MUREL: Multimodal Relational Reasoning for Visual Question Answering

Remi Cadene [1*] Hedi Ben-younes [1,2*] Matthieu Cord [1] Nicolas Thome [3]
[1] Sorbonne Université, CNRS, LIP6, 4 place Jussieu, 75005 Paris
[2] Heuritech, 110 avenue de la République, 75011 Paris
[3] Conservatoire National des Arts et Métiers, 75003 Paris
remi.cadene@lip6.fr, hedi.ben-younes@lip6.fr, matthieu.cord@lip6.fr, nicolas.thome@cnam.fr

Abstract

Multimodal attentional networks are currently state-of-the-art models for Visual Question Answering (VQA) tasks involving real images. Although attention allows to focus on the visual content relevant to the question, this simple mechanism is arguably insufficient to model complex reasoning features required for VQA or other high-level tasks.

In this paper, we propose MuRel, a multimodal relational network which is learned end-to-end to reason over real images. Our first contribution is the introduction of the MuRel cell, an atomic reasoning primitive representing interactions between question and image regions by a rich vectorial representation, and modeling region relations with pairwise combinations. Secondly, we incorporate the cell into a full MuRel network, which progressively refines visual and question interactions, and can be leveraged to define visualization schemes finer than mere attention maps.

We validate the relevance of our approach with various ablation studies, and show its superiority to attention-based methods on three datasets: VQA 2.0, VQA-CP v2 and TDIUC. Our final MuRel network is competitive to or outperforms state-of-the-art results in this challenging context.

Our code is available: github.com/Cadene/murel.bootstrap.pytorch

1. Introduction

Since the success of Convolutional Neural Networks (ConvNets) at the ILSVRC 2012 challenge [29], Deep Learning has become the baseline approach for any computer vision problem. Beyond their outstanding performances for perception tasks, *e.g.* classification or detection [14], deep ConvNets have also been successfully used for new artificial intelligence tasks like Visual Question Answering (VQA) [4, 17, 23]. VQA requires a high level understanding of images and questions, and is often considered to be a good proxy for visual reasoning. However, it is not straightforward to use ConvNets in a context where a high level of reasoning is required. The question of leveraging the perception power of deep CNNs for reasoning tasks is crucial if we want to go further in visual scene understanding [21, 11].

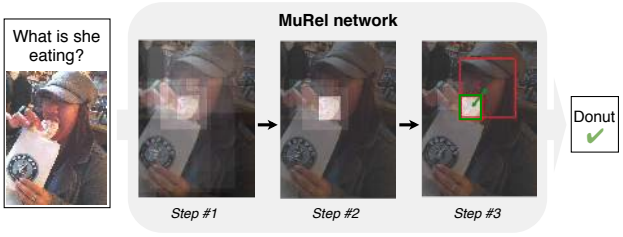

Figure 1. **Visualization of the MuRel approach.** Our MuRel network for VQA is an iterative process based on a rich vectorial representation between the question and visual information explicitly modeling pairwise region relations. MuRel is thus able to express complex analysis primitives beyond attention maps: here the two regions corresponding to the head and the donuts are selected based on their visual cues and semantic relations to properly answer the question "what is she eating?"

It is also not trivial to define nor evaluate a model's capacity to reason about the visual modality in VQA. To fill this need, synthetic datasets have been released, *e.g.* CLEVR [21], which specific structure controls the exact reasoning primitives required to give the answer [22, 19, 34]. However, methods that tackle the VQA problem on real data struggle to integrate this explicit reasoning procedure. Instead, state-of-the-art methods often rely on the much simpler attentional framework [16, 8, 25, 12]. Despite its effectiveness, this mechanism restricts visual reasoning to a soft selection of regions that are relevant to answer the question. This arguably limits the modeling power of such models to bridge the gap between the perceptual strengths of ConvNets and the high-level reasoning demand for VQA.

In this paper, we propose MuRel, a multimodal relational network that goes one step further towards reasoning about questions and images. Our first contribution is to introduce the MuRel cell, an atomic reasoning primitive enabling to

*Equal contribution

represent rich interactions between question and image regions. It is based on a vectorial representation that explicitly models relations between regions. Our second contribution is to embed this MuRel cell into an iterative reasoning process, which progressively refines the internal network representation to answer the question. The rationale of MuRel is illustrated in Figure 1: for the question "what is she eating", our model focuses on two main regions (the head and the donut) with important visual cues and semantic relations between them to provide the correct answer ("donut"). The visual reasoning of our MuRel system is formed by this multi-step relational module that discards useless information to focus on the relevant regions.

In the experiments, we show additional results for explaining the behaviour of MuRel. We also provide various ablative studies to validate the relevance of the MuRel cell and the iterative reasoning process, and show that MuRel is highly competitive or even outperforms state-of-the-art results on three of the most common VQA datasets: the VQA 2.0 dataset [17], VQA-CP v2 [1] and TDIUC [23].

2. Related work and contributions

Recently, the deep learning community started to tackle complex visual reasoning problems such as relationship detection [31], object recognition [11], multimodal retrieval [10, 15], abstract reasoning [38], visual causality [30], or visual dialog [13, 48], while more theoretical work attempt to formalize relational reasoning [7].

But the most popular image reasoning task is certainly Visual Question Answering (VQA), which has been a hot research topic for the last five years [33, 4, 17, 23]. Since the seminal work of [33], different sub-problems have been identified for the resolution of VQA. In particular, explicit reasoning techniques have been developed relying on synthetic datasets [21, 41]. Meanwhile, real-data VQA systems are the test bed for more practical approaches based on high quality visual representations or multimodal fusion schemes.

