HO CHI MINH CITY

# CERTIFICATE

**Bui Cao Doanh**

**Vietnam National University Ho Chi Minh City**

**University of Information Technology**

**Participated in the final round of**

**Student Scientific Research Prize**

Euréka Lần XXIII 2021

*Ho Chi Minh City, 23rd November 2021*

**ON BEHALF OF STANDING COMMITTEE OF**
**HO CHI MINH COMMUNIST YOUTH UNION OF HO CHI MINH CITY**
SECRETARY

(signed)

**Phan Thi Thanh Phuong**

THÀNH PHỐ HỒ CHÍ MINH

# CHỨNG NHẬN

**Bùi Cao Doanh**

**Trường Đại học Công nghệ Thông tin**

**Đại học Quốc gia Thành phố Hồ Chí Minh**

**Tham gia và vào vòng Chung kết**

**Giải thưởng Sinh viên Nghiên cứu Khoa học**

Euréka Lần XXIII 2021

*TP. Hồ Chí Minh, ngày 23 tháng 11 năm 2021*

**TM. BAN THƯỜNG VỤ THÀNH ĐOÀN**
BÍ THƯ

BCH ĐOÀN THANH PHỐ HỒ CHÍ MINH

**Phan Thị Thanh Phương**

| TT | Họ và Tên SV CNĐT | MSSV (CNĐT) | Họ và Tên SV tham gia (nếu có) | Khoa | GVHD | Tên đề tài | Đợt | Kinh phí (triệu đồng) |
|---|---|---|---|---|---|---|---|---|
| *(1)* | *(2)* | *(3)* | *(4)* | *(5)* | *(6)* | *(7)* | *(8)* | *(9)* |
| 41 | VŨ NGỌC TÚ | 18520184 | | KHOA KHOA HỌC MÁY TÍNH | NGUYỄN TẤN TRẦN MINH KHANG | PHÁT HIỆN SỰ KIỆN BẤT THƯỜNG TRONG VIDEO GIÁM SÁT SỬ DỤNG DEEP LEARNING | 5/2021 | 6 |
| 42 | ĐỖ NGUYỄN THUẬN PHONG | 18520126 | | KHOA KHOA HỌC MÁY TÍNH | NGUYỄN VĂN KIỆT | NGHIÊN CỨU ĐỌC HIỂU TỰ ĐỘNG VĂN BẢN DỰA TRÊN CẤP ĐỘ CÂU CHO TIẾNG VIỆT | 5/2021 | 6 |
| 43 | NGUYỄN ĐỨC TOÀN | 18521506 | | KHOA KHOA HỌC MÁY TÍNH | VÕ DUY NGUYÊN | KHỬ SƯƠNG MỜ CHO PHÁT HIỆN ĐỐI TƯỢNG TRONG KHÔNG ẢNH | 5/2021 | 6 |
| 44 | BÙI CAO DOANH | 19521366 | | KHOA KHOA HỌC MÁY TÍNH | VÕ DUY NGUYÊN | PHÁT HIỆN PHƯƠNG TIỆN GIAO THÔNG TRONG KHÔNG ẢNH DỰA TRÊN PHƯƠNG PHÁP TĂNG CƯỜNG DỮ LIỆU BẰNG CÁCH CẮT NGẪU NHIÊN | 5/2021 | 6 |
| 45 | BÙI TRẦN NGỌC DŨNG | 19521385 | | KHOA KHOA HỌC MÁY TÍNH | VÕ DUY NGUYÊN | TÌM HIỂU PHƯƠNG PHÁP PHÂN LOẠI PHƯƠNG TIỆN GIAO THÔNG Ở VIỆT NAM BẰNG HÌNH ẢNH TRÊN KHÔNG SỬ DỤNG MACHINE LEARNING VÀ DEEP LEARNING. | 5/2021 | 6 |
| 46 | PHẠM XUÂN TRÍ | 18521530 | | KHOA KHOA HỌC MÁY TÍNH | NGÔ ĐỨC THÀNH | NGHIÊN CỨU CÁC PHƯƠNG PHÁP PHÁT HIỆN CÔNG THỨC TOÁN HỌC TRONG ẢNH KĨ THUẬT SỐ. | 5/2021 | 6 |

