Text-to-Text Pre-Training with Paraphrasing for Improving Transformer-based Image Captioning

Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, Shota Orihashi
NTT Computer & Data Science Laboratories, NTT Corporation
ryo.masumura@ntt.com

***Abstract*—In this paper, we propose a novel training method for the transformer encoder-decoder based image captioning, which directly generates a captioning text from an input image. In general, many image-to-text paired datasets need to be prepared for robust image captioning, but such datasets cannot be collected in practical cases. Our key idea for mitigating the data preparation cost is to utilize text-to-text paraphrasing modeling, i.e., a task to convert an input text into different expressions without changing the meaning. In fact, paraphrasing deals with a similar transformation task to image captioning even though paraphrasing tasks have to handle texts instead of images. In our proposed method, an encoder-decoder network trained via the paraphrasing task is directly leveraged for image captioning. Thus, an encoder-decoder network pre-trained by a text-to-text transformation task is transferred into an image-to-text transformation task even though a different modal must be handled in the encoder network. Our experiments using the MS COCO caption datasets demonstrate the effectiveness of the proposed method.**

***Index Terms*—image captioning, transformer encoder-decoder, paraphrasing, pre-training**

I. Introduction

Image captioning is a task that generates a captioning text to explain content from an input image. Researchers have actively studied image captioning as a technical area linked to symbol grounding. In particular, since the advent of deep learning, neural image captioning has successfully learned image-to-text transformation in an end-to-end manner using neural networks [1], [2].

For neural image captioning, various modeling methods have been proposed. One promising approach is to utilize object regions found by object detectors such as faster regions with convolutional neural networks (Faster R-CNN) [3] as the inputs [4]. The object regions are converted into text mainly by using attention-based encoder-decoders [5]–[7], which are widely used in various natural language generation tasks. For the encoder-decoder networks, initial studies mainly introduced recurrent neural networks [4]. In addition, recent studies have used the transformer, which achieves more powerful encoder-decoder modeling [8]–[10].

Modeling neural image captioning generally requires a large amount of image-to-text paired data. In particular, multiple captions need to be annotated for each image because image captioning is a one-to-many mapping problem. However, collecting a large amount of paired data is difficult because such annotations are very costly. Therefore, a method for building accurate modeling is required even from limited image-to-text paired data. To mitigate the data scarcity problem, we focus on a paraphrasing task, which converts the input text into different expressions without changing the meaning. In a previous study, the paraphrasing task was utilized for the post-processing of image captioning to improve diversity of captions [11]. We consider that paraphrasing deals with a similar transformation task to image captioning using encoder-decoder networks [12]–[15] except that the encoder in a paraphrasing model handles texts instead of images. Therefore, we can expect that the performance of image captioning can be improved by using paraphrasing training datasets.

In this paper, we propose a novel training method for state-of-the-art transformer encoder-decoder based neural image captioning. Our key idea is to share one transformer encoder-decoder network between paraphrasing and image captioning and to train the network using both paraphrasing datasets and image captioning datasets. In our proposed method, a transformer encoder-decoder trained via the paraphrasing task is directly leveraged for transformer encoder-decoder based image captioning. In other words, encoder-decoder networks pre-trained using text-to-text transformation tasks are transferred into image-to-text transformation tasks even though the individual inputs are clearly inconsistent. We expect that the network structure obtained by the paraphrasing task will be directly useful for image captioning because both image captioning and paraphrasing tasks are done to convert the input into different expressions without changing the meaning of the input. Previously, pre-training of image-to-text encoder-decoder networks has been examined using large-scale image-to-text paired datasets [16]–[20] (see Section 2), but these methods require a large amount of image-to-text paired data. To the best of our knowledge, this is the first study to utilize text-to-text encoder-decoder networks for enhancing image-to-text ones.