Visual reasoning The research efforts towards VQA models that are able to reason about a visual scene is mainly conducted using the CLEVR dataset [21]. This artificial dataset provides questions that require spatial and relational reasoning on simple images coming from a visual world with low variability. An important line of work attempts to solve this task through explicit reasoning. In such methods [22, 19, 34], a neural network reads the question and generates a program, corresponding to a graph of elementary neural operations that process the image. However, there are two major downsides to these techniques. First, their performance strongly depends on whether or not program annotations are used to learn the program generator; and second, they can be matched or surpassed by simpler models that implicitly learn to reason without requiring program annotation. In particular, FiLM [37] modulates the visual feature map with an affine transformation whose parameters depend on the question. In more recent work, the MAC network [20] draws inspiration from the Model-View-Controller paradigm to design the trainable MAC cell on which the network iterates. Finally, in [39], they reason over all the possible pairs of objects in the picture, thus introducing relationship modeling in visual question answering.

VQA on real data An important part of the research in VQA is focused on designing functions that can represent high-level correlations between two vector spaces. Among these multimodal fusion algorithms, the most effective ones use second order (or higher [45]) interactions, made tractable through sketching methods [16], or with more success using the tensor decomposition framework [25, 8, 44].

This line of work is often considered orthogonal to visual reasoning contributions. In a setup involving real data, complex methods such as explicit or relational reasoning are much more challenging to implement than with artificial images and questions. This is certainly why the most widely used reasoning framework involves soft attention mechanisms [5, 42]. Given a question, these models assign an importance score to each region, and use them to weight-sum pool the visual representations. Multiple attention maps (also called *glimpses*) can even be computed in parallel [25, 8, 44, 45] or sequentially [43]. More complex attention strategies have been explored, such as the Structured Attention [12], where a locally-connected graphical structure is considered to infer the region saliency scores. [47] also leverages a graphical structure between regions to address weaknesses of the soft-attention mechanism, improving the VQA model's ability to count. In [36], the image representation is computed using pairwise semantic attention and spatial graph convolutions. The soft attention framework is questioned in [32], where regions are hardly selected based on the norm of their feature. Finally, recent work of [24] simultaneously attends over regions and word tokens through a bilinear attention network.

Importantly, the type of visual features used to feed the VQA system has an large impact on performance. While early work have been using fixed-grid representation given by a fully-convolutional network (such as ResNet-152 [18]), performance can be improved using predictions from an object detector [3]. Recently, a crucial component in the VQA Challenge 2018 winning entry was the mix of multiple types of visual features [46].

MuRel contributions In this work, we move away from the classical attention framework [25, 16, 8, 45] widely used in real-data VQA systems. Instead, we use a vectorial rep-

resentation, more expressive than scalar attention maps, to model the semantic interaction between each region's visual content and the question. In addition, we include a notion of spatial and semantic context in the representations by representing pairs of image regions through interactions between their visual embeddings and spatial coordinates. Differently than the approach followed in [36] where a locally connected graph structure is built, we use the relations between all possible pairs of regions.

Our MuRel network embodies an iterative process with inspiration from works driven by the synthetic reasoning CLEVR dataset, *e.g.*, MAC [20] or FiLM [37], which we adapt to the real data VQA purpose. In particular, we improve the interactions between image regions and questions by using richer bilinear fusion models and by explicitly incorporating relations between regions.

3. MuRel approach

Our VQA approach is depicted in Figure 3. Given an image $v \in \mathcal{I}$ and a question $q \in \mathcal{Q}$ about this image, we want to predict an answer $\hat{a} \in \mathcal{A}$ that matches the ground truth answer $a^\star$. As very common in VQA, the prediction $\hat{a}$ is given by classification scores:

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

where p_θ is our trainable model. In our system, the image is represented by a set of vectors $\{\boldsymbol{v}_i\}_{i\in[1,N]}$, where each $\boldsymbol{v}_i \in \mathbb{R}^{d_v}$ corresponds to an object detected in the picture. We also use the spatial coordinates of each region $\boldsymbol{b}_i = [x, y, w, h]$, where (x, y) are the coordinates of the top-left point of the box, and h and w correspond to the height and the width of the box. Note that x and w (respectively y and h) are normalized by the width (resp. height) of the image. For the question, we use a gated recurrent unit network to provide a sentence embedding $\boldsymbol{q} \in \mathbb{R}^{d_q}$.

In Section 3.1, we present the MuRel cell, a neural module that learns to perform elementary reasoning operations by blending question information into the set of spatially-grounded visual representations. Next, in Section 3.2, we leverage the power of this cell using the MuRel network, a VQA architecture that iterates through a MuRel cell to reason about the scene with respect to a question.

3.1. MuRel cell

The MuRel cell takes as input a bag of N visual features $\boldsymbol{s}_i \in \mathbb{R}^{d_v}$, along with their bounding box coordinates $\boldsymbol{b}_i$. As shown in Figure 2, it is a residual function consisting of two modules. First, an efficient bilinear fusion module merges question and region feature vectors to provide a local multimodal embedding. This fusion is directly followed by a pairwise modeling component, designed to update each multimodal representation with respect to its own spatial and visual context.

