GCN layer, we could calculate the position embedding not merely based on coordinate of the corresponding node but also the spatial relationship between all the proposed nodes.



(a) Global Embedding

(b) Regional Embedding

Figure 5: Learned Position Embedding from GCN

4 EXPERIMENT

In this section, we provide experimental results to illustrate the efficiency of our model. In section 4.1 and section 4.2, we respectively give the description of the datasets and implementation details. In section 4.3, we compare the performance of our proposal with previous approaches. And we conduct an ablation study in section 4.4 to explore how position embedding will effect image caption.

4.1 Datasets

Similar to Bottom-Up paper [2], we pre-train our model on Visual Genome Dataset [14], which contains 108K image samples with densely annotated objects, attributes, relationships, and 1.7M visual questions. However, for pre-training, we only use the annotated objects and attributes label. The Visual Genome dataset is divided into 90%, 5% and 5% respectively for training, validation and testing. Please note that since Visual Genome dataset has intersection images with COCO dataset [19], we avoid images from COCO validation and COCO testing in our training data. The COCO 2014 captions dataset [19] is used to test the performance of our model. We use the 'Karpathy' splits [12] to separate our training and testing/validation set for a fair comparison, where we have 11.3K training images with five captions each. A standard text processing is implemented for this experiment, where the captions are set to lower case, then tokenized sequence of tokens with 10K vocabularies.

4.2 Implementation Details

For basic image extractor, we use a similar Faster-RCNN model as mentioned in Bottom-Up and Top-Down [2], where the regional proposal threshold and classification threshold are respectively set to 0.7 and 0.3. However, since the BERT context encoder requires a fixed-length input, we set the maximum number of both regional proposals and global proposals as 16, which simultaneously means the global pooled feature map is sliced by a 4 × 4 grid mask. Each of the visual embedding (regional and global) is pooled to the same size 512. For BERT visual encoder, we set embedding size as 512 for visual, position and segmentation embedding while the hidden feature dimension is set to 512 with 2 layers of standard transformer blocks [31]. For transformer decoder, we set the hidden dimension to 512 with 3 layers of transformer blocks. For pre-training on Visual Genome, we use the Adam [13] optimizer with linear learning rate decay from 1e-3 to 1e-5. For the training procedure on COCO dataset, we use Adam optimizer with linear learning rate decay from 2e-4 to 1e-5. All model is trained on 2 Nvidia V100 GPUs.

4.3 Evaluation Metrics

In this section, we compare our proposed methods with traditional approaches as it shows in Table 1 on COCO dataset [19]. Since our image captioning proposal use transformer-based structures for both image encoder and caption generator with position embedding, we use TPE-Cap (Transformer with Position Embedding) as the abbreviation of our proposal. We take 3 widely known previous approaches as the compared approaches including Show-Tell

Table 1: Automatic evaluation metrics on COCO dataset. Here, TPE-Cap* denotes our model with proposed position embedding learned from grid mask while TPE-Cap^o denotes our model with position embedding learned from bounding boxes. TPE-Cap† denotes the position embedding learned from GCN.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
ShowTell [34]	70.4	54.7	38.9	29.8	23.8	50.7	86.3	17.2
Adaptive [21]	74.2	58.3	42.7	32.5	26.1	52.6	108.5	19.5
Att2all [27]	-	-	-	34.2	26.7	55.7	114.0	-
Up-Down [2]	79.8	62.6	47.1	36.3	27.2	56.3	117.6	21.4
Visual-policy [20]	80.4	62.9	48.5	38.6	28.3	58.5	126.3	21.6
GCN-LSTM [39]	80.5	63.1	48.2	38.2	28.5	58.3	127.6	22.0
ORT [10]	80.5	63.0	48.7	38.6	28.7	58.4	128.3	22.6
TPE-Cap*	79.7	62.6	48.4	38.2	28.2	57.9	126.1	21.5
TPE-Cap^o	80.2	62.9	48.4	38.3	28.4	58.1	126.7	21.9
TPE-Cap†	**80.9**	**63.2**	**48.9**	**38.7**	28.6	**58.6**	127.4	22.4