In our experiments, we use the MS COCO caption datasets [21]. In our experimental setups, we split the datasets into image captioning datasets and paraphrasing datasets. Thus, we use the MS COCO caption datasets for not only evaluating image captioning but also building paraphrasing models as seen in previous work [12], [13]. We show that our proposed pre-training yields higher captioning performance than no pre-training. We also show an ablation study in which either a pre-trained encoder network or pre-trained decoder network is only utilized for improving the image captioning task. Furthermore, we qualitatively analyze the experimental results.

II. Related Work

Data scarcity in image captioning: Many image-to-text paired datasets need to be prepared for robust image captioning, but such datasets cannot be collected in practical cases. To mitigate the data scarcity problem, several methods have been examined. Semi-supervised learning methods that utilize not only image-to-text paired datasets but also unpaired image datasets and unpaired text datasets are one main solution [22], [23]. In addition, unsupervised learning methods that only leverage unpaired datasets have been investigated [24], [25]. Different from these studies, our proposed method utilizes text-to-text paired datasets collected from a paraphrasing task for mitigating the data scarcity problem in image captioning.

Pre-training for image captioning: For pre-training of image captioning networks, vision-language pre-training has attracted much attention. Most modern studies focus on jointly embedding texts and images in the same continuous space by using image-to-text paired datasets [16]–[20]. The pre-trained networks are usually used for image captioning, video captioning, and visual question answering. These previous methods required image-to-text paired datasets to be prepared for pre-training. In contrast, our proposed pre-training uses text-to-text paired datasets to improve image-to-text transformation tasks, i.e., image captioning. Our key contribution is to utilize pre-trained text-to-text encoder-decoder networks for improving image-to-text ones even though the individual inputs are clearly inconsistent.

III. Proposed Method

This section details our proposed pre-training and fine-tuning methods for neural image captioning. Our key idea is to share one transformer encoder-decoder network between paraphrasing and image captioning and to train the network using both paraphrasing datasets and image captioning datasets.

Our objective is to construct an image captioning encoder-decoder network from both paraphrasing datasets and image captioning datasets. We first define training datasets for paraphrasing as

$$D_{\text{para}} = \{\{\boldsymbol{W}_1^n, \cdots, \boldsymbol{W}_{K^n}^n\} \mid n \in \{1, \cdots, N\}\}, \quad (1)$$

where $\boldsymbol{W}_k^n = \{w_{k,1}^n, \cdots, w_{k,T_k^n}^n\}$ is the k-th sentence (token sequence) in the n-th paraphrase set. N represents the number of paraphrase sets, K^n represents the number of sentences in the n-th paraphrase set, and T_k^n represents the number of tokens in the k-th sentences in the n-th paraphrase set.

Next, we define training datasets for image captioning as

$$D_{\text{cap}} = \{\{\boldsymbol{C}^m, \{\boldsymbol{W}_1^m, \cdots, \boldsymbol{W}_{K^m}^m\}\} \mid m \in \{1, \cdots, M\}\}, \quad (2)$$

where $\boldsymbol{C}^m$ is the m-th red-green-blue image data, and $\boldsymbol{W}_k^m$ is the k-th captioning text for the m-th image. M represents the number of images, and K^m represents number of captioning text for the m-th image. Note that data collections for the paraphrasing datasets are not the same as those for the image captioning datasets[1]. In our proposed pre-training, both datasets are used for building an image captioning model.

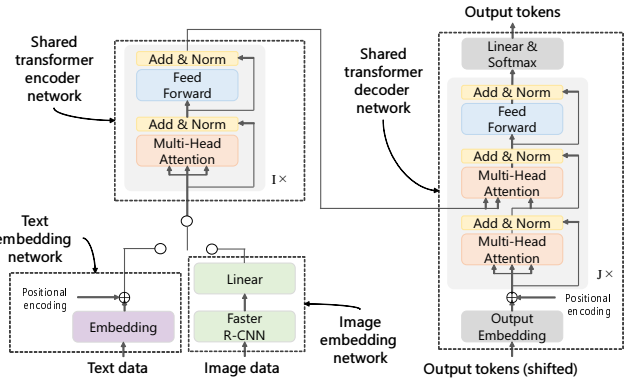

Fig. 1. Joint transformer encoder-decoder network for modeling both paraphrasing and image captioning.