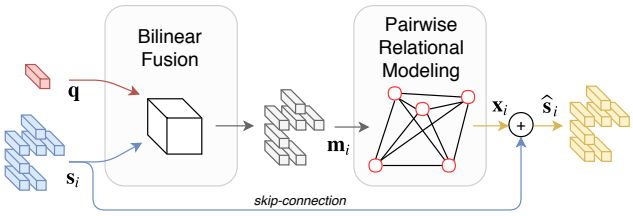

Figure 2. **MuRel cell.** In the MuRel cell, the bilinear fusion represents rich and fine-grained interactions between question and region vectors $\boldsymbol{q}$ and $\boldsymbol{s}_i$. All the resulting multimodal vectors $\boldsymbol{m}_i$ pass through a pairwise modeling block to provide a context-aware embedding $\boldsymbol{x}_i$ per region. The cell's output $\hat{\boldsymbol{s}}_i$ is finally computed as a sum between $\boldsymbol{s}_i$ and $\boldsymbol{x}_i$, acting as residual function of $\boldsymbol{s}_i$.

Multimodal fusion We want to include question information within each visual representation $\boldsymbol{s}_i$. Multiple multimodal fusion strategies have been recently proposed [25, 16, 8, 44, 45] to model the relevant interactions between two modalities. One of the most efficient technique is the one proposed by [9], based on the Tucker decomposition of third-order tensors. This bilinear fusion model learns to focus on the relevant correlations between input dimensions. It models rich and fine-grained multimodal interactions, while keeping a relatively low number of parameters. Each input vector $\boldsymbol{s}_i$ is fused with the question embedding $\boldsymbol{q}$ using the same bilinear fusion:

$$\boldsymbol{m}_i = \mathrm{B}\left(\boldsymbol{s}_i, \boldsymbol{q}; \Theta\right) \tag{2}$$

where Θ are the trainable parameters of the fusion module. Each dimension m of $\boldsymbol{m}_i$ can be written as a bilinear function in the form $\sum_{s,q} w^{s,q,m} \boldsymbol{s}_i^s \boldsymbol{q}^q$. Thanks to the Tucker decomposition, the tensor $\{w_{s,q,m}\}$ is factorized into the list of parameters Θ. We set the number of dimensions in $\boldsymbol{m}_i$ to d_v to facilitate the use of residual connections throughout our architecture.

In classical attention models, the fusion between image region and question features $\boldsymbol{s}$ and $\boldsymbol{q}$ only learns to encode whether a region is relevant. In the MuRel cell, the local multimodal information is represented within a richer vectorial form $\boldsymbol{m}_i$ which can encode more complex correlations between both modalities. This allows to store more specific information about what precise characteristic of a particular region is important in a given textual context.

Pairwise interactions To answer certain types of question, it can be necessary to reason over multiple object that interact together. More generally, we want each representation to be aware of the spatial and semantic context around it. Given that our features are structured as a bag of localized vectors [3], modeling the visual context of each region is not straightforward. Similarly to the recent work of [36], we opt for a pairwise relationship modeling where each region receives a message based on its relations to its neighbours. In their work, a region's neighbours correspond to

the K most similar regions, whereas in the MuRel cell the neighbourhood is composed of every region in the image. Besides, instead of using scalar pairwise attention and graph convolutions with Gaussian kernels as they do, we merge spatial and semantic representations to build relationship vectors. In particular, we compute a context vector $\check{\boldsymbol{e}}_i$ for every region. It consists in an aggregation of all the pairwise links $\boldsymbol{r}_{i,j}$ coming into i. We define it as $\check{\boldsymbol{e}}_i = \max_j \boldsymbol{r}_{i,j}$, where $\boldsymbol{r}_{i,j}$ is a vector containing information about the content of both regions, but also about their relative spatial positioning. We use the $\max$ operator in the aggregation function to reduce the noise that can be induced by average or sum poolings, which oblige all the regions to interact with each other. To encode the relationship vector, we use the following formulation:

$$\boldsymbol{r}_{i,j} = \mathrm{B}\left(\boldsymbol{b}_i, \boldsymbol{b}_j; \Theta_b\right) + \mathrm{B}\left(\boldsymbol{m}_i, \boldsymbol{m}_j; \Theta_m\right) \tag{3}$$

Through the $\mathrm{B}(.,.;\Theta_b)$ operator, the cell is free to learn spatial concepts such as *on top of, left, right, etc.* In parallel, $\mathrm{B}(.,.;\Theta_s)$ encodes correlations between multimodal vectors $(\boldsymbol{s}_i, \boldsymbol{s}_j)$, corresponding to semantic visual concepts conditioned on the question representation. By summing up both spatial and semantic fusions, the network can learn high-level relational concepts such as *wear, hold, etc.*

The context representation $\check{\boldsymbol{e}}_i$ is obtained by aggregating the representations $\boldsymbol{r}_{i,j}$ provided by its neighbours through an element-wise max pooling. Using this operator, the network can learn to filter our irrelevant interactions for each features dimension. Then, the multimodal vector $\boldsymbol{m}_i$ is updated in an additive manner:

$$\boldsymbol{x}_i = \boldsymbol{m}_i + \check{\boldsymbol{e}}_i \tag{4}$$

This formulation of the pairwise modelling is actually closer to the Graph Networks [7], where the notion of relational inductive biases is formalized.

Finally, the MuREL cell's output is computed as a residual function of its input, to avoid the vanishing gradient problem. Each visual feature $\boldsymbol{s}_i$ is updated as: $\hat{\boldsymbol{s}}_i = \boldsymbol{s}_i + \boldsymbol{x}_i$.

