

Multi-Layer Content Interaction Through Quaternion Product For Visual Question Answering

Shi, Lei; Geng, Shijie; Shuang, Kai; Hori, Chiori; Liu, Songxiang; Gao, Peng; Su, Sen

TR2020-046 April 11, 2020

Abstract

Multi-modality fusion technologies have greatly improved the performance of neural network-based Video Description/Caption, Visual Question Answering (VQA) and Audio Visual Scene-aware Dialog (AVSD) over the recent years. Most previous approaches only explore the last layers of multiple layer feature fusion while omitting the importance of intermediate layers. To solve the issue for the intermediate layers, we propose an efficient Quaternion Block Network (QBN) to learn interaction not only for the last layer but also for all intermediate layers simultaneously. In our proposed QBN, we use the holistic text features to guide the update of visual features. In the meantime, Hamilton quaternion products can efficiently perform information flow from higher layers to lower layers for both visual and text modalities. The evaluation results show our QBN improve the performance on VQA 2.0, furthermore surpasses the approach using large scale BERT or visual BERT pre-trained models. Extensive ablation study has been carried out to examine the influence of each proposed module in this study.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

MULTI-LAYER CONTENT INTERACTION THROUGH QUATERNION PRODUCT FOR VISUAL QUESTION ANSWERING

Lei Shi[†], Shijie Geng[◇], Kai Shuang[†], Chiori Hori^{*}, Songxiang Liu[‡], Peng Gao[‡], Sen Su[†]

[†]Beijing University of Posts and Telecommunications [◇] Rutgers University
[‡]CUHK ^{*}Mitsubishi Electric Research Laboratories (MERL)

ABSTRACT

Multi-modality fusion technologies have greatly improved the performance of neural network-based Video Description/Caption, Visual Question Answering (VQA) and Audio Visual Scene-aware Dialog (AVSD) over the recent years. Most previous approaches only explore the last layers of multiple layer feature fusion while omitting the importance of intermediate layers. To solve the issue for the intermediate layers, we propose an efficient Quaternion Block Network (QBN) to learn interaction not only for the last layer but also for all intermediate layers simultaneously. In our proposed QBN, we use the holistic text features to guide the update of visual features. In the meantime, Hamilton quaternion products can efficiently perform information flow from higher layers to lower layers for both visual and text modalities. The evaluation results show our QBN improve the performance on VQA 2.0, furthermore surpasses the approach using large scale BERT or visual BERT pre-trained models. Extensive ablation study has been carried out to examine the influence of each proposed module in this study.

Index Terms— Visual question answering, quaternion product, multi-layer interaction, self-attention, co-attention, representation learning

1. INTRODUCTION

Multi-modality fusion technologies have been applied to neural network-based Video Description/Caption [1], Visual Question Answering (VQA) [2] and Audio Visual Scene-aware Dialog (AVSD) [3] over the recent years. Especially, VQA connecting vision and language becomes a standard benchmark for testing fundamental multimodality fusion algorithm. Multi-layer Intra- and Inter- layer network has been commonly applied for multimodality feature fusion on VQA. However, most approaches including Dynamic Fusion with Intra and Inter Modality Attention Flow (DFAF) [4] and Modular Co-attention Network (MCAN) [5] only uses the last layer for down-stream tasks such as VQA classification. We argue that the interaction captured in the last layer is not enough to model multimodality fusion. In fact, interaction in the intermediate layer can capture some important aspects of the interplay between intra- and inter- modality as shown in the final visualization part.

To address the issue of missing interaction information in intermediate layers, we propose a novel Quaternion Block Networks (QBN) which can capture multi-layer interaction with a novel Hamilton quaternion product. Our proposed QBN composes every intermediate four layers as a quaternion and performs a quaternion Hamilton product between two modalities visual and language quaternions (Visual and language). The complex interaction between all layers can be captured efficiently through our proposed approaches.

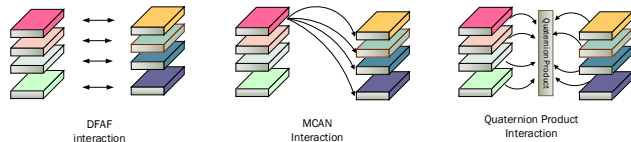


Fig. 1. Three ways of multimodal information interaction. DFAF updates information between the same hierarchy of multimodal layers; MCAN only utilizes one high-level modality to update all layers in another modality; Our QBN enables updates among all layers both intra- and inter- two modalities and builds better information flow for multiple layers.