A. Modeling

Paraphrasing Model: In a paraphrasing task, input text is converted into different expressions without changing the meaning. In this work, we use the transformer encoder-decoder based autoregressive paraphrasing model. It predicts the generation probability of a paraphrased text $\boldsymbol{W} = \{w_1, \cdots, w_T\}$ given an input text $\bar{\boldsymbol{W}} = \{\bar{w}_1, \cdots, \bar{w}_L\}$, where w_t is the t-th token in the paraphrased text, and $\bar{w}_l$ is the l-th token in the input text [13]. In the autoregressive paraphrasing model, the generation probability of $\boldsymbol{W}$ is defined as

$$P(\boldsymbol{W}|\bar{\boldsymbol{W}}; \boldsymbol{\Theta}_{\text{t2t}}) = \prod_{t=1}^{T} P(w_t|w_{1:t-1}, \bar{\boldsymbol{W}}; \boldsymbol{\Theta}_{\text{t2t}}), \quad (3)$$

where $\boldsymbol{\Theta}_{\text{t2t}} = \{\boldsymbol{\theta}_{\text{text}}, \boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}\}$ represents the trainable model parameter sets, and $w_{1:t-1} = \{w_1, \cdots, w_{t-1}\}$. In this work, $P(w_t|w_{1:t-1}, \bar{\boldsymbol{W}}; \boldsymbol{\Theta}_{\text{t2t}})$ is computed using a text embedding network $\boldsymbol{\theta}_{\text{text}}$, a shared transformer encoder network $\boldsymbol{\theta}_{\text{enc}}$, and a shared transformer decoder network $\boldsymbol{\theta}_{\text{dec}}$.

Image Captioning Model: In an image captioning task, a captioning text to explain content is generated from an input image. In this work, we use the transformer encoder-decoder based autoregressive image captioning model. It predicts the generative probability of a captioning text $\boldsymbol{W}$ given input image data $\boldsymbol{C}$ [8]. In the autoregressive image captioning model, the generation probability of $\boldsymbol{W}$ is defined as

$$P(\boldsymbol{W}|\boldsymbol{C}; \boldsymbol{\Theta}_{\text{i2t}}) = \prod_{t=1}^{T} P(w_t|w_{1:t-1}, \boldsymbol{C}; \boldsymbol{\Theta}_{\text{i2t}}), \quad (4)$$

where $\boldsymbol{\Theta}_{\text{i2t}} = \{\boldsymbol{\theta}_{\text{image}}, \boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}\}$ represents the trainable model parameter sets. In this work, $P(w_t|w_{1:t-1}\ \boldsymbol{C}; \boldsymbol{\Theta}_{\text{i2t}})$ is computed using an image embedding network $\boldsymbol{\theta}_{\text{image}}$, a shared transformer encoder network $\boldsymbol{\theta}_{\text{enc}}$, and a shared transformer decoder network $\boldsymbol{\theta}_{\text{dec}}$.

[1]In our experiments described in Section 4, we split MS COCO datasets [21] into datasets for building image captioning and those for building paraphrasing to produce this situation.

B. Joint Network

We construct one joint transformer encoder-decoder network for modeling both paraphrasing and image captioning. Figure 2 shows the joint network structure of our method. It is composed of a text embedding network, an image encoder network, a shared transformer encoder network, and a shared transformer decoder network.