The chain of operations that updates the set of localized region embeddings $\{\boldsymbol{s}_i\}_{i\in[1,N]}$ using the multimodal fusion with $\boldsymbol{q}$ and the pairwise modeling operator is noted:

$$\{\hat{\boldsymbol{s}}_i\} = \mathrm{MurelCell}\left(\{\boldsymbol{s}_i\}; \{\boldsymbol{b}_i\}, \boldsymbol{q}\right) \tag{5}$$

3.2. MuRel network

The MuRel network mimics a simple form of iterative reasoning by leveraging the power of bilinear fusions to iteratively merge visual information into context-aware visual embeddings. As we can see in Figure 3, the region state vectors $\{\boldsymbol{s}_i\}$ are updated by a MuRel cell through multiple steps, each time refining the representations with contextual and question information. More specifically, for each step $t = 1..T$ where T is the total number of steps, a MuRel cell processes and updates the state vectors as follows:

$$\{\boldsymbol{s}_i^t\} = \mathrm{MurelCell}\left(\{\boldsymbol{s}_i^{t-1}\}; \{\boldsymbol{b}_i\}, \boldsymbol{q}\right) \tag{6}$$

The state vectors are initialized with the features outputted by the object detector; for each region i, $\boldsymbol{s}_i^0 = \boldsymbol{v}_i$.

The MuRel network represents each region regarding the question, but also using its own visual context. This representation is done iteratively, through multiple steps of a MuRel cell. The residual nature of this module makes it possible to align multiple cells without being subject to gradient vanishing. Moreover, the weights of our model are shared across the cells, which enables compact parametrization and good generalization.

At step $t = T$, the representations $\{\boldsymbol{s}_i^T\}$ are aggregated with a global max pooling operation to provide a single vector $\boldsymbol{s} \in \mathbb{R}^{d_v}$. This scene representation contains information about the objects, the spatial and semantic relations between them, with respect to a particular question.

The scene representation $\boldsymbol{s}$ is merged with the question embedding $\boldsymbol{q}$ to compute a score for every possible answer $\hat{\boldsymbol{y}} = \mathrm{B}\left(\boldsymbol{s}, \boldsymbol{q}; \Theta_y\right)$. Finally, $\hat{a}$ is the answer with maximum score in $\hat{\boldsymbol{y}}$.

Visualizing MuRel network Our model can also be leveraged to define visualization schemes finer than mere attention maps. Especially, we can highlight important relations between image regions for answering a specific question. At the end of the MuRel network, the visual features $\{\boldsymbol{s}_i^T\}$ are aggregated using a $\max$ operation, yielding a d_v-dimensional vector $\boldsymbol{s}$. Thus, we can compute a *contribution map* by measuring to what extent each region contributes to the final vector. To do so, we compute the point-wise $\boldsymbol{c} = \mathrm{argmax}_i\{\boldsymbol{s}_i^T\} \in [1, N]^{d_v}$, and measure the occurrence frequency of each region in this vector $\boldsymbol{c}$. This provides a value for each region that estimates its contribution to the final vector. Interestingly, this process can be done after each cell, and not exclusively at the last one. Intuitively, it measures what the contribution map would have been if the iterative process had stopped at this point. As we can see in Figures 1,3,5, these relevance scores match human intuition and can be used to explain the model's decision, even if the network has not been trained with any selection mechanism.

Similarly, we are able to visualize the pairwise relationships involved in the prediction of the MuRel cell. The first step is to find $i^\star$, which is the region that is the most impacted by the pairwise modeling. It is the region such that $\|\frac{\check{\boldsymbol{e}}_i}{\boldsymbol{x}_i}\|_2$ is maximal (cf. Equation (4)). This bounding box is shown in green in all our visualizations. We then measure the contribution of every other region to $i^\star$ using the occurrence frequencies in $\mathrm{argmax}_j \boldsymbol{r}_{i,j}$. We show in red the regions whose contribution to $i^\star$ is above a certain threshold (0.2 in our visualizations). If there is no such region, the green box is not shown.

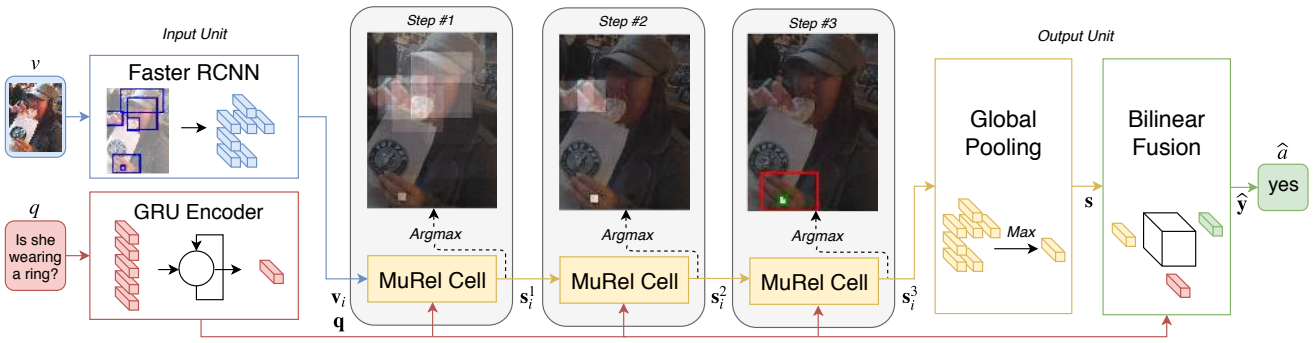

Figure 3. **MuRel network.** The MuRel network merges the question embedding $\boldsymbol{q}$ into spatially-grounded visual representations $\{\boldsymbol{v}_i\}$ by iterating through a single MuRel cell. This module takes as input a set of localized vectors $\{\boldsymbol{s}_i\}$ and updates their representation using a multimodal fusion component. Moreover, it models all the possible pairwise relations between regions by combining spatial and semantic information. To construct the importance map at step t, we count the number of time each region provides the maximal value of $\max_i\{\boldsymbol{s}_i^t\}$ (over the 2048 dimensions).

