Neural Networks for Detecting Irrelevant Questions during Visual Question Answering

Mengdi Li, Cornelius Weber, and Stefan Wermter

University of Hamburg, Department of Informatics,
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
{mli, weber, wermter}@informatik.uni-hamburg.de
http://www.informatik.uni-hamburg.de/WTM

Abstract. Visual question answering (VQA) is a task to produce correct answers to questions about images. When given an irrelevant question to an image, existing models for VQA will still produce an answer rather than predict that the question is irrelevant. This situation shows that current VQA models do not truly understand images and questions. On the other hand, producing answers for irrelevant questions can be misleading in real-world application scenarios. To tackle this problem, we hypothesize that the abilities required for detecting irrelevant questions are similar to those required for answering questions. Based on this hypothesis, we study what performance a state-of-the-art VQA network can achieve when trained on irrelevant question detection. Then, we analyze the influences of reasoning and relational modeling on the task of irrelevant question detection. Our experimental results indicate that a VQA network trained on an irrelevant question detection dataset outperforms existing state-of-the-art methods by a big margin on the task of irrelevant question detection. Ablation studies show that explicit reasoning and relational modeling benefits irrelevant question detection. At last, we investigate a straight-forward idea of integrating the ability to detect irrelevant questions into VQA models by joint training with extended VQA data containing irrelevant cases. The results suggest that joint training has a negative impact on the model's performance on the VQA task, while the accuracy on relevance detection is maintained. In this paper we claim that an efficient neural network designed for VQA can achieve high accuracy on detecting relevance, however integrating the ability to detect relevance into a VQA model by joint training will lead to degradation of performance on the VQA task.

Keywords: Visual question answering · Irrelevant question detection · Multimodality · Deep neural networks.

1 Introduction

Visual question answering (VQA) [3] is an important multimodal task in the field of artificial intelligence in recent years. Given an image and a natural language question about the image, the task is to provide an accurate natural language

answer. This task has received significant interest from researchers because it not only can be utilized to examine the development of multimodal and crossmodal technologies [6], but also has great potentials in real-world application scenarios [8].

Despite significant progress in recent years, the majority of conducted research focuses on improving accuracy on current hand-curated VQA datasets [3, 7, 10], in most of which questions are relevant to corresponding images by default. When given an irrelevant question to an image, current state-of-the-art models would still produce an answer with a high probability score rather than predict that the question is irrelevant and cannot be answered correctly. Obviously, it is not what we expect for an intelligent VQA system. On the one hand, this situation shows that current VQA models do not truly understand visual information of images and what questions are asking about. On the other hand, producing answers to irrelevant questions would be a harm to user experience and mislead users by conveying misinformation that premises in questions are all correct.

More formally, irrelevant questions in the context of VQA can be defined by premises [15], which are facts implied by questions. For instance, the question "What's the black cat on the table doing?" implies the presence of a black cat, a table, and that the cat is on the table. Mahendru et al. [15] categories premises into three classes of order. The first-order premises mean the presence of objects (e.g. a cat). The second-order premises reflect attributes of objects (e.g. a black cat) and the third-order premises are about relations and interactions between objects (e.g. a cat on a table). Once there is at least one false premise in a question, the question should be classified as an irrelevant question to the paired image. In the previous example, if there is a dog instead of a cat, or the cat is under the table in the image, the question is irrelevant to the image. In this case, if a VQA model still gives an answer like "sleeping", misinformation that there is "a black cat on the table" in the image would be conveyed to the asker.

Current approaches treat the VQA task as a multiclass classification problem. Given a question $q \in Q$ and an image $v \in V$, a VQA model is expected to give the ground truth answer $a^* \in A$ with the highest classification score

$$\hat{a} = \underset{a \in A}{\operatorname{argmax}}\, p_\theta(a|v, q), \tag{1}$$

where $\hat{a}$ is the predicted answer, and θ are the parameters of the trained model. The task of irrelevant question detection can be defined as a binary classification task. For a question-image pair (q, v), the task is to classify whether the question q is relevant to the image v.