The key contributions of the paper can be outlined as three-fold. (1) We propose Quaternion Block Networks(QBN) to explore the impact of multi-layer content and multi-layer relationship for visual question answering. (2) We carry out spatial scaling of visual features conditional on question feature in VQA, while the existing VQA model neglects discussion on this. (3) We execute extensive ablation studies for each component of QBN and achieve state-of-the-art performance on VQA v2.0 [6]. Surprisingly, our proposed QBN can even surpass BERT pretrained models like ViBERT.

2. RELATED WORK

2.1. Representation Learning

Representation learning has revolutionized both Computer Vision(CV) and Natural Language Processing(NLP). In CV, VGGNet [7], ResNet [8], SENet [9] and DenseNet [10] raise up to be the fundamental building block for downstream tasks. While in NLP, word2vec [11], GloVe [12], Skipthought [13], ELMo [14], and BERT [15] play a fundamental role in pushing the research boundary. In Visual Question Answering, BERT and Faster RCNN has been treated as a fundamental feature extractor.

2.2. Self-attention & Co-attention

Attention mechanism imitates human vision by putting large weight on important regions. Neural Attention has been popularized by [16]. Subsequently, [17] used the attention mechanism to perform machine translation task. Motivated by attention model in machine translation, [18] proposed self-attention mechanism, which has been a fundamental backbone for many CV and NLP applications.

After its success on pure CV and NLP tasks, researchers [4] [5] [19] start to apply self-attention and co-attention mechanism for solving multi-modality fusion among which Visual Question Answering has been served as a standard benchmark. DFAF [4] firstly introduce

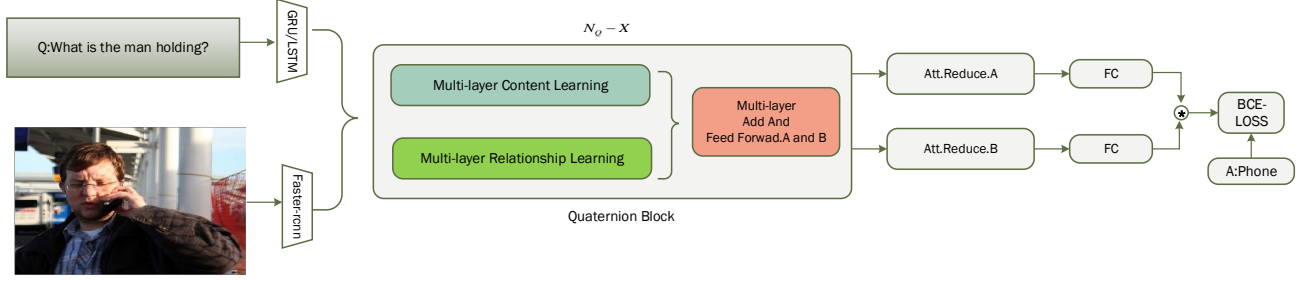


Fig. 2. The overall framework of Quaternion Block Networks. We only show one Quaternion block here, the X in $N_Q - X$ is the number of blocks. Each block is composed of three parts: 1) *Multi-layer Content Learning module*, which learns the holistic text features to guide the update of visual features. 2) *Multi-layer Relationship Learning module*, which builds connections between multi-layers of the two modalities. 3) *Feed Forward module*, which integrates information of multiple layers and outputs the fused information to next Quaternion block.

a intra- and inter- modality attention flow pipeline for approaching VQA. Separate intra-attention will be applied for message passing inside text and image, then inter-modality attention will enable cross-modality interaction. MCAN [5] performs intra modality attention several times followed by inter modality attention while DFAC interleaves Intra and Inter modality attention. MCAN can achieve slightly better performance than DFAC. MLIN [19] proposes a latent interaction mechanism which can reduce the quadratic computation of self attention into linear.

2.3. Quaternion Product

Quaternion connectionist architectures have been widely applied in various areas such as speech recognition [20], computer vision [21], and natural language processing [22]. Our proposed QBN will utilize Hamilton product to learn multi-layer interaction efficiently which result in improved robustness and better interpretability.