Connection to previous work We can draw a comparison between our MuRel network and the FiLM network proposed in [37]. Beyond the fact that their model is built for the synthetic CLEVR dataset [21] and ours processes real data, some connections can be found between both models. In their work, the image passes through multiple residual cells, whereas we only have one cell through which we iterate. In FiLM, the multimodal interaction is modeled with a feature-wise affine modulation, while we use a bilinear fusion strategy [8] which seems better suited to real world data. Finally, both MuRel and FiLM leverage the spatial structure of the image representation to model the relations between regions. In FiLM, the image is represented with a fully-convolutional network which outputs a feature map disposed in a fixed spatial grid. With this structure on image features, the relations between regions are modeled with a 3×3 convolution inside each residual block. Thus, the representation of each region depends on its neighbours in the locally-connected graph induced by the fixed grid structure. In our MuRel network, the image is represented as a set of localized features. This makes the relational modeling non trivial. As we want to model relations between regions that are potentially far apart, we consider that the set of regions forms a complete graph, where each region is connected to all the others.

4. Experiments

4.1. Experimental setup

Datasets: We validate the benefits of the MuRel cell and the MuRel network on three recent datasets. VQA 2.0 [17] is the most used dataset. It comes with a training set, a validation set and an online testing set. We provide a fine grained analysis on the validation set, while we compare MuRel to the state-of-the-art models on the testing set. Then, we use VQA Changing Priors v2 [2] to demonstrate the generalization capacity of MuRel. VQA-CP v2 uses the same data as in VQA 2.0, but proposes different distribution of answers per question between training and validation splits. Finally, we use the TDIUC dataset [23] to construct a more detailed analysis of our model's performance on 12 well-defined types of question. TDIUC is currently the biggest dataset for visual question answering.

Model	VQA 2.0	VQA CP v2	TDIUC
Attention baseline	63.44	38.04	86.96
MuRel	**65.14**	**39.54**	**88.20**

Table 1. **Comparing MuRel to Attention.** Comparison of the MuRel strategy against a strong Attention-based model on the VQA 2.0 *val*, VQA-CP v2 and TDIUC datasets. Both models have an equivalent number of parameters (∼60 million) and are trained on the same features following the same experimental setup.

Hyper-parameters: We use standard features extraction, preprocessings and loss function [16]. We use the recent Bottom-up features provided by [3] to represent our image as a set of 36 localized regions. For the question embedding, we use the pretrained Skip-thought encoder from [27]. Inspired by recent works, we use Adam as optimizer [26] with a learning scheduler [46]. More details about the experimental setup are given in appendix.

4.2. Model validation

We compare MuRel against models trained on the same Bottom-up features [3] which are required to reach the best performances.

Comparison to Attention-based model In Table 1, we compare MuRel against a strong attentional model based on bilinear fusions [8], which encompasses a multi-glimpses attentional process [16]. The goal of this experiments is to compare our approach with strong baselines for real VQA in controlled conditions. In addition to using the same bottom-

Pairwise	Iter.	VQA 2.0	VQA CP v2	TDIUC
✗	✗	64.13	38.88	87.50
✓	✗	64.57	39.12	87.86
✗	✓	64.72	39.37	87.92
✓	✓	**65.14**	**39.54**	**88.20**

Table 2. **Ablation study of MuRel**. Experimental validation of the pairwise module and the iterative processing on the VQA 2.0 *val*, VQA-CP v2 and TDIUC datasets.

up features, which are crucial for fair comparisons, we also dimension the attention-based baseline to have an equivalent amount of learned parameters than MuRel (∼60 millions including those from the GRU encoder). Also, we train it following the same experimental setup to insure competitiveness. MuRel reaches a higher accuracy on the three datasets. We report a significant gain of +1.70 on VQA 2.0 and +1.50 on VQA CP v2. Not only these results validate the ability of MuRel to better model interactions between the question and the image, but also to generalize when the distribution of the answers per question are completely different between the training and validation set as in VQA CP v2. A gain of +1.24 on TDIUC demonstrates the richer modeling capacity of MuRel in a fine-grained context of 12 well delimited question types.

Ablation study In Table 2, we compare three ablated instances of MuRel to its complete form. First, we validate the benefits of the pairwise module. Adding it to a vanilla MuRel without iterative process leads to higher accuracy on every datasets. In fact, between line 1 and 2, we report a gain of +0.44 on VQA 2.0, +0.24 on VQA CP v2 and +0.36 on TDIUC. Secondly, we validate the interest of the iterative process. Between line 1 et 3, we report a gain of +0.59 on VQA 2.0, +0.49 on VQA CP v2 and +0.42 on TDIUC. Notably, this modification does not add any parameters, because we iterate over a single MuRel cell. Unsharing the weights by using a different MuRel cell for each step gives similar results. Finally, the pairwise module and the iterative process are added to create the complete MuRel network. This instance (in line 4) reaches the highest accuracy on the three datasets. Interestingly, the gains provided by the combination of the two methods are sometimes larger than those of each one separately. For instance, we report a gain of +1.01 on VQA 2.0 between line 1 and 4. This attests to the complementary of the two modules.