# An Augmented Embedding Spaces approach for Text-based Image Captioning

Doanh C. Bui, Truc Trinh, Nguyen D. Vo, Khang Nguyen
*University of Information Technology, Ho Chi Minh City, Vietnam*
*Vietnam National University, Ho Chi Minh City, Vietnam*
{19521366, 19521059}@gm.uit.edu.vn, {nguyenvd, khangnttm}@uit.edu.vn

***Abstract*—Scene text-based Image Captioning is the problem that generates caption for an input image using both contexts of image and scene text information. To improve the performance of this problem, in this paper, we propose two modules, Objects-augmented and Grid features augmentation, to enhance spatial location information and global information understanding in images based on M4C-Captioner architecture for text-based Image Captioning problems. Experimental results on the TextCaps dataset show that our method achieves superior performance compared with the M4C-Captioner baseline approach. Our highest result on the Standard Test set is 20.02% and 85.64% in the two metrics BLEU4 and CIDEr, respectively.**

***Index Terms*—image captioning, text-based image captioning, relative geometry, grid features, region features, bottom up top down**

## I. Introduction

The image content sometimes depends not only on the objects, but also on the text appearing around in the image. Some tasks about automatic document understanding on document images, identity cards, receipts, scientific papers heavily depend on texts on images [1], [2]. With Image Captioning, taking advantage of the text present could help generate image descriptions more realistic automatically. In that way, people could better understand the content of an image via predicted description. For the mentioned purpose, the TextCaps dataset [3] was form to promote research and development on text-based image captioning, which requires artificial intelligence systems to read and infer meaning from text in the image to generate coherent descriptions. Hardly had a method that paid attention to comprehending text in the context of an image but focused on the objects or general features to generate description before the TextCaps dataset was published. After the introduction of TextCaps, the M4C-Captioner [4] method (improved by M4C for the VQA problem) was considered as a baseline for solving this problem, and later studies on scene text-based image captioning were mostly improved from M4C-Captioner.

It would seem that M4C-Captioner ignored the location information of objects in the image. With that observation, in this paper, we conduct experiments and contributions with two simple but effective modules as follows:

1) We propose the **objects-augmented module** for the addition of spatial location information between objects and OCR tokens.
2) We propose **grid features augmentation module** that suggest to augment global semantic information of the image by combining grid features.
3) We achieve the better results compared to M4C-Captioner baseline and competitive results versus other methods.

Some comparisons results between our method and M4C-Captioner baseline are shown in Figure 1.

The rest of the paper is structured as follows: Section II provides an overview of image captioning; Section III describes clearly our proposed method; Section IV shows our experiments and results. Finally, conclusions are drawn in Section V.

## II. Related works

### A. Overview

*1) Image Captioning:* It is a function that automatically generates textual description of an image. Currently, there have been many studies showing high BLEU4 results on the MS-COCO dataset. The common approach of Image Captioning is to use a CNN architecture to extract image features, then apply RNNs as a sequence decoder to generate output word by word at time $t$. Therefore, previous studies on this problem often suggest improving image features understanding, language models as well as applying other techniques such as: RL training [5], Model Language Mask [6] or apply BERT-like architectures to combine image and language features, or combine object tags which are predicted by object detectors with image features [7]–[9], etc.

*2) Scene text-based Image Captioning:* Although having a good BLEU4 performance metric on MS-COCO dataset, traditional Image Captioning approaches are only trained to generate sentences based on objects in the image which totally ignore textual information. To promote research on Scene text-based Image Captioning, Sidorov *et al.* has published the TextCaps dataset, which requires the generation of textual description for images to be depended on the text feature contained in the image. A currently well-known method for this problem is M4C-Captioner which we will introduce later in this section. The existing studies on Scene text-based Image Captioning are now mostly improved from M4C-Captioner.


Figure 1: The figure shows some visualizations that compare our method with the M4C-Captioner baseline. The red text indicates that M4C-Captioner's predictions are not suitable with the image's context or does not have enough words to describe the image. The green text indicates that our predictions seem better.

### B. Visual presentation

Currently, the Image Captioning problem has two main ways of how images is represented which is grid features and region features.