Text embedding network: In the text embedding network that is only used for the paraphrasing model, an input text $\bar{\boldsymbol{W}}$ is converted into continuous vectors $\boldsymbol{Q} = \{\boldsymbol{q}_1, \cdots, \boldsymbol{q}_L\}$ as

$$\boldsymbol{q}_l = \texttt{AddPosEnc}(\bar{\boldsymbol{w}}_l), \quad (5)$$

$$\bar{\boldsymbol{w}}_l = \texttt{Embedding}(\bar{w}_l; \boldsymbol{\theta}_{\text{text}}), \quad (6)$$

where $\texttt{AddPosEnc}()$ is a function that adds a continuous vector in which position information is embedded. $\texttt{Embedding}()$ is a linear layer that inserts the input token in a continuous vector.

Image embedding network: In the image embedding network that is only used for the image captioning, input image data $\boldsymbol{C}$ is converted into continuous vectors $\boldsymbol{H} = \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_D\}$ as

$$\boldsymbol{H} = \texttt{Linear}(\boldsymbol{R}; \boldsymbol{\theta}_{\text{image}}), \quad (7)$$

$$\boldsymbol{R} = \texttt{FasterRCNN}(\boldsymbol{C}; \boldsymbol{\theta}_{\text{image}}), \quad (8)$$

where $\texttt{FasterRCNN}()$ is a function that converts an image into object-wise continuous vectors on the basis of Faster R-CNNs [3], and $\texttt{Linear}()$ is a linear transformational function. D is the number of objects detected by the Faster R-CNN.

Shared transformer encoder network: The transformer encoder network converts input continuous vectors into hidden representations $\boldsymbol{S}^{(I)}$ using I transformer encoder blocks. The i-th transformer encoder block composes the i-th hidden representations $\boldsymbol{S}^{(i)}$ from the lower layer inputs $\boldsymbol{S}^{(i-1)}$ as

$$\boldsymbol{S}^{(i)} = \texttt{TransformerEnc}(\boldsymbol{S}^{(i-1)}; \boldsymbol{\theta}_{\text{enc}}), \quad (9)$$

where $\texttt{TransformerEnc}()$ is a transformer encoder block that consists of a scaled dot product multi-head self-attention layer and a position-wise feed-forward network [26]. The hidden representations $\boldsymbol{S}^{(0)}$ are defined as

$$\boldsymbol{S}^{(0)} = \begin{cases} \boldsymbol{Q} & \text{if the input is text} \\ \boldsymbol{H} & \text{if the input is image.} \end{cases} \quad (10)$$

Shared transformer decoder network: The transformer decoder network computes the generation probability of a token from the preceding tokens and the hidden representations generated in the transformer encoder. The predicted probabilities of the t-th token w_t are calculated as

$$P(w_t | w_{1:t-1}, \boldsymbol{O}) = \texttt{Softmax}(\boldsymbol{u}_{t-1}^{(J)}; \boldsymbol{\theta}_{\text{dec}}), \quad (11)$$

$$\boldsymbol{O} = \begin{cases} \bar{\boldsymbol{W}} & \text{if the input is text} \\ \boldsymbol{C} & \text{if the input is image,} \end{cases} \quad (12)$$

where $\texttt{Softmax}()$ is a softmax layer with a linear transformation. The input hidden vector $\boldsymbol{u}_{t-1}^{(J)}$ is computed from J transformer decoder blocks. The j-th transformer decoder block composes the j-th hidden representation $\boldsymbol{u}_{t-1}^{(j)}$ from the lower layer inputs $\boldsymbol{U}_{1:t-1}^{(j-1)} = \{\boldsymbol{u}_1^{(j-1)}, \cdots, \boldsymbol{u}_{t-1}^{(j-1)}\}$ as

$$\boldsymbol{u}_{t-1}^{(j)} = \texttt{TransformerDec}(\boldsymbol{U}_{1:t-1}^{(j-1)}, \boldsymbol{S}^{(I)}; \boldsymbol{\theta}_{\text{dec}}), \quad (13)$$