Works of Ray et al.[16] and Mahendru et al. [15] are most related to ours. Ray et al. [16] firstly introduce the problem of irrelevant question detection in the context of VQA. They construct a dataset named VTFQ (Visual True and False Question) by showing annotators with images paired with randomly selected questions and asking them to annotate whether the question is relevant to the corresponding image or not. Mahendru et al. [15] propose a premise extraction pipeline to automatically extract premise information from questions. In their

paper, they give a formal definition of question premises and classify premises into mentioned three orders according to their complexity. A new dataset named QRPE (Question Relevance Prediction and Explanation) is constructed by them for the task of irrelevant question detection based on premises of questions. This dataset encompasses more and ambiguous examples in comparison to the VTFQ dataset, which makes it more challenging. Several different methods have been proposed for detecting question relevance in these works. Their experimental results indicate that methods based on image captioning models have the best performance on this task. Thought both of these papers briefly mention the benefits of integrating relevance detection to existing VQA systems, less attention has been devoted to relations between the relevance detection task and the VQA task.

In this paper, we have a hypothesis that the abilities required for detecting irrelevant questions are similar to those required for answering visual questions. In contrast to answering visual questions, judging whether a question is relevant to an image also requires a model to have a thorough and comprehensive understanding of both images and questions. To achieve this task, a model has to acquire information about classes of objects, colors, relative locations, counts, etc. Based on this hypothesis, using an end-to-end network architecture designed for the VQA task for detecting irrelevant questions is a more natural approach, in contrast to existing best-performing methods [15, 16] which utilize separated image captioning models and MLP networks. In this work, we investigate the possibility of solving the task of irrelevant question detection with a neural network designed for the VQA task.

To integrate the ability to detect irrelevant questions into a VQA model, a straight-forward idea is training a VQA model jointly on a dataset containing both relevant cases and irrelevant cases by treating answers of irrelevant cases as "irrelevant". However, interference between these two tasks is still unclear when jointly training them together. Therefore we conduct several experiments to investigate this issue. We expect the performance of the joint model on both of two tasks could be boosted based on our hypothesis. Our main contributions are as follows:

1. We find that the task of irrelevant question detection could be solved well by a neural network designed for the VQA task.
2. We set a new baseline accuracy on the QRPE dataset.
3. We find that the task of irrelevant question detection benefits from iterative reasoning and relational modeling.
4. We find that jointly training a VQA model on extended VQA data containing irrelevant cases impairs the accuracy on the VQA task while the performance on the task of irrelevant question detection is maintained.

2 Model

We choose the Multimodal Relational reasoning (MuRel) network [5], one of the current state-of-the-art models on the VQA task, as our basic model. Explicit it-

erative reasoning and relational modeling abilities distinguish MuRel from other networks. Two components associated with iterative reasoning and relational modeling in the network are the MuRel cell and the pairwise module. It has been shown that visual features fed into a VQA model play an important role in VQA performance [2, 9]. MuRel uses the Bottom-up features [2] to represent images. An object detector Faster R-CNN [17] is used to extract region feature vectors to generate the Bottom-up features of images. A pretrained skip-thought encoder [12] is used for the question features extraction.

MuRel cell takes visual features and question features as inputs and produces updated visual features. The MuRel cell could be invoked several times to update visual features interactively. A pairwise module is an element of the MuRel cell. It obtains region features and coordinates of regions to model relations between them. An efficient bilinear fusion module [4] is used as the multimodal fusion strategy to combine visual and language information. A running process of the MuRel cell with the pairwise module is formalized as

$$\{s_i^t\} = MuRelCell(\{s_i^{t-1}\}, \{b_i\}, q), \tag{2}$$

where $t \in \{1, ..., T\}$ is the step number of the current process, s_i^t represents the updated representation of region i, b_i is the coordinate of region i and q is the representation of the input question. In the first step of the process (when $t = 1$), $s_i^0 = v_i$ exists, where v_i is the feature of region i of the visual features provided by the Bottom-up features. After the last step of this process, when $t = T$, all s_i^T are aggregated together to provide a single vector s, which is then fused with question features q to produce a probability distribution $\hat{y}$ over all possible answers. This process can be formalized as

$$\hat{y} = B(s, q, \Theta_c), \tag{3}$$

where Θ_c are trainable parameters of the classifier.

We term MuRel2 the MuRel relational reasoning network trained for relevance detection with a binary classifier. The network architecture of MuRel2 is illustrated in Figure. 1. The network is applied to the task of irrelevant question detection. Inputs of MuRel2 are Bottom-up features of images and question features extracted by a skip-thought encoder. Labels of "relevant" and "irrelevant" in irrelevant question detection datasets are treated as two answers and correspond to two output neurons. Cross entropy loss is calculated to supervise the learning process.