Number of reasoning steps In Figure 4, we perform an analysis of the iterative process. We train four different MuRel networks on the VQA 2.0 *train* split, each with a different number of iterations over the MuRel cell. Performance is reported on *val* split. Networks with two and three steps respectively provides a gain of +0.30 and +0.57 in overall accuracy on VQA 2.0 over the network with a single step. An interesting aspect of the iterative process

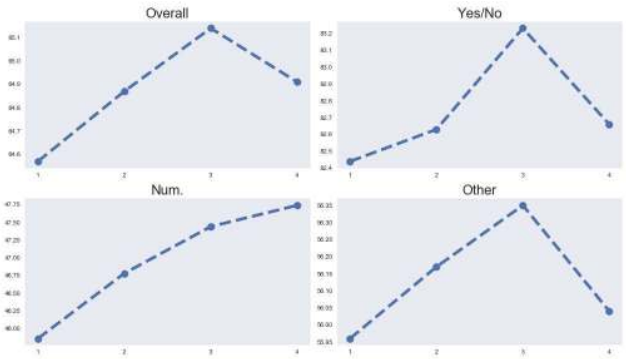

Figure 4. **Number of iterations**. Impact of the number of steps in the iterative process on the different question types of VQA 2.0 *val*.

of MuRel is that the four networks have exactly the same amount of parameters, but the accuracy significantly varies with respect to the number of steps. While the accuracy for the answer type involving numbers keeps increasing, we report a decrease in overall accuracy at four reasoning steps. Counting is a challenging task: not only does the model need to detect every occurrence of the desired object, but also the representation computed after the final aggregation must keep the information of the number of detected instances. The complexity of this question may require deeper relational modeling, and thus benefit from a higher number of iterations over the MuRel cell.

4.3. State of the art comparison

VQA 2.0 In Table 3, we compare MuRel to the most recent contributions on the VQA 2.0 dataset. For fairness considerations, all the scores correspond to models trained on the VQA 2.0 *train+val* split, using the Bottom-up visual features [3]. Interestingly, our model surpasses both MUTAN [8] and MLB [25], which correspond to some of the latest development in visual attention and bilinear models. This tends to indicate that VQA models can benefit from retaining local information in mulitmodal vectors instead of scalar coefficients. Moreover, our model greatly improves over the recent method proposed in [36] where the regions are structured using pairwise attention scores, which are leveraged through spatial graph convolutions. This shows the interest of our spatial-semantic pairwise modeling between all possible pairs of regions. Finally, even though we did not extensively tune the hyperparameters of our model, our overall score on the *test-dev* split is highly competitive with state-of-the-art methods. In particular, we are comparable to Pythia [46] who won the VQA Challenge 2018. Please note that they improve their overall scores up to 70.01% when they include multiple types of visual features and more training data. Also, we did not report the score of 69.52% obtained by BAN [24] as they train their model on extra data from the Visual Genome dataset [28].

Model	*test-dev* Yes/No	Num.	Other	All	*test-std* All
Bottom-up [3]	81.82	44.21	56.05	65.32	65.67
Graph Att. [36]	-	-	-	-	66.18
MUTAN† [8]	82.88	44.54	56.50	66.01	66.38
MLB† [25]	83.58	44.92	56.34	66.27	66.62
DA-NTN [6]	84.29	47.14	57.92	67.56	67.94
Pythia [46]	-	-	-	68.05	-
Counter [47]	83.14	**51.62**	**58.97**	**68.09**	**68.41**
MuRel	**84.77**	49.84	57.85	68.03	**68.41**

Table 3. **State-of-the-art comparison on the VQA 2.0 dataset.** Results on *test-dev* and *test-std* splits. All these models were trained on the same training set (VQA 2.0 *train+val*), using the Bottom-up features provided by [3]. No ensembling methods have been used. † have been trained by [6].

	RAU* [35]	MCB* [16]	QTA [40]	MuRel
Bottom-up	✗	✗	✓	✓
Scene Reco.	93.96	93.06	93.80	**96.11**
Sport Reco.	93.47	92.77	95.55	**96.20**
Color Attr.	66.86	68.54	60.16	**74.43**
Other Attr.	56.49	56.72	54.36	**58.19**
Activity Reco.	51.60	52.35	60.10	**63.83**
Pos. Reasoning	35.26	35.40	34.71	**41.19**
Object Reco.	86.11	85.54	86.98	**89.41**
Absurd	96.08	84.82	**100.00**	99.8
Util. and Afford.	31.58	**35.09**	31.48	21.43
Object Presence	94.38	93.64	94.55	**95.75**
Counting	48.43	51.01	53.25	**61.78**
Sentiment	60.09	**66.25**	64.38	60.65
Overall (A-MPT)	67.81	67.90	69.11	**71.56**
Overall (H-MPT)	59.00	**60.47**	60.08	59.30
Overall Accuracy	84.26	81.86	85.03	**88.20**

Table 4. **State-of-the-art comparison on the TDIUC dataset.** * trained by [23].