where $\texttt{TransformerDec}()$ is a transformer decoder block that consists of a scaled dot product multi-head masked self-attention layer, a scaled dot product multi-head source-target attention layer, and a position-wise feed-forward network. The hidden representations $\boldsymbol{U}_{1:t-1}^{(0)} = \{\boldsymbol{u}_1^{(0)}, \cdots, \boldsymbol{u}_{t-1}^{(0)}\}$ are produced by

$$\boldsymbol{u}_{t-1}^{(0)} = \texttt{AddPosEnc}(\boldsymbol{w}_{t-1}), \quad (14)$$

$$\boldsymbol{w}_{t-1} = \texttt{Embedding}(w_{t-1}; \boldsymbol{\theta}_{\text{dec}}). \quad (15)$$

C. Training

In our method, the transformer encoder-decoder trained via the paraphrasing task is leveraged for transformer encoder-decoder based image captioning. In a pre-training phase, a paraphrasing model composed of the encoder-decoder network and a text embedding network is trained using the paraphrasing datasets. In the fine-tuning phase, an image captioning model composed of the encoder-decoder network and an image embedding network is trained. Note that text embedding network and image embedding network are not shared because the inputs are definitely different.

Pre-training phase: In a pre-training phase, the paraphrasing model is trained using D_{para}. The training loss function L_{p} is defined as

$$L_{\text{p}} = -\frac{1}{N} \sum_{n=1}^{N} \frac{1}{K^n (K^n - 1)} \sum_{k'=1}^{K^n} \sum_{k=1, k \neq k'}^{K^n} \log P(\boldsymbol{W}_k^n | \bar{\boldsymbol{W}}_{k'}^n; \boldsymbol{\Theta}_{\text{t2t}}). \quad (16)$$

Fine-tuning phase: In the fine-tuning phase, the image captioning model is trained using pre-trained encoder-decoder parameters. Thus, $\boldsymbol{\theta}_{\text{enc}}$ and $\boldsymbol{\theta}_{\text{dec}}$ are pre-trained using the aforementioned loss function and are unfrozen in the fine-tuning. Note that the parameters for the object detection in $\boldsymbol{\theta}_{\text{image}}$ are also pre-trained using object detection modeling and are frozen during fine-tuning [8]. The training loss function L_{f} is defined as

$$L_{\text{f}} = -\frac{1}{M} \sum_{m=1}^{M} \frac{1}{K^m} \sum_{k=1}^{K^m} \log P(\boldsymbol{W}_k^m | \boldsymbol{C}_k^m; \boldsymbol{\Theta}_{\text{i2t}}). \quad (17)$$

In the experiments, we additionally examined decoder pre-training that only transfers the pre-trained decoder parameter $\boldsymbol{\theta}_{\text{dec}}$ and encoder pre-training that only transfers the pre-trained encoder parameter $\boldsymbol{\theta}_{\text{enc}}$ in the fine-tuning phase.

IV. Experiments

The effectiveness of our proposed method was evaluated using MS COCO caption datasets [21]. We used the datasets for not only evaluating image captioning but also building paraphrasing models as seen in previous work [12], [13]. In the datasets, multiple captioning texts were annotated into each

TABLE I
RESULTS IN TERMS OF BLEU (B-1–B-4), METEOR, ROUGE-L, AND CIDER.