3 Validation of MuRel2

3.1 Dataset

We use the QRPE dataset [1] [15] to evaluate the MuRel2 model and compare it against other approaches on the task of irrelevant question detection. The

[1] https://virajprabhu.github.io/qpremise/dataset/

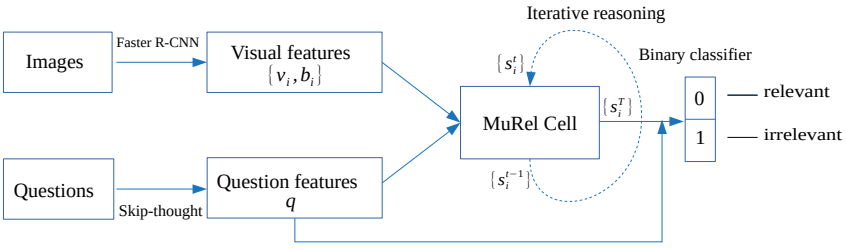

Fig. 1. Illustration of the network architecture of MuRel2.

QRPE dataset is curated automatically based on the MSCOCO [14], Visual Genome [13] and VQA v2 dataset [7]. First-order and second-order premises are firstly extracted from questions through a semantic tuple extraction pipeline used in the SPICE metric [1] for evaluating the quality of image captions. For first-order premises, irrelevant images for a question are selected by checking the absence of the appropriate class label in the MSCOCO annotations. For second-order premises, images that contain a matching object but a different attribute to the question premise according to annotations of Visual Genome are determined as irrelevant images. To ensure that the irrelevant image is similar enough to the relevant image, the one with the closest visual distance to the relevant image has been selected from irrelevant candidate images. In the end, every question in the QRPE dataset is paired with a relevant image and an irrelevant image. Compared to the VTFQ dataset, which is the first dataset for the task of irrelevant question detection, the QRPE dataset is balanced in the label space, larger and constructed in finer granularity.

The training set of the QRPE dataset contains 35,486 irrelevant question-image pairs which are generated from the training set of the VQA v2 dataset. The test set of the QRPE dataset contains 18,425 irrelevant question-image pairs which are generated from the validation set of the VQA v2 dataset. Based on the order of the false premise, irrelevant cases can be divided into a first-order part and a second-order part. The number of irrelevant question-image pairs in the QRPE dataset is shown in Table 1.

Table 1. Number of irrelevant question-image pairs in the QRPE dataset.

Split	Overall	First-order	Second-order
Training set	35,486	32,939	2,547
Test set	18,425	17,096	1,329

3.2 Experimental setup

Matching the experimental setup of existing methods we compare, we randomly select 90% of the training set of the QRPE dataset for training and the rest for validation. To avoid bias resulting from random division, we train 5 models independently and report the average accuracy of them on the test set as final results. All MuRel2 models are trained from scratch on the QRPE dataset. We performed some preliminary study for training strategy and critical hyperparameters. We observed that overfitting problems can easily arise when inappropriate learning rates applied. Finally, a similar learning scheduler as [5] with different settings is used in our training. We begin with a learning rate of $5e-6$, linearly increasing it at each epoch till it reaches $2e-5$ at epoch 6. Then we decrease the learning rate by a factor of 0.25 every 2 epochs from epoch 8 to epoch 14, at which we stop training. In our experiments, the batch size is set to 80, and experiments are conducted on 2 $\times$ NVIDIA Geforce 1080 TI.

3.3 Comparison to state-of-the-art approaches

We compare MuRel2 against state-of-the-art approaches on the QRPE dataset. The goal of this experiment is to evaluate whether a well-performing network designed for the VQA task can solve the task of irrelevant question detection well.

QC-Sim, PC-Sim, and QPC-Sim [15] are existing best-performing approaches on the QRPE dataset. QC-Sim uses an image captioning model NeuralTalk2 [11] pretrained on the MSCOCO dataset to automatically generate natural language descriptions for images. An LSTM network is used to encode both the generated image captions and corresponding questions into vector representations. Then, question and caption representations are concatenated and fed into an MLP network to predict the relevance between questions and images. PC-Sim and QPC-Sim are variants of QC-Sim. PC-Sim uses automatically generated image captions and premises extracted from questions for relevance prediction. QPC-Sim considers all the three sources, including questions, premises, and captions, for relevance prediction, and achieved the highest overall accuracy.