*1) Grid features:* Grid features are semantic features extracted from existing CNN network architectures such as ResNet [10], or VGG [11]. This form of image representation has shown impressive results in the early stages of the Image Captioning problem. In recent years, the emergence of region features has made grid features no longer be used much. However recently, Jiang *et al.* revisited the grid features by extracting the grid features at the same layer of object detector which was used to extract region features. This approach is less time consuming but gives more competitive performance versus region features.

*2) Region features:* Grid features usually focus only on global semantic information, which means that the model does not really pay attention to any particular location in the image. to overcome this issue, Anderson *et al.* proposed a bottom-up and top-down method [13] that uses Faster R-CNN to extract region features. Specifically, the Regional Proposal Network (RPN) proposes areas on feature maps that have a high possibility of the object appearing in them. These regions are then passed through RoI Pooling to be transformed into same-size vectors. After that, these vectors will be used to represent an image. Correct use of semantic vectors of potential regions means that the image features will include more valuable information, and the model could learn more things about the image.

### C. Multimodal Multi-Copy Mesh (M4C)

Proposed by Hu *et al.*, this model is originally built to solve the VQA problem by being based on a pointer-augmented multimodal transformer architecture with iterative answer prediction [4]. In particular, the authors use all three information: question, visual objects and text in order to represent images which question is represented by vector word embedding, visual objects features are extracted from object detector, and OCR token features represent texts. The coordinates to retrieve the OCR feature are determined by an external OCR system. The authors also propose a Dynamic Pointer Network to decide at which point $t$ a word in vocabulary or an OCR token should be selected. However, M4C-Captioner only takes the information of visual objects and text regions that are presented in an image in the text-based Image Captioning problem. Still, location information of objects is not exploited in this architecture.

## III. Methodology

In this section, we present our proposed modules in the text-based image captioning problem. Figure 2 shows our general architecture.

### A. Objects-augmented module

Originally, M4C-Captioner use an object detector at [13] to obtain a set of $M$ features of visual objects ($x_m^{fr}$). The authors additionally use a set of coordinates ($x_m^b$) $[x_{\min}/W_{\text{im}}, y_{\min}/H_{\text{im}}, x_{\max}/W_{\text{im}}, y_{\max}/H_{\text{im}}]$ to present location information between them. The final visual objects presentation used to train the Mutilmodal Transformer model is the combination of $x_m^{fr}$ and $x_m^b$. With texts that appear in images, the authors use the **Rosetta-en** OCR system to obtain $N$ coordinates of text regions ($x_n^b$), then extract theirs features ($x_n^{fr}$) by using the same detector at the same layer used to extract visual objects features. Sub-words in these text regions are embedded using FastText [14] ($x_n^{ft}$), and characters are embedded using PHOC [15] ($x_n^P$). The final OCR tokens presentation is the combination of $x_n^{fr}$, $x_n^{ft}$, $x_n^P$ and $x_n^b$. Nevertheless, we suppose that combining bounding boxes information still does not show spatial location information, so we propose **Objects-augmented module** that helps interpolate

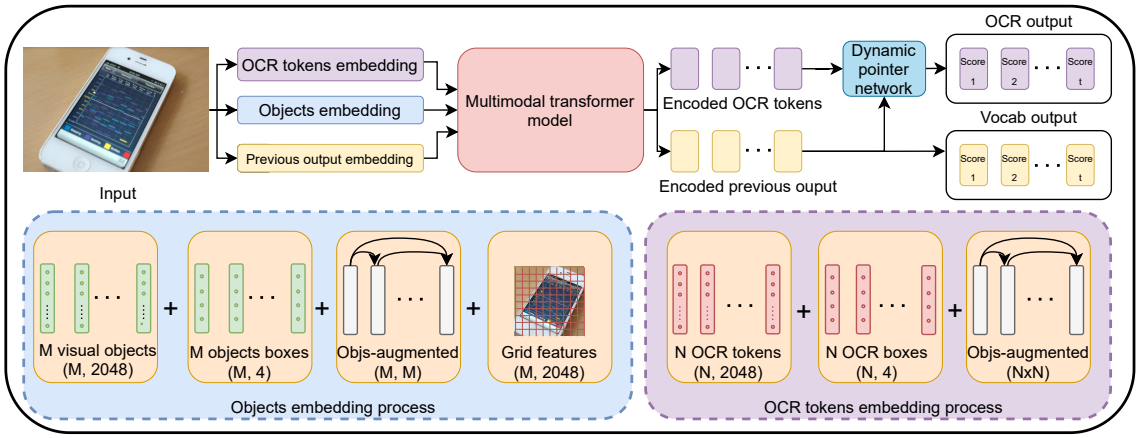



Figure 2: An overview of our two proposed modules based on M4C-Captioner. We propose an objects-augmented module for augmenting spatial location information between objects as well as OCR tokens. Besides, we also propose grid features augmentation module for augmenting the global semantic feature of an image.

relative geometry relationships between visual objects and OCR tokens.