Method		#captions	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
(1). Baseline	No pre-training	1	0.689	0.520	0.386	0.290	0.253	0.522	0.926
(2). Proposed	Decoder pre-training	1	0.701	0.532	0.393	0.298	0.256	0.529	0.948
(3). Proposed	Encoder pre-training	1	0.696	0.525	0.390	0.296	0.254	0.526	0.940
(4). Proposed	**Encoder-decoder pre-training**	1	**0.715**	**0.547**	**0.402**	**0.305**	**0.257**	**0.535**	**0.966**
(1). Baseline	No pre-training	2	0.719	0.551	0.414	0.313	0.258	0.538	0.984
(2). Proposed	Decoder pre-training	2	0.726	0.558	0.420	0.317	0.261	0.540	1.014
(3). Proposed	Encoder pre-training	2	0.723	0.555	0.415	0.314	0.260	0.538	1.005
(4). Proposed	**Encoder-decoder pre-training**	2	**0.730**	**0.562**	**0.423**	**0.319**	**0.262**	**0.541**	**1.023**
(1). Baseline	No pre-training	5	0.732	0.564	0.425	0.322	0.265	0.544	1.047
(2). Proposed	Decoder pre-training	5	0.738	0.568	0.430	0.327	0.268	0.548	1.054
(3). Proposed	Encoder pre-training	5	0.736	0.565	0.426	0.324	0.266	0.547	1.050
(4). Proposed	**Encoder-decoder pre-training**	5	**0.745**	**0.575**	**0.439**	**0.334**	**0.269**	**0.550**	**1.062**

image. We regarded the multiple annotations as paraphrases. To this end, we first split all the datasets into 113,287 training datasets, 5,000 validation datasets, and 5,000 test datasets in accordance with previous work [2]. In addition, we split 113,287 training datasets into 40,000 image captioning training datasets and 73,287 paraphrasing training datasets. We used five captioning texts for the paraphrasing training datasets. In other words, we used 1,465,740 ($5 \times 4 \times 73{,}287$) text-to-text paired datasets for pre-training. In addition, for the image captioning training datasets, we varied the number of annotated captions to construct image captioning models in order to evaluate the cases of data limitation. We lowercased all text, and all words were registered in the dictionary.

A. Setups

We evaluated four training methods by changing the number of annotated captions for each image.

1) **No pre-training:** The image captioning model was trained from only the image captioning training datasets. This is our baseline method.
2) **Decoder pre-training:** The paraphrasing model was first trained from the paraphrasing datasets. Then, the pre-trained decoder parameters were transferred into the image captioning model, and the encoder-decoder was fine-tuned using the image captioning datasets.
3) **Encoder pre-training:** The paraphrasing model was first trained from the paraphrasing datasets. Then, the pre-trained encoder parameters were transferred into the image captioning model, and the encoder-decoder was fine-tuned by using the image captioning datasets.
4) **Encoder-decoder pre-training:** The paraphrasing model was first trained from the paraphrasing datasets. Then, the parameters of both the pre-trained encoder and the pre-trained decoder were transferred into the image captioning model, and the encoder-decoder was fine-tuned by using the image captioning datasets. This is our proposed method.

Note that the latter two methods used the pre-trained encoder familiar with text inputs to enhance image captioning models.

For the image encoder, we used the Faster R-CNN [3] with a ResNet-101 backbone [27]. For the transformer encoder-decoder, we stacked three encoder blocks and three decoder blocks. The transformer blocks were constructed under these conditions: the number of dimensions of the output continuous representations was set to 256, the number of dimensions of the inner outputs in the position-wise feed-forward networks was set to 2,048, and the number of heads in the multi-head attention layer was set to 4. In the nonlinear transformational functions, a Gaussian error linear unit activation was used. We used the Radam optimizer for the training. The training steps were stopped early using the validation sets. We set the mini-batch size to 64 and the dropout rate in the transformer blocks to 0.1. We introduced label smoothing, where the smoothing parameter was set to 0.1. For testing, we used a beam search algorithm in which the beam size was set to 20.

B. Results

Table 1 shows image captioning evaluation results in terms of major image captioning evaluation metrics: BLEU [28], METEOR [29], ROUGE [30], and CIDEr [31]. Note that we varied the number of annotated captions (#captions) to construct image captioning models in order to evaluate the cases of data limitation.

We verified the effectiveness of our proposed method. First, the results show the performance of image captioning was affected by the number of captioning texts. This indicates that the number of image-to-text paired datasets needs to be increased for robust image captioning modeling. Our proposed encoder-decoder pre-training yielded higher performance than no pre-training in each setup. This proves that our proposed encoder-decoder pre-training via paraphrasing effectively improves the performance of image captioning when image-to-text paired datasets are limited.