Results of MuRel2 in Table 2 are achieved when the number of reasoning steps is set to 3 and the pairwise module is not used. Figure. 2 shows the training curves of MuRel2 under this setting. The figure indicates that the model converges soon. After 8 epochs, an evaluation accuracy of around 90% is reached. We can notice that from epoch 8 on, the evaluation loss starts to increase slightly, which indicates that the model tends to overfit. A comparison of accuracy between MuRel2 and other approaches on the overall and two splits of the test set of the QRPE dataset is shown in Table 2.

Results of QC-Sim, PC-Sim, and QPC-Sim are reported in their original paper [15]. The accuracy on the test set would be 50% if chosen at random since every question in the test set of QRPE is paired with a relevant and an irrelevant image. From Table 2, we can see that MuRel2 outperforms existing best performing approaches by a big margin (over 10%) both on the overall

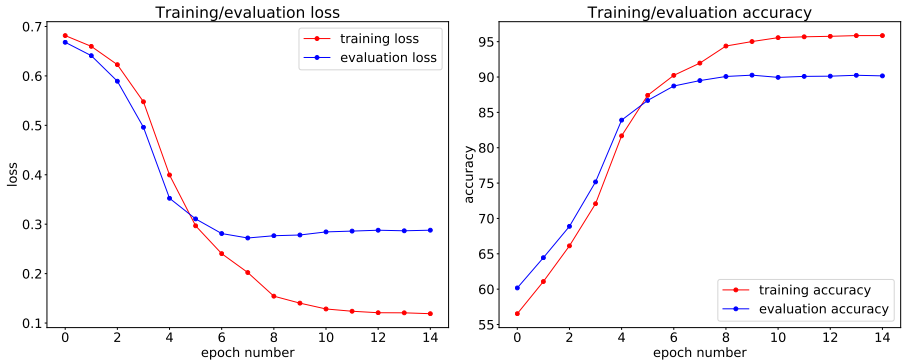

Fig. 2. Training curves of MuRel2 on the QRPE dataset.

Table 2. Comparison of accuracies on the QRPE dataset.

Models	Overall	First-order	Second-order
QC-Sim	74.35	75.82	55.12
PC-Sim	75.05	76.47	56.04
QPC-Sim	75.35	76.67	55.95
MuRel2	**86.62**	**88.13**	**67.02**

test set and each split divided according to the order of false premises. We can conclude from this experiment that a network architecture designed for the VQA task can solve the task of irrelevant question detection well.

3.4 Ablation study

In this part, we investigate the effects of multi-step reasoning and relational modeling on irrelevant question detection. Their contributions to the VQA task have been well proven [5]. In Table 3, we compare four MuRel2 models with different settings. To ensure comparability, we train them following the same experimental setup. The setting "Pairwise" means whether the pairwise module is used and the setting "Iter." means whether iterative reasoning is used. In our experiments, the number of reasoning steps is set to 3 when iterative reasoning is used.

The results in Table 3 show that a MuRel2 model with iterative reasoning but without the pairwise relational module achieves the best overall performance and the highest accuracies on both the first-order and second-order part. The first three rows of Table 3 show that both iterative reasoning and relational modeling contribute to MuRel2 network's performance on the QRPE dataset, which is consistent with their benefits on the VQA task. However, comparing row 3 and row 4, we find adding the pairwise module to a model with iterative reasoning results in a loss of accuracy. We observed that the distance between

training and evaluation loss curves increases when the pairwise module is used in this case, thus a possible explanation for this situation is using the iterative reasoning process and the pairwise module together leads to overfitting.

Table 3. Accuracies in the ablation study of MuRel2.

Pairwise	Iter.	Overall	First-order	Second-order
✗	✗	85.64	87.20	65.69
✓	✗	86.15	87.84	64.35
✗	✓	**86.62**	**88.13**	**67.02**
✓	✓	86.16	87.72	66.27

4 Joint training

In this part, we investigate the idea of integrating the ability to detect irrelevant questions into a VQA model by joint training a VQA model on a training set containing also irrelevant cases. For handling irrelevant cases, the model treats answers of irrelevant cases as a special answer "irrelevant". Based on our hypothesis that abilities required for detecting irrelevant questions are similar to those required for answering visual questions, we expect that training data for these two tasks could benefit each other by joint training. The approach to joint training the MuRel network on an extended VQA dataset containing irrelevant cases is illustrated in Figure. 3.