3. QUATERNION BLOCK NETWORKS

3.1. Overview

The proposed method includes a series of Quaternion Product Block. The whole pipeline is demonstrated in Fig.2, where visual and language features from the two modalities are first pre-processed. In the second step, the two modalities consisting of the pre-processed visual and language features are serially encoded three times by the self-attention mechanism; In this way, we obtain four layers of content information in visual and text modalities respectively. In the third step, Multi-layer textual information will be a better guide multi-modal information interaction at the appropriate level using our proposing multi-layer content operation as described in the Section 3.3 to learn the overall presentation of multiple layers of content in text modality. In the fourth step, we apply Quaternion Product mechanism in the Section 3.4 performing multi-layer relationship learning to better learn the relationship between text modality and visual modality in multiple layers of content.

Finally, we use the co-attention mechanism to implement multi-modal information to achieve information updates in the corresponding layer, taking into account the overall content of the multi-layer and the overall relationship of the multi-layer. Such Quaternion Product Block can be stacked multiple times to pass information iteratively flows between multi-layer words and regions to model potential alignments for visual problem-solving.

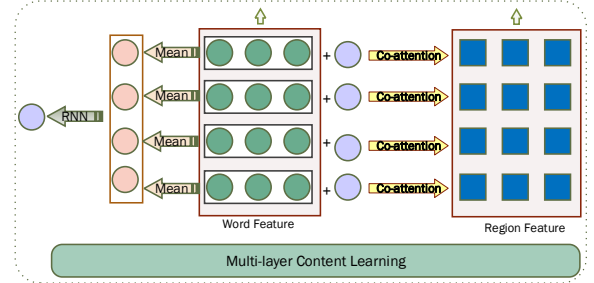


Fig. 3. Illustration of multi-layer content learning in a Quaternion block. In a Quaternion block, there are four layers of word features. We first average the words features of each layer, then input the average sentence vector of each layer’s word vectors into an RNN model, and then take the output hidden vector as the holistic textual feature of the whole block. Finally, we use co-attention to update the visual information of each layer in the quaternion block based on the holistic textual feature.

3.2. Visual and language feature preprocess

To obtain basic language features, we first embed each word in question Q using pre-trained Glove [12]. We then use an LSTM, generating a sequence of the hidden state $\{w_1, \dots, w_t\}$. Note that we use the last hidden state of LSTM w_t as a question feature, signified as $q_t \in \mathbb{R}^{512}$ and all questions are padded and truncated to the same length 14.

$$w = \text{LSTM}(Q), q_t = w_t \quad (1)$$

We extract image features, which are obtained from a Faster-RCNN [23] model pre-trained on Visual Genome [24] dataset. For each image I , we extract 100 region proposals and associated region features. However, different from bottom-up & top-down attention [25], We select the $R \in \mathbb{R}^{u \times 2 \times 2 \times 2048}$ image region feature as input. We map the dynamically changing question vector to the scaling factor and bias term of the channel feature through the fully connected layer f_c and h_c . The image region features are pre-processed by the sentence vector of Question to achieve effective scaling about image spatial-region features.

$$R = \text{RCNN}(I) \quad (2)$$

$$\gamma_{i,c} = f_c(q_t) \quad (3)$$

$$\beta_{i,c} = h_c(q_t) \quad (4)$$

$$v = \text{AveragePooling}(\gamma_{i,c} * R_{i,c} + \beta_{i,c}) \quad (5)$$

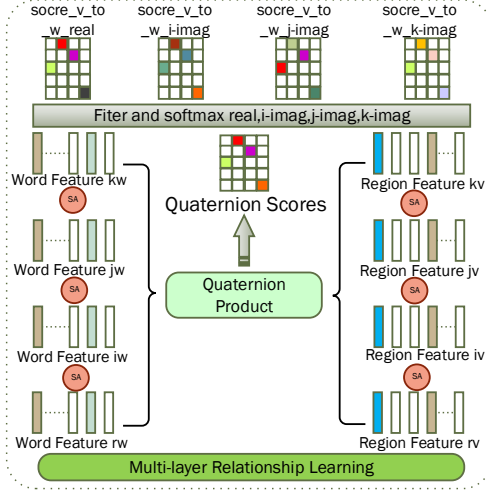


Fig. 4. Illustration of multi-layer relationship learning in a Quaternion block. In a Quaternion block, we regard the textual and visual information as different parts in two Quaternions. Then we utilize Quaternion Product to conduct relationship learning among different layers across the two modalities.