Table 1 also shows an ablation study in which either a pre-trained encoder network or pre-trained decoder network is only utilized for improving the image captioning task. The results show the performance was improved by introducing each pre-training method. Although encoder pre-training used the pre-trained encoder network familiar with text inputs, the performance was improved. In particular, encoder-decoder pre-training outperformed decoder pre-training and encoder pre-training in each setup. This indicates that the paraphrasing model handles a similar transformation task to the image captioning model in not only the decoder network but also the encoder network. Thus, we can conclude that knowledge extracted from text-to-text transformation tasks is effective for improving image-to-text transformation tasks even though the individual inputs were inconsistent.

V. Conclusions

This paper proposed encoder-decoder pre-training using a paraphrasing task to improve neural image captioning. In our method, a transformer encoder-decoder trained via a paraphrasing task is leveraged for transformer encoder-decoder based image captioning. The key advance is to effectively utilize the text-to-text encoder-decoder for improving not only the text decoder but also the image encoder of image captioning. Thus, our proposed method can leverage knowledge extracted from text-to-text transformation tasks for directly improving image-to-text transformation tasks. Experimental results demonstrated that our proposed encoder-decoder pre-training effectively improves image captioning performance by utilizing paraphrasing datasets even though limited image-to-text paired datasets can be collected.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and Tell: A neural image caption generator," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2015.

[2] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *In Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 91–99, 2015.

[4] 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 Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077–6086, 2018.

[5] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image captioning with semantic attention," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4651–4659, 2016.

[6] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 375–383, 2017.

[7] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5659–5667, 2017.

[8] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares, "Image captioning: Transforming objects into words," *In Proc. Conference on Neural Information Processing Systems (NeurIPS)*, pp. 11137–11147, 2019.

[9] Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault, "Image captioning through image transformer," *arXiv preprint arXiv:2004.14231*, 2020.

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara, "Meshed-memory transformer for image captioning," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10578–10587, 2020.

[11] Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo, "Generating diverse and descriptive image captions using visual paraphrases," *In Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4240–4249, 2019.

[12] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri, "Neural paraphrase generation with stacked residual lstm networks," *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 2923–2934, 2016.

[13] Elozino Egonmwan and Yllias Chali, "Transformer and seq2seq model for paraphrase generation," *In Proc. Workshop on Neural Generation and Translation (WNGT)*, pp. 249–255, 2019.

[14] Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge, "A task in a suit and a tie: Paraphrase generation with semantic augmentation," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 7176–7183, 2019.

[15] Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu, "Decomposable neural paraphrase generation," *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3403–3414, 2019.

[16] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid, "VideoBERT: A joint model for video and language representation learning," *In Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7464–7473, 2019.

[17] Hao Tan and Mohit Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," *In Proc. Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

[18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *In Proc. Conference on Neural Information Processing Systems (NeurIPS)*, pp. 13–23, 2019.

[19] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao, "Unified vision-language pre-training for image captioning and VQA," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13041–13049, 2020.

[20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, "UNITER: UNiversal Image-TExt Representation Learning," *In Proc. European Conference on Computer Vision (ECCV)*, pp. 104–120, 2020.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollar, "Microsoft COCO: Common objects in context," *In Proc. European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

[22] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," *In Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 521–530, 2017.

[23] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon, "Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach," *In Proc. Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2012–2023, 2019.

[24] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo, "Unsupervised image captioning," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4125–4134, 2019.

[25] Iro Laina, Christian Rupprecht, and Nassir Navab, "Towards unsupervised image captioning with shared multimodal embeddings," *In Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7414–7424, 2020.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *In Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation," *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

[29] Satanjeev Banerjee and Alon Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," *In Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

[30] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," *In Proc. Text Summarization Branches Out*, pp. 74–81, 2004.

[31] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh, "CIDEr: Consensus-based image description evaluation," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.