TDIUC One of the core aspect of VQA models lies in their ability to address different tasks. The TDIUC dataset enables a detailed analysis of the strengths and limitations of a model by evaluating its performance on different types of question. We show in Table 4 a detailed comparison of recent models to our MuRel. We obtain state-of-the-art results on the Overall Accuracy and the arithmetic mean of per-type accuracies (A-MPT), and surpass by a significant margin the second best model proposed by [40]. Interestingly, we improve over this model even though it uses a combination of Bottom-up and fixed-grid features, as well as a supervision on the question types (hence its 100% result on the *Absurd* task). MuRel notably surpasses all previous methods on the Positional reasoning (+5.9 over MCB), Counting (+8.53 over QTA) questions. These improvements are likely due to the pairwise structure induced within the MuRel cell, which makes the answer prediction depend on the spatial and semantic relations between regions. The effectiveness of our per-region context modelling is also demonstrated by our the improvement on Scene recognition questions. For these questions, representing the image as a collection of independent objects shows lower performance than replacing each of them in its spatial and semantic context. Interestingly, our results on the harmonic mean of per-type accuracies (H-MPT) are lower than state-of-the-art. For MuRel, this harmonic metric is significantly harmed by our low score of 21.43% on the *Utility and Affordances* task. As these questions concern the possible usages of objects present in the scene (such as *Can you eat the yellow object?*), and are not directly related to the visual understanding of the scene.

VQA-CP v2 This dataset has been proposed to evaluate and reduce the question-oriented bias in VQA models. In particular, the distributions of answers with respect to question types differ from *train* to *val* splits. In Table 5, we report the scores of two recent baselines [1, 32], on which we improve significantly. In particular, we demonstrate an important gain over GVQA [1], whose architecture is designed to focus on Yes/No questions. However, since both methods do not use the Bottom-up features, the fairness of the comparison can be questioned. So we also train an attention model similar to [8] using these Bottom-up region representation. We observe that MuRel provides a substantial gain over this strong attention baseline. Given the distribution mismatch between *train* and *val* splits, models that only focus on linguistic biases to answer the question are systematically penalized on their *val* scores. This property of VQA-CP v2 implies that the pairwise iterative structure of MuRel is less prone to question-based overfitting than classical attention architectures.

4.4. Qualitative results

In Figure 5 we illustrate the behaviour of a MuRel network with three shared cells. Iterations through the MuRel cell tend to gradually discard regions, keeping only the most relevant ones. As explained in Section 3.2, the regions that are most involved in the pairwise modeling process are shown in green and red. Both region contributions and pair-

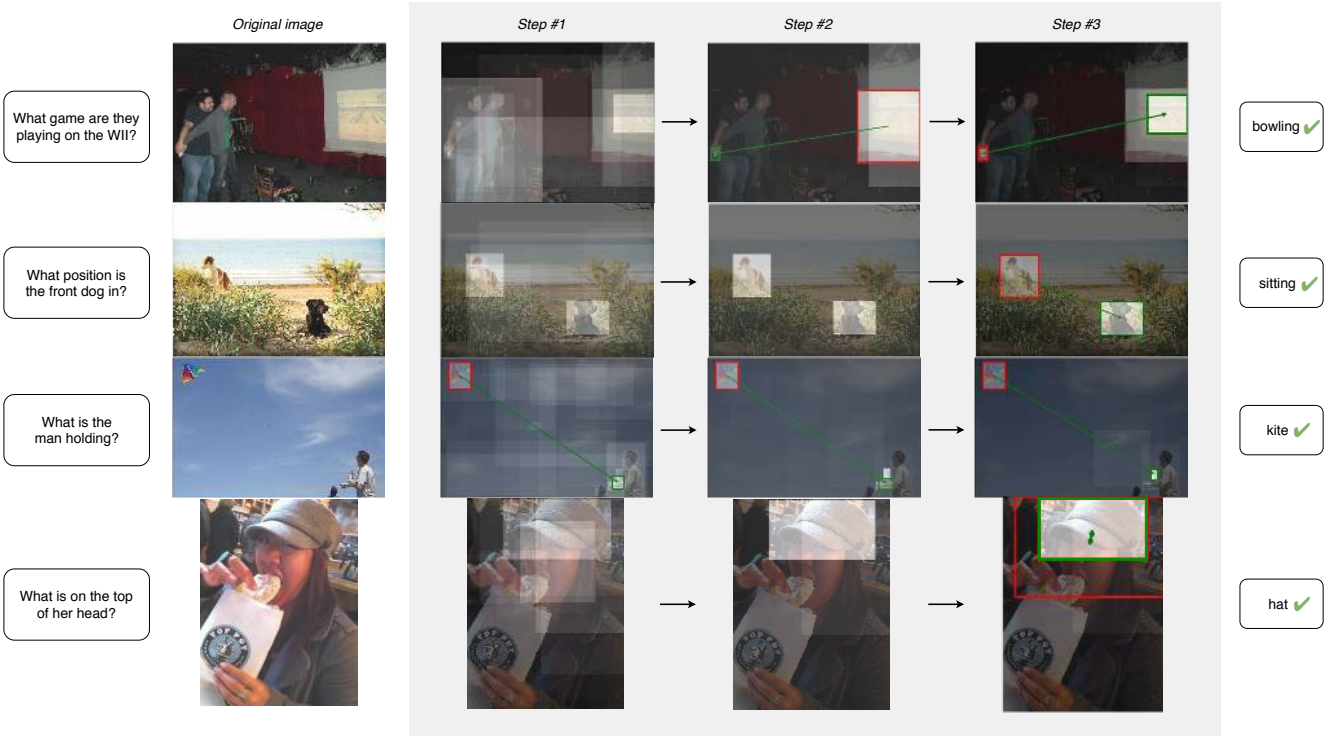

Figure 5. **Qualitative evaluation of MuRel**. Visualization of the importance maps with colored regions related to the relational mechanism. As in Figure 3, the most selected regions by the implicit attentional mechanism are shown in brighter. The green region is the most impacted by the pairwise modeling, while the red regions impact the green regions the most. These colored regions are only represented if they are greater than a certain threshold.