Table 2: Ablation study of different position embedding representation

Spatial Info.	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Without Position	35.5	27.2	55.7	120.9	21.4
Grid Mask	38.2	28.2	57.9	126.2	21.5
Bounding Box	38.3	28.4	58.1	126.7	21.9
GCN	38.7	28.6	58.6	127.4	22.4

[34], Adaptive [21], Bottom-Up, Top-Down [2]. Here, 1) Show-Tell [34] denotes baseline with LSTM generator, global pooled visual feature and basic soft attention model, 2) Adaptive [21] denotes baseline with LSTM generator and attention alignment between words and visual regions, 3) Up-Down [2] denotes baseline with regional visual features. Besides, ORT [10] provides comparison baseline with transformer-based methods while GCN-LSTM [39] provides comparison baseline which also utilize GCN into image captioning task. Widely used automatic evaluation metrics, BLEU [23], METEOR [15], ROUGE-L [18], CIDEr [33] and SPICE [1] used to measure relevance between generated caption and ground truth as shown in Table 1.

It is observed that, the proposed transformer-based image captioning model could reach competitive performance on standard image captioning tasks. Both three kinds of position embedding methods in section 3.4 could reach competitive results on auto-metrics, which illustrate the efficiency of the proposed transformer-based image captioning model. Our proposed TPE-Cap model with position embedding constructed by GCN achieve our beset performance, where TPE-Cap† outperforms the previous state of the art (SOTA) model on metrics including BLUE-14̃ and ROUGE-L while keeping METEOR, CIDEr and SPICE the same high as previous SOTA approaches. Position embedding learned from bounding boxes performs similarly to previous SOTA model while position embedding learned form grid mask performs slightly lower than the baseline model (around 1%). In that case, we argue that the proposed transformer-caption model could achieve competitive performance on image captioning task. It is rational because transformer-based models should perform better at capturing compared to LSTM-based models. However, the different methods of dealing spatial information may lead to final performance differences. We moreover deploy ablation study in section 4.4 to reveal effects bring by position embeddings.

Table 3: Auto-Metric decreasing by reducing training data.

Data	B-2	B-4	METEOR	ROUGE-L	CIDEr	SPICE
10%	8.2	6.7	8.0	9.4	12.7	8.2
20%	6.1	4.9	6.3	7.1	10.5	5.9
50%	4.3	4.2	5.8	5.3	8.3	4.1
70%	2.4	1.9	2.9	2.4	4.7	1.8
100%	0.6	0.5	0.4	0.7	1.2	0.9

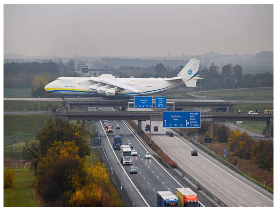
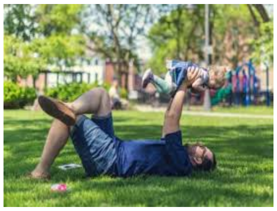
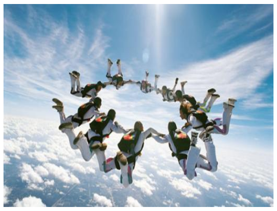
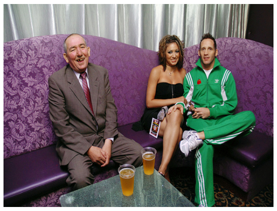
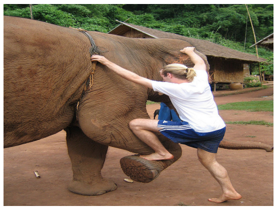
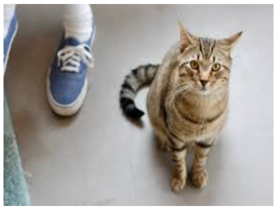
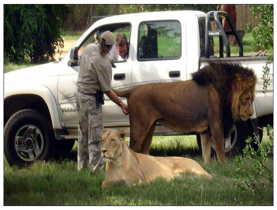
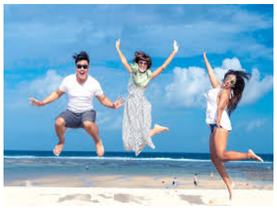
4.4 Ablation Study on Spatial Representation

4.4.1 Different Position Embedding Representation. To reveal impact bring by position embedding, we firstly compare the evaluation metric mentioned in section 4.3 between our proposal with or without position embedding as it shows in Table 2. It is observed that the proposed transformer model without position embedding performs significantly lower than those with position embedding, which suggests that align each proposed visual feature with spatial information could improve the final performance of image captioning. This observation supports the effectiveness of position embedding proposed in section 3.2.2, i.e., without position embedding, the transformer would loss the ability of representing the spatial relationships between visual embeddings.