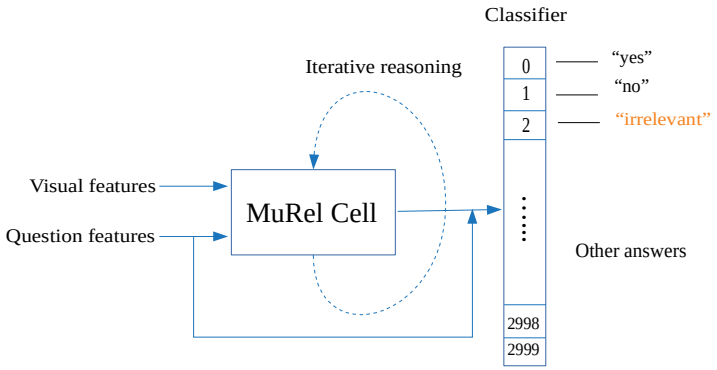

Fig. 3. Illustration of the approach to joint training the MuRel network on an extended VQA dataset containing irrelevant cases.

4.1 Dataset

In extended training sets, irrelevant cases are annotated with answer "irrelevant" for fitting VQA networks. In our experiments, we construct extended training sets based on the VQA v2 dataset, which is the most widely used VQA dataset. The VQA v2 dataset contains 443K, 214K, and 453K question-image pairs for training, evaluation, and testing respectively. We denote the training set of the VQA v2 dataset as $VQAv2$ in our experiments. We assume that all questions in the VQA v2 dataset are relevant to their corresponding images since human annotators are instructed to ask questions about the image that can be answered. First, we add 90% of irrelevant question-image pairs in the training set of the QRPE dataset to $VQAv2$ to build an extended training set $VQAv2 + QRPE$. The reason why we only select 90% of irrelevant cases is to match the training setting in Section 3 for fair comparisons on the test set of the QRPE dataset. In $VQAv2{+}QRPE$, irrelevant cases account for 6.7% of all cases. To investigate the impact of different proportions of irrelevant cases, we construct another training set by adding all irrelevant cases in the training set of both the QRPE dataset and the VTFQ dataset. We denote this training set as $VQAv2{+}QRPE{+}VTFQ$, of which irrelevant cases account for 9.0%.

For $VQAv2$, 3000 most frequent answers are selected as candidate answers. The top two most frequent answers are "yes" and "no", both of which occur over 80K times in the training set. Following them are answers "1" and "2", both of which occur over 10K times.

For $VQAv2 + QRPE$ and $VQAv2 + QRPE + VTFQ$, the special answer "irrelevant" is included in the 3000 candidate answers. In these two training sets, counts of the answer "irrelevant" are 31938 and 44024 respectively, which matches the numbers of irrelevant cases in them. Thus, in both of these two training sets, the answer "irrelevant" ranks between "no" and "1". The count of answer "irrelevant" is about half the count of answer "yes" and in the same order of magnitude as some other frequent answers.

4.2 Experimental setup

For experiments on joint training, we use a MuRel network with a pairwise module and a 3-step iterative reasoning process, because this setting achieves the best performance on the VQA v2 dataset. We adopt the same learning schedulers with the original MuRel model [5] trained on the VQA v2 dataset. The starting learning rate is set to $1.5e-4$ with a batch size of 160. Models are trained for 25 epochs. Our experiments are conducted on 4 $\times$ NVIDIA Geforce 1080 TI. We train all models on different training sets following the same experimental setup to ensure comparability.

4.3 Results

Three MuRel models are trained on $VQAv2$, $VQAv2 + QRPE$, $VQAv2 + QRPE + VTFQ$ respectively and evaluated on the validation set of the VQA

v2 dataset at every epoch. Checkpoints with the highest top 1 accuracy on the validation set are selected and tested on the *test-dev* split of the VQA v2 dataset for comparison. Scores of accuracy in Table 4 are calculated by the evaluation metric of the VQA Challenge [2] for all questions, "yes/no" questions, "number" questions, and other questions that are neither answered "yes/no" nor number.

Table 4. Resulting accuracies on the *test-dev* split of the VQA v2 dataset after joint training on different training sets

Training set	Yes/No	Num.	Other	All
$VQAv2$	82.70	48.32	**56.13**	**66.19**
$VQAv2 + QRPE$	**83.03**	47.95	54.79	65.64
$VQAv2 + QRPE + VTFQ$	82.91	**48.35**	54.69	65.59

From the accuracies shown in Table 4, we derive that jointly training a VQA model on training sets containing also irrelevant cases has a negative impact on its overall performance on the normal VQA data. As the proportion of irrelevant cases increases, the overall accuracy gradually decreases. We notice that the accuracy of "yes/no" questions can be improved when training on extended training sets.