First, we calculate centre coordinates of bounding boxes $(cx_i, cy_i)$, width $w_i$ and height $h_i$ by Equation 1, 2, 3 below:

$$(cx_i, cx_i) = (\frac{x_i^{min} + x_i^{max}}{2}, \frac{y_i^{min} + y_i^{max}}{2}), \tag{1}$$

$$w_i = (x_i^{max} - x_i^{min}) + 1, \tag{2}$$

$$h_i = (y_i^{max} - y_i^{min}) + 1, \tag{3}$$

Finally, we follow [16], [17] to obtain the relative geometry features between two objects/OCR tokens $i$ and $j$ by Equation 1, 2, 3:

$$r_{ij} = \begin{pmatrix} \log(\frac{|cx_i - cx_j|}{h_i}) \\ \log(\frac{|cy_i - cy_j|}{h_i}) \\ \log(\frac{w_i}{h_j}) \\ \log(\frac{h_i}{h_j}) \end{pmatrix}, \tag{4}$$

$$G_{ij} = FC(r_{ij}), \tag{5}$$

$$\lambda_{ij}^g = ReLU(w_g^T G_{ij}), \tag{6}$$

Where $r \in \mathbb{R}^{N \times N \times 4}$ is relative geometry relationship between grids; $FC$ is a fully-connected layer with activation function; $G \in \mathbb{R}^{N \times N \times d_g}$ is a high-dimensional presentation of $r$, in which $d_g = 64$; $w_g$ is learned weight matrix;

By above operations, we obtain relative geometry features of visual objects features ($\lambda_{objs}^g$) and OCR tokens ($\lambda_{ocr}^g$) in an image. Then $\lambda_{objs}^g$ is combined with $x_m^{fr}$ and $x_m^b$; $\lambda_{ocr}^g$ is combined with $x_n^{fr}$, $x_n^{ft}$, $x_n^{PHOC}$ and $x_n^b$ by Equation 7, 8:

$$x_m^{\text{obj}} = \text{LN}(W_1 x_m^{\text{fr}}) + \text{LN}(W_2 x_m^{\text{b}}) + \mathbf{LN}(\mathbf{W_3} \lambda_{\mathbf{objs}}^{\mathbf{g}}) \tag{7}$$

$$x_n^{\text{ocr}} = \text{LN}(W_4 x_n^{\text{ft}} + W_5 x_n^{\text{fr}} + W_6 x_n^{\text{PHOC}}) + \text{LN}(W_7 x_n^{\text{b}}) + \mathbf{LN}(\mathbf{W_8} \lambda_{\mathbf{ocr}}^{\mathbf{g}}) \tag{8}$$

### B. Grid features augmentation

Although region features help the Multimodal Transformer model pay attention to specific regions that can infer the description, we suppose that grid features contain the global semantic of the image can augment the ability to represent image semantics, helping the model learn more information; therefore, we proposed **Grid Features Augmentation module**. We follow [12] to extract grid features; in detail, Jiang *et al.* use bottom-up, top-down architecture [13] to compute feature maps from lower blocks of ResNet to block $C_4$. But instead of using $14 \times 14$ `RoIPooling` to compute $C_4$ output features, then feed to $C_5$ block and apply `AveragePooling` to compute per-region features, they convert the detector in [13] back to the ResNet classifier and compute grid features at the same $C_5$ block. By experiments, they observe that using converted $C_5$ block directly helps reduce computational time but achieve surprising results. After extracting, grid features are $2048-d$ matrices that have the shape of $(H, W)$; we apply `AdaptiveAvgPool2d` $(m, m)$ to reshape grid features to $(m, 2048)$, where $m$ is the number of visual objects.