According to Table 2, GCN based spatial representation significantly outperforms other approaches, where the results acquired by GCN is It is rational because the by introducing GCN, the model could integrates all positions/shapes from all proposed visual features rather than merely use a simple projection to transfer coordinates information to position embedding. The relative spatial information representation is also closer to how people really see in real life.

4.4.2 Performance Under Various Training Data Volume. The position embedding learned from bounding boxes outperforms which learned from the tokenized mask grid. Since both approaches contain position coordinates corresponding to each visual embedding,

Table 4: Results Comparison The Proposed Spatial Relation Transformer

			
Standard: a plane is on the top of a street	a man lying on the grass with a Frisbee	a group of people on a ski lift	a group of people sitting on a coach.
Ours: a large plane is taking off from a run way	a man lying on the grass with a kid in his hand	a group of people jumping in the sky	a group of people sitting next to each other
			
Standard: a man is standing next to an elephant	a cat is sitting by a pair of shoes	a man and his dog are standing on the back of a truck	a group of women jumping on a beach
Ours: a man is standing next to a large elephant	a cat is sitting on the ground next to a pair of shoes	a man and two lions are standing in front of a large truck	a group of people jumping on the top of a beach

we think that the performance difference reals that the continuous coordinates are easier to learn than discrete grid indices. To prove our assumption, we further compare the performance difference between the two position embedding methods under different dataset scale. We respectively use 10%, 20%, 50%, 70%, 100% samples from the training split for training and use the same evaluation dataset to evaluate, then report the performance difference in Table 3.

It is observed that the difference between position embedding learned from bounding boxes and from grid mask increases while we reduce the training sample. This phenomenon reveals that the proposed transformer-based image captioning model is easier to learn spatial relations from bounding boxes than grid mask tokens. We argue this observation is rational because (1) coordinates of bounding boxes have already implicitly contained relative positions of each visual proposal, even just a affine transformation could maintain the spatial information, which is easier to learn (2) coordinates is continuous in numerical-wise, which is easier to be optimized during training process. In that case, projection from continuous bounding box coordinates to position embedding is easier to learn compared to projection from discrete tokens.

4.4.3 *Case Study of Image Captioning Performance.*

In order to directly show the performance of our proposed model, we show generated captions in Table 4. We found that fine-grained learned position representation which implicitly contain spatial information of each visual embedding could facilitate the image captioning model describing the objects in the image, especially when those objects contains spatial relations. To illustrate our model performance, we use blue color to denote caption words that contain position or shape information as it shows in Table 4.

4.4.4 *Graph Density of GCN Spatial Representation.*

While constructing the graph for GCN to calculate position embedding, we consider each pair of the nodes in the graph has an edge, which means we use a complete graph to calculate the position embedding. However, previous GCN papers also figure out that while constructing scenery graphs, it's possible to only connect two nodes while the pair fit into certain condition [40]. In that case, we conduct ablation study of different graph density, where we define the density of the graph as the ratio α between the number of current graph edges and the number of a complete graph edges. We control the density of the scenery graph by control the number of edges connected to each node. For example, we connect every other nodes for a graph of density 1, while we only connect top α% closest nodes for graph with density α, where we rank the distance of each nodes pair by the center of the bounding boxes.