We also test the MuRel model trained on $VQAv2 + QRPE$ on the test set of the QRPE dataset to see the impacts of joint training on the task of irrelevant question detection. To get the accuracy of this model on an irrelevant question detection dataset, we treat the answer of "irrelevant" as a prediction of irrelevance and other answers as a prediction of relevance. For a fair comparison, we take the same checkpoint that produces in scores in Table 4 for testing on the QRPE test set. The overall accuracy achieved by this MuRel model on the test set of QRPE is 86.24%. This accuracy is a bit higher than the accuracy of 86.16% achieved by the MuRel2 model with the same setting shown in row 4 of Table 3. It shows that joint training can maintain accuracy on the task of irrelevant question detection well.

To avoid degradation on the VQA task when jointly training a model on a training set containing data for both VQA and relevance detection, we would like to suggest an alternative architecture. In this architecture, network layers for processing features of images and questions are shared for two tasks, while the output layers are separated. When the network is trained on irrelevant cases, parameters in output layers for the VQA task are not updated. This separation procedure might avoid unexpected interference of those tasks and reduce the overfitting problem.

[2] https://visualqa.org/evaluation.html

5 Conclusion

In this paper, we investigate networks designed for VQA on the task of irrelevant question detection. A multimodal relational network for VQA is used for experiments. We demonstrate that the network adapted as a binary classifier outperforms the existing state-of-the-art methods by a large margin on the task of irrelevant question detection. From the ablation study, we derive that the relevance prediction task has the same requirement for the reasoning ability and relational modeling ability as the VQA task has. We also investigate the idea of integrating the ability to detect irrelevant questions into a VQA model by training a VQA model on a training set containing also irrelevant cases. It is interesting that joint training leads to degradation of performance on the normal VQA data, while the accuracy on the task of irrelevant question detection is maintained compared with models trained for each specific task.

Future work may include building a larger and more difficult dataset for the task of irrelevant question detection. Though compared with the VTFQ dataset, the QRPE dataset is collected in a finer granularity by concerning different orders of false premises, it only contains irrelevant questions with false first-order and second-order premises and ignores irrelevant cases with false third-order premises concerning relations and interactions between objects. That makes current datasets unsuitable for the true challenges of the relevance detection task. In addition to the dataset building necessity, it is also promising to study methods of improving models' performance on the VQA task by taking advantage of the task of irrelevant question detection and vice versa. While we observed that jointly training a model on extended datasets containing also irrelevant cases leads to degradation of accuracy on the VQA task, we hypothesize that it may be possible to improve the performance by using other training methods, such as the method mentioned that shared layers are trained jointly while output layers are trained separately.

Acknowledgement

We gratefully acknowledge support from the China Scholarship Council (CSC) and the German Research Foundation (DFG) under project Crossmodal Learning (TRR 169).

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic propositional image caption evaluation. In: European Conference on Computer Vision. pp. 382–398. Springer (2016)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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. pp. 6077–6086 (2018)

3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
4. Ben-Younes, H., Cadene, R., Thome, N., Cord, M.: BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8102–8109 (2019)
5. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: MUREL: Multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1989–1998 (2019)
6. Fu, D., Weber, C., Yang, G., Kerzel, M., Nan, W., Barros, P., Wu, H., Liu, X., Wermter, S.: What can computational models learn from human selective attention? A review from an audiovisual unimodal and crossmodal perspective. Frontiers in Integrative Neuroscience **14**, 10 (2020)
7. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
8. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: VizWiz Grand Challenge: Answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3617 (2018)
9. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0. 1: The winning entry to the VQA challenge 2018. arXiv preprint arXiv:1807.09956 (2018)
10. Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1965–1973 (2017)
11. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3128–3137 (2015)
12. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in Neural Information Processing Systems. pp. 3294–3302 (2015)
13. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
15. Mahendru, A., Prabhu, V., Mohapatra, A., Batra, D., Lee, S.: The Promise of Premise: Harnessing question premises in visual question answering. arXiv preprint arXiv:1705.00601 (2017)
16. Ray, A., Christie, G., Bansal, M., Batra, D., Parikh, D.: Question Relevance in VQA: Identifying non-visual and false-premise questions. arXiv preprint arXiv:1606.06622 (2016)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)