In the formula, γ and β represent the scaling factor and the bias term respectively; q_t represents question feature. In $R_{i,c} \in \mathbb{R}^{2 \times 2 \times 2048}$, i represents the i -th region feature; c represents the c -th channel of the 2048 channel. In $R = \{R_1, \dots, R_\mu\}$, μ represents an image with a total of μ region features. Finally, we average pooling the scaled region features.

3.3. Multi-layer content learning for text information

In the previous use of the self-attention [4][19] model to solve the VQA task, there are two common methods. One method is to use self-attention to encode each layer of text information and visual information, and then use the co-attention mechanism to realize information interaction between text modality and visual modality[4]. Another [5] method is to use self-attention mechanism to encode text information several times and learn high-level text information; Then high-level text information is used to interact with visual modality information by co-attention mechanism. These two methods are not considered enough because they do not reflect the impact of global text information formed by shallow and deep levels on multi-modal information interaction. So we use multi-layer global information of text to help multi-modal information interaction.

We first encode the text feature and the visual region feature three times using the self-attention mechanism. Considering the original text and visual information, so we have four layers of information in the text modality and the visual modality respectively. As shown in the following formula rv, iv, jv, kv and rw, iw, jw, kw represent the features of different layers; Sa stands for self-attention encoding.

$$\begin{aligned} rv = v ; iw = Sa(rv) ; jv = Sa(iv) ; kv = Sa(jv) \quad (6) \\ rw = w ; iw = Sa(rw) ; jw = Sa(iw) ; kw = Sa(jw) \quad (7) \end{aligned}$$

In order to reflect the overall information of each layer, we first average the text information of each layer. Mean represents the function of averaging.

$$xw_{\text{mean}} = \text{Mean}(xw), x \in \{r, i, j, k\} \quad (8)$$

$rw_{\text{mean}}, iw_{\text{mean}}, jw_{\text{mean}}, kw_{\text{mean}}$ represents the overall information of each layer of text features, in order to reflect the correlation between the layers of information, we use RNN to process the mean information of each layer. Multi_{q_n} represents the overall information of the multi-layer text feature in a Quaternion Block.

$$\text{Multi}_{q_n} = \text{RNN}(rw_{\text{mean}}, iw_{\text{mean}}, jw_{\text{mean}}, kw_{\text{mean}}) \quad (9)$$

We use the co-attention mechanism to achieve the fusion of visual information and text information, but unlike the previous method, we first add overall text features to each layer, so that the visual information of each layer is fused by the text information of the corresponding layer considering the overall feature of multi-layered text.

$$\begin{aligned} \text{Multi_attention}(query, key, value) \\ = \text{Concat}(head_1, \dots, head_n)w \quad (10) \end{aligned}$$

$$head_i = \text{Attention}(Q, K, V) \quad (11)$$

$$\begin{aligned} \text{Attention}(Q, K, V) = \\ \text{Softmax}\left(\frac{Q * K}{\sqrt{\text{dim}}}\right) * G_{\text{quaternion_score_softmax}} * V \quad (12) \end{aligned}$$

$$xv_{\text{update}} =$$

$$\begin{aligned} \text{Multi_attention}(xv, xw + \text{Multi}_{q_n}, xv + \text{Multi}_{q_n}) \quad (13) \\ x \in \{r, i, j, k\} \end{aligned}$$

3.4. Multi-layer feature relationship learning

For self-attention or co attention, only the same level of the attention map can be established. In order to establish the relationship between text modality and visual modal multi-layer features, we use the quaternion product method to learn this relationship and apply this relationship to Long interest in the attention.

$$W = rw + iw \vec{I} + jw \vec{J} + kw \vec{K} \quad (14)$$

$$V = rv + iv \vec{I} + jv \vec{J} + kv \vec{K} \quad (15)$$

$$\begin{aligned} W \otimes V = (rv * rw - iv * iw - jv * jw - kv * kw) \\ + (iv * rw + rv * iw - kv * jw + jv * kw) \vec{I} \\ + (jv * rw + kv * iw + rv * jw - iv * kw) \vec{J} \\ + (kv * rw - jv * iw + iv * jw + rv * kw) \vec{K} \quad (16) \end{aligned}$$

We separate the real and imaginary parts of the Hamilton Product's results; the real layer is represented by r , i represents the first imaginary layer relationship, j represents the second imaginary layer relationship, and k represents the third imaginary layer relationship. We apply these relationships to information updates between the image and text modalities [26].