Table 5: Impact of graph density

Density	B-2	B-4	METEOR	ROUGE-L	CIDEr	SPICE
1	63.2	38.7	28.6	58.6	127.4	22.4
0.75	63.1	38.5	28.6	58.6	127.4	22.3
0.5	63.0	38.5	28.5	58.4	127.1	22.0
0.25	63.0	38.4	28.4	58.2	126.9	21.9
0	62.9	38.3	28.4	58.1	126.7	21.9

It is observed from experimental results shown in Table 5 that the density of the scenery graph only has a negligible impact on the final evaluation metrics. Yet, the graph with higher density still slightly better than lower density. We speculate that the reason why of this observation is that the caption in COCO dataset normally is compared short (around 10-20 words), which may not contains enough spatial relation description. In that case, the final metrics won't show much differences on COCO caption dataset. However, we did observe that complete graph (density 1) outperforms the graph with no connections (in fact, the graph with no connection means directly project bounding boxes coordinates to embedding, which is the same as second position embedding defined in section 3.4.2). We think this is also rational because the position representation which contains relative spatial information from other objects absolutely contains more information than those merely has coordinates information.

5 CONCLUSION & FUTURE WORK

In this paper, we present a novel transformer-based image captioning model and the corresponding multi-stage training framework to solve cross-domain caption style synthesis problems. We explore and exploit the position embedding in the visual context, which properly represents spatial relation between visual embeddings to facilitate the image captioning tasks. Experiment results show the efficiency of our proposed methods. Forcing the image captioning model to simultaneously present multiple styles cross-domain is still a hard question. We will explore how to generate more detailed or object-oriented captions in future work. Also, in this paper, we only use a simple position embedding for transformers. However, we think how to present spatial relations to the language model properly is a promising task, too.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9909. Springer, 382–398. https://doi.org/10.1007/978-3-319-46454-1_24

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

[3] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. 2019. Improving Image Captioning with Conditional Generative Adversarial Nets. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 8142–8150.

[4] Junyoung Chung, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014). arXiv:1412.3555 http://arxiv.org/abs/1412.3555

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186.

[6] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2017. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 677–691.

[7] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised Image Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 4125–4134. https://doi.org/10.1109/CVPR.2019.00425

[8] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. MSCap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4204–4213.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778.

[10] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 11137–11147.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 664–676. https://doi.org/10.1109/TPAMI.2016.2598339

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[14] 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. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[15] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, Chris Callison-Burch, Philipp Koehn, Cameron S. Fordyce, and Christof Monz (Eds.). Association for Computational Linguistics, 228–231.

[16] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:cs.CV/1908.03557

[17] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 3538–3545. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16098

[18] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8693. Springer, 740–755.

[20] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-Aware Visual Policy Network for Sequence-Level Image Captioning. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. 1416–1424.

[21] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 3242–3250. https://doi.org/10.1109/CVPR.2017.345

[22] Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating Image Descriptions with Sentiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 3574–3580.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318.

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans.*

Pattern Anal. Mach. Intell. 39, 6 (2017), 1137–1149.

[26] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 1179–1195. https://doi.org/10.1109/CVPR.2017.131

[27] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Vivek Kumar Rangarajan Sridhar. 2015. Unsupervised Text Normalization Using Distributed Representations of Words and Phrases. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, Phil Blunsom, Shay B. Cohen, Paramveer S. Dhillon, and Percy Liang (Eds.). The Association for Computational Linguistics, 8–16. https://doi.org/10.3115/v1/w15-1502

[29] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *CoRR* abs/1908.08530 (2019). arXiv:1908.08530 http://arxiv.org/abs/1908.08530

[30] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. arXiv:cs.CV/1908.08530

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[33] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 4566–4575.

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[35] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 652–663.

[36] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position Focused Attention Network for Image-Text Matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 3792–3798. https://doi.org/10.24963/ijcai.2019/526

[37] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748* (2019).

[38] Linjie Yang, Kevin D. Tang, Jianchao Yang, and Li-Jia Li. 2017. Dense Captioning with Joint Inference and Visual Context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 1978–1987.

[39] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *The European Conference on Computer Vision (ECCV)*.

[40] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11218. Springer, 711–727. https://doi.org/10.1007/978-3-030-01264-9_42

[41] Yue Zheng, Yali Li, and Shengjin Wang. 2019. Intention Oriented Image Captions With Guiding Objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 8395–8404. https://doi.org/10.1109/CVPR.2019.00859

[42] Luowei Zhou, Chenliang Xu, Parker A. Koch, and Jason J. Corso. 2016. Image Caption Generation with Text-Conditional Semantic Attention. *CoRR* abs/1606.04621 (2016). arXiv:1606.04621