Then we combine grid features with $x_m^{\text{obj}}$ by the following equation:

$$x_m^{finalobj} = x_m^{obj} + \mathbf{LN}(\mathbf{W_9 x_m^{grids}}) \tag{9}$$

Where $x_m^{obj}$ is computed from Equation 7, $\{W_i\}_{i=1:9}$ are learned projection matrices and $\text{LN}(\cdot)$ is layer normalization.

## IV. Experiment

### A. Machine configuration

Our machine configuration: 1) Processor: Intel(R) Core(TM) i9-10900X CPU @ 3; 2) Memory: 64GiB; 3)

Table I: Evaluation results on TextCaps Validation set

| # | Method | Proposed module | | | TextCaps validation set metrics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RG features | | Grid features | B4 | M | R | S | C |
| | | Objs | OCR | | | | | | |
| 1 | BUTD [13] | | | | 20.1 | 17.8 | 42.9 | 11.7 | 41.9 |
| 2 | AoANet [23] | | | | 20.4 | 18.9 | 42.9 | 13.2 | 42.7 |
| 3 | M4C-Captioner [3] | | | | 23.3 | 22.0 | 46.2 | 15.6 | 89.6 |
| **4** | **Ours** | ✓ | ✓ | ✓ | **23.79** | **22.7** | **46.77** | **16.34** | **93.97** |

Table II: Evaluation results on TextCaps Test set

| # | Method | Proposed module | | | TextCaps test set metrics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RG features | | Grid features | B4 | M | R | S | C |
| | | Objs | OCR | | | | | | |
| 1 | BUTD [13] | | | | 14.9 | 15.2 | 39.9 | 8.8 | 33.8 |
| 2 | AoANet [23] | | | | 15.9 | 16.6 | 40.4 | 10.5 | 34.6 |
| 3 | M4C-Captioner [3] | | | | 18.9 | 19.8 | 43.2 | 12.8 | 81.0 |
| 4 | Ours | ✓ | | | 19.32 | 20.46 | 43.82 | 13.27 | 82.32 |
| 5 | Ours | ✓ | | ✓ | 19.83 | 20.82 | 44.25 | 13.77 | 84.69 |
| **6** | **Ours** | ✓ | ✓ | ✓ | **20.02** | **20.89** | **44.41** | **13.74** | **85.64** |
| 7 | Human [3] | | | | 24.4 | 26.1 | 47.0 | 18.8 | 125.5 |

GPU: 1× GeForce RTX 2080 Ti 11GiB; 4) OS: Ubuntu 20.04.1 LTS. We train the model in 12000 iterations with batch size = 64.

### *B. Dataset*

We evaluate experiments of our proposed modules on the TextCaps dataset[3]. It contains 28,408 images from Open-Images; one image has five ground-truth captions, so there are 142,040 captions in total. Besides, the 6th caption is also prepared per image for comparing performance between AI model with human. Before TextCaps, there was COCO dataset, which is also used for Image Captioning or TextVQA tasks, but the statistics show that there are only 2.7% of captions and 12.7% of images have at least one OCR token; obviously, it is not suitable for Text-based Image Captioning. These numbers of the TextCaps dataset are 81.3% and 96.9%, respectively. Furthermore, some images in the TextCaps dataset that OCR tokens are not presented directly in ground-truth captions, but they should be used to infer descriptions of these images. Therefore, formulating the predicted caption based on heuristic approaches is impossible.

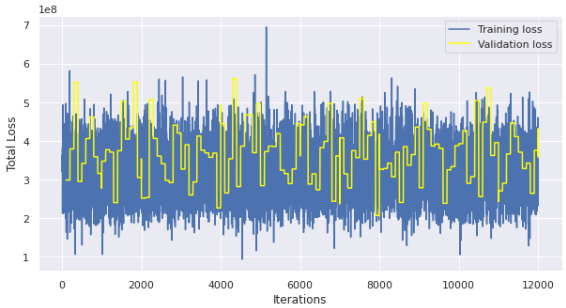
After training, we export the output and submit it to eval.ai (https://eval.ai/web/challenges/challenge-page/906). The results on the Validation set and Test set are reported in Tables I and II.