$$\text{quaternion_score}_r = rv * rw - iv * iw - jv * jw - kv * kw \quad (17)$$

$$\text{quaternion_score}_i = iv * rw + rv * iw - kv * jw + jv * kw \quad (18)$$

$$\text{quaternion_score}_j = jv * rw + kv * iw + rv * jw - iv * kw \quad (19)$$

$$\text{quaternion_score}_k = kv * rw - jv * iw + iv * jw + rv * kw \quad (20)$$

$$\begin{aligned} \text{Quaternion_softmax}_{(r,i,j,k)} \\ = \text{Softmax}\left(\text{quaternion_score}_{(r,i,j,k)}\right) \quad (21) \end{aligned}$$

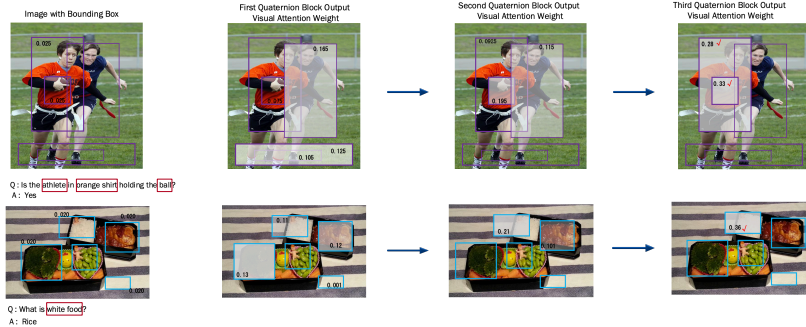


Fig. 5. Visualization of the updating process of attention weights learned by our Quaternion Block Networks. If the weight of an object is higher, the color of that bounding box will be more “gray”.

Model	test-dev				test-std
	Y/N	No.	Other	All	All
Bottom-up [25]	81.82	44.21	56.05	65.32	65.67
MFH [6]	-	-	-	66.12	-
DCN [27]	83.51	46.61	57.26	66.87	66.97
MFH+Bottom-Up [6]	84.27	49.56	59.89	68.76	-
BAN+Glove [28]	85.46	50.66	60.50	69.66	-
DFAF [4]	86.09	<u>53.32</u>	60.49	70.22	70.34
DFAF-Bert [4]	86.73	52.92	61.04	70.59	70.81
MCAN [5]	<u>86.8</u>	53.26	60.72	<u>70.63</u>	70.90
MLIN w/o Bert [19]	85.96	52.93	60.40	70.18	70.28
ViLBERT [29]	-	-	-	70.55	<u>70.92</u>
QBN (ours)	87.12	53.48	<u>60.8</u>	70.81	71.02

Table 1. Experimental results on VQA 2.0 test-dev and test-std splits. In each sub-category, the best number is labeled as bold while the second largest number is underlined. Overall, our model achieves the best performance against previous strong baselines including pretraining-based approach ViLBERT.

4. EXPERIMENTS AND DISCUSSIONS

The evaluation results comparing with the conventional approaches are shown in Table 1. Table 2 shows the ablation study on the VQA 2.0 validation dataset to examine the performance by combining each module. we tested our QBN on the test dataset by combining the VG dataset, the train and test dataset of VQA v2.0. All experiments use ADAM optimizer with 0.0001 learning rate for 13 epochs.

We mainly explore the effects of Multi-layer content learning, Multi-relationship content, dynamic sentence vector features effects on visual channel features, number of heads, number of stacks, and visual question-and-answer models in multi-layer content.

Table 2 shows the performance is improved after stacking the models, such as model BAN, DFAF, MCAN. This is because multi-level features are used to match the attention mechanism, which is helpful for improving the power of the model. The DFAF is a representative model in which the alignment levels are updated with each other. As shown in Fig.1, MCAN is a representative model that uses high-level information to match; Shown in Figure 3, our model first understands the multi-layer content of the text features in a quaternion-block, and obtains a representative multi-layer feature $Multi_{qn}$, and then adds this feature, $Multi_{qn}$, to the text information of each layer; When the image feature uses co-attention for information interaction, each layer feature of the text can be considered, and the corresponding layer feature is selected in the case of the character high-level feature. And in combination with the Quaternion-block gate, the model can achieve better results, because

the $quaternion_{score}$ in the quaternion-block gate can reflect the effective relationship between the features of other layers in the form of correlation coefficients, and the content of each layer. Importance, so that the image and the fine-grained head feature space in the text establish a more efficient relationship. The Scaled 2x2 region feature can bring richer visual information, which helps the performance.