Model	Bottom up	Yes/No	Num.	Other	All
HAN [32]	✗	52.25	**13.79**	20.33	28.65
GVQA [1]	✗	**57.99**	13.68	22.14	31.30
Attention	✓	41.56	12.19	43.29	38.04
MuRel	✓	42.85	13.17	**45.04**	**39.54**

Table 5. **State-of-the-art comparison on the VQA-CP v2 dataset.** The Attention model was trained by us using the Bottom-up features.

wise links match human intuition. In the first row, the most relevant relations according to our model are between the player's hand, containing the WII controller, and the screen, which explains the prediction *bowling*. In the third row, the model answers *kite* using the relation between the man's hand and the kite he is holding. Finally, in the last row, our model is able to address a third question on the same image as in Figure 1 and 3. Here, the relation between the head of the woman and her hat is used to provide the right answer. As VQA models are often subject to linguistic bias [17, 1], these visualizations tend to show the ability of the MuRel network to rely on visual information to answer questions.

5. Conclusion

In this paper, we introduced MuRel, a multimodal relational network for Visual Question Answering task. Our system is based on rich representations of visual image regions that are progressively merged with the question representation. We also included region relations with pairwise combinations in our fusion, and the whole system can be leveraged to define visualization schemes helping to interpret the decision process of MuRel.

We validated our approach on three challenging datasets: VQA 2.0, VQA-CP v2 and TDIUC. We exhibited various ablation studies, clearly demonstrating the gain of our vectorial representation to model the attention, the use of pairwise combination, and the multi-step iterations in the whole process. Our final MuRel network is very competitive and outperforms state-of-the-art results on two of the most widely used datasets.

Acknowledgments

This work has been supported within the Labex SMART supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-LABX-65.

References

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7, 8

[2] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Dont just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 5

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, June 2018. 2, 3, 5, 6, 7

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR 2015*. 2

[6] Y. Bai, J. Fu, T. Zhao, and T. Mei. Deep attention neural tensor network for visual question answering. In *The European Conference on Computer Vision (ECCV)*, September 2018. 7

[7] P. W. Battaglia et al. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261, 2018. 2, 4

[8] H. Ben-Younes, R. Cadène, N. Thome, and M. Cord. Mutan: Multimodal tucker fusion for visual question answering. *ICCV*, 2017. 1, 2, 3, 5, 6, 7

[9] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the 33st Conference on Artificial Intelligence (AAAI)*, 2019. 3

[10] M. Carvalho, R. Cadene, D. Picard, L. Soulier, N. Thome, and M. Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2018. 2

[11] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[12] Z. Chen, Z. Yanpeng, H. Shuaiyi, T. Kewei, and M. Yi. Structured attentions for visual question answering. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[13] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[14] T. Durand, T. Mordan, N. Thome, and M. Cord. WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[15] M. Engilberge, L. Chevallier, P. Pérez, and M. Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018. 2

[16] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*. The Association for Computational Linguistics, 2016. 1, 2, 3, 5, 7

[17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2017. 1, 2, 5, 8

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2016. 2

[19] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[20] D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018. 2, 3

[21] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2017. 1, 2, 5

[22] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, 2017. 1, 2

[23] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 5, 7

[24] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear Attention Networks. *arXiv preprint arXiv:1805.07932*, 2018. 2, 6

[25] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017. 1, 2, 3, 6, 7

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[27] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3294–3302, Cambridge, MA, USA, 2015. MIT Press. 5

[28] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 6

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In

F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[30] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, Piscataway, NJ, USA, July 2017. IEEE. 2

[31] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 2

[32] M. Malinowski, C. Doersch, A. Santoro, and P. Battaglia. Learning visual question answering by bootstrapping hard attention. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 7, 8

[33] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc., 2014. 2

[34] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, June 2018. 1, 2

[35] H. Noh and B. Han. Training recurrent answering units with joint loss minimization for vqa. *arXiv preprint arXiv:1606.03647*, 2016. 7

[36] W. Norcliffe-Brown, E. Vafeias, and S. Parisot. Learning conditioned graph structures for interpretable visual question answering. 2018. 2, 3, 6, 7

[37] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 2, 3, 5

[38] A. Santoro, F. Hill, D. G. T. Barrett, A. S. Morcos, and T. P. Lillicrap. Measuring abstract reasoning in neural networks. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 4477–4486. JMLR.org, 2018. 2

[39] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4974–4983, 2017. 2

[40] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar. Question type guided attention in visual question answering. In *The European Conference on Computer Vision (ECCV)*, September 2018. 7

[41] D. Teney, L. Liu, and A. van den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2017. 2

[42] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057. JMLR.org, 2015. 2

[43] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2016. 2

[44] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848, 2017. 2, 3

[45] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2018. 2, 3

[46] Yu Jiang*, Vivek Natarajan*, Xinlei Chen*, M. Rohrbach, D. Batra, and D. Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 2, 5, 6, 7

[47] Y. Zhang, J. Hare, and A. Prgel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations*, 2018. 2, 7

[48] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. van den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261, 2018. 2