### *C. Metrics*

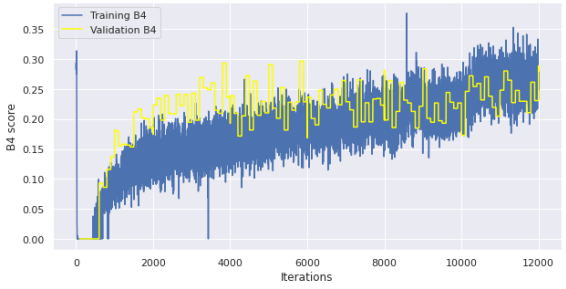
We use five standard metrics for Machine-Translation or Image Captioning to measure the performance of our proposed modules: BLEU (B) [18], METEOR (M) [19], ROUGE_L (R) [20], SPICE (S) [21] and CIDEr (C) [22]. We focus on BLEU and CIDEr scores. BLEU score is popular and always used to evaluate the difference between two sequences. Besides, the CIDEr score is a new metric that will put more weights on more informative tokens so that it is more suitable for Text-based Image Captioning.

### *D. Main results*

Experimental results in Table I and II obviously witness previous methods in Image Captioning such as BUTD[13] or AoANet[23] do not achieve expected results due to their limitations of paying attention to OCR tokens. M4C-Captioner based on M4C architecture improves the performance conspicuously when compared to BUTD (**B4 +4%**) and AoANet (**B4 +3%**). Nevertheless, exactly what we hypothesize, lacking spatial information make M4C-Captioner does not achieve the expected performance. Our Objects-augmented module applied in visual objects features at embedding step achieves higher scores when compared with M4C-Captioner (**B4 +0.42%** and **CIDEr +1.32%**). When combined with Grid features augmentation, the performance witnessed an obvious improvement (**B4 +0.93%** and **CIDEr +3.69%**). Finally, combining our two proposed modules, which means applying objects-augmented on both visual objects features and OCR tokens and adding Grid features to Visual objects features, achieves the highest performance (**B4 20.02%** and **CIDEr 85.64%**). Besides, we also plot the loss function values (Figure 3a) and BLEU4 (Figure 3b) on the training and validation sets over the entire 12000 iterations. Figure 3b shows that BLEU4 gradually increases (unstable) during the first 6000 iterations, then tends to fluctuate around the 20% to 25% range but does not reach a new peak.




(a) Variation of the value of loss function




(b) Variation of the value of B4 score

Figure 3: The change in the value of the loss function and B4 score during training time.

## V. Conclusion

In conclusion, we propose two simple but effective modules: **Objects-augmented** and **Grid features augmentation**. Objects-augmented is used for enhancing spatial information and Grid features augmentation is used to augment the global semantic of images. Our experimental results show that combining our two proposed modules is more effective than the

original M4C-Captioner, and the performance can be further improved if training time increases. In the future, we plan to collect the Vietnamese dataset for the Text-based Image Captioning problem and use more valuable information such as object tags and classified objects in the embedding process, which are hoped to increase the results.

## ACKNOWLEDGMENT

This work was supported by the Multimedia Processing Lab (MMLab) and UIT-Together research group at the University of Information Technology, VNUHCM.

## REFERENCES

[1] D. C. Bui, D. Truong, N. D. Vo, and K. Nguyen, "Mc-ocr challenge 2021: Deep learning approach for vietnamese receipts ocr," Accepted as regular paper in RIVF2021 conference.

[2] M. Li, Y. Xu, L. Cui, *et al.*, *Docbank: A benchmark dataset for document layout analysis*, 2020. arXiv: 2006.01038 [cs.CL].

[3] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: A dataset for image captioning with reading comprehension," in *European Conference on Computer Vision*, Springer, 2020, pp. 742–758.

[4] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9992–10 002.

[5] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[6] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," *arXiv preprint arXiv:1904.09324*, 2019.

[7] W. Su, X. Zhu, Y. Cao, *et al.*, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[8] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 041–13 049.

[9] X. Li, X. Yin, C. Li, *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*, Springer, 2020, pp. 121–137.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 267–10 276.

[13] P. Anderson, X. He, C. Buehler, *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[15] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[16] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 327–10 336.

[17] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *arXiv preprint arXiv:1906.05963*, 2019.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[19] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

[20] L. C. ROUGE, "A package for automatic evaluation of summaries," in *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004.

[21] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*, Springer, 2016, pp. 382–398.

[22] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[23] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4634–4643.