Component	Setting	Accuracy
Bottom-up [25]	Bottom-up	63.37
Bilinear	BAN-1	65.36
Attention [28]	BAN-12	66.04
DFAF [4]	DFAF-1	66.21
	DFAF-8	66.66
MCAN [5]	MCAN-9	67.16
# of Stacked Blocks	QB-3	67.20
	QB-4	67.35
Embedding Dimension	512	67.45
Region feature	$1 \times 1 \times 2048$	67.35
	Scaled $2 \times 2 \times 2048$	67.45
W/o Relationship Learning	QB-4	67.28
W/o Content Learning	QB-4	67.25
Parallel Heads	8 heads	67.26
	12 heads	67.30
	16 heads	67.35

Table 2. Ablation studies of Quaternion Block Networks on VQA v2.0 validation dataset. The N in QB-N represents the number of a Quaternion block.

5. CONCLUSION

Our proposed Multi-Layer Content Interaction can be a fundamental approach to capture high-level interaction between different modalities. In this paper, the performance on Visual Question Answering can even surpass models with BERT pre-training. In the future, we will test our model on more diverse applications like visual grounding and vision language pre-training.

Acknowledgement

This work was supported in part by National key research and development program of China(2016QY01W0200), The Foundation for Innovative Research Groups of the National Natural Science Foundation of China(Grant No. 61921003). The work of C. Hori was fully supported by MERL.

6. REFERENCES

- [1] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi, “Attention-based multimodal fusion for video description,” in *ICCV*, 2017.
- [2] Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven CH Hoi, and Xiaogang Wang, “Question-guided hybrid convolution for visual question answering,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 469–485.
- [3] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al., “End-to-end audio visual scene-aware dialog using multimodal attention-based video features,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2352–2356.
- [4] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li, “Dynamic fusion with intra- and inter-modality attention flow for visual question answering,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [5] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, “Deep modular co-attention networks for visual question answering,” in *CVPR*, 2019.
- [6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” in *CVPR*, 2017.
- [7] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *ICLR*, 2015.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.
- [9] Jie Hu, Li Shen, and Gang Sun, “Squeeze-and-excitation networks,” in *CVPR*, 2018.
- [10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, “Densely connected convolutional networks,” in *CVPR*, 2017.
- [11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, “Distributed representations of words and phrases and their compositionality,” in *NIPS*, 2013.
- [12] Jeffrey Pennington, Richard Socher, and Christopher Manning, “Glove: Global vectors for word representation,” in *EMNLP*, 2014.
- [13] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, “Skip-thought vectors,” in *NIPS*, 2015.
- [14] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, “Deep contextualized word representations,” *NAACL*, 2018.
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *NAACL*, 2019.
- [16] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., “Recurrent models of visual attention,” in *NIPS*, 2014.
- [17] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” *ICLR*, 2015.
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *NIPS*, 2017.
- [19] Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li, “Multi-modality latent interaction network for visual question answering,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [20] Dario Pavllo, David Grangier, and Michael Auli, “Quaternion: A quaternion-based recurrent model for human motion,” *BMVC*, 2018.
- [21] Titouan Parcollet, Ying Zhang, Mohamed Morchid, Chiheb Trabelsi, Georges Linarès, Renato De Mori, and Yoshua Bengio, “Quaternion convolutional neural networks for end-to-end automatic speech recognition,” *INTERSPEECH*, 2018.
- [22] Yi Tay, Aston Zhang, Luu Anh Tuan, Jinfeng Rao, Shuai Zhang, Shuohang Wang, Jie Fu, and Siu Cheung Hui, “Lightweight and efficient neural natural language processing with quaternion networks,” *ACL*, 2019.
- [23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015.
- [24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *IJCV*, 2017.
- [25] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei ZBothang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *CVPR*, 2018.
- [26] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, “Vqa: Visual question answering,” in *ICCV*, 2015.
- [27] Duy-Kien Nguyen and Takayuki Okatani, “Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering,” in *CVPR*, 2018.
- [28] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang, “Bilinear attention networks,” in *NIPS*, 2018.
- [29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” *arXiv*, 2019.