Visual Question Answering using Attentive Fusion of Objects, Relations and Text (A-FORT)

[redacted]

Abstract

Natural image understanding and compositional reasoning are two important lines of research required to develop models capable of answering questions about images. Instead of focusing on both of these requirements, current VQA models focus on only one of the aspects and evaluate their algorithms either only on natural VQA datasets or only on synthetic compositional reasoning datasets. Our initial experiments showed that none of these algorithms excel on both kinds of tasks. While natural VQA algorithms have focused on attentive fusion mechanisms without considering relations between image regions required for compositional reasoning, models designed for compositional reasoning use inadequate fusion strategies to combine visual and linguistic information, thereby limiting their performance on natural VQA datasets. In this work, we propose a new model "A-FORT" that utilizes all the key pieces of information: visual objects, relations and question and captures rich interactions between them using bilinear attention mechanism to excel on both natural and synthetic datasets.

1. Introduction

Visual question answering (VQA) requires understanding of visuo-linguistic concepts and reasoning with those concepts to answer questions about images. However, current VQA models either focus only on natural VQA datasets that test natural image understanding [3, 8, 16, 18, 21] or, only on synthetic datasets such as that test compositional reasoning skills [14, 4, 11, 19]. While both lines of research are essential for VQA, it is not clear if any of these models can perform well on both kinds of datasets. To examine this, we ran natural and compositional VQA algorithms on both kinds of VQA datasets and found that current natural image based VQA models perform poorly when asked to perform compositional reasoning. State-of-the-art natural-image VQA model: Bilinear Attention Network (BAN) [16] achieved only 75.56% accuracy on CLEVR (the synthetic dataset to test compositional reasoning), while models with explicit compositional reasoning mechanisms have already achieved over 98% accuracy on CLEVR [19, 12].

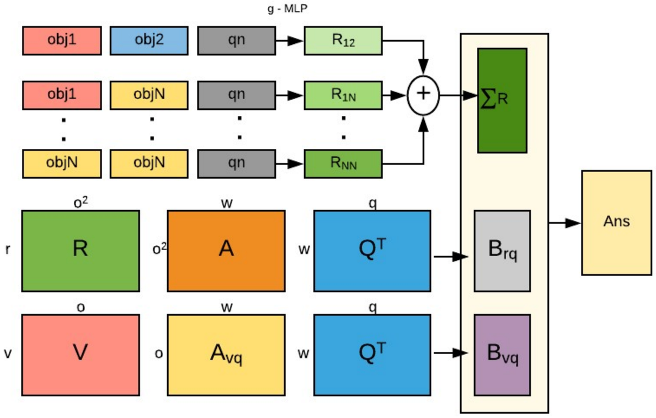

Figure 1. A-FORT model consists of a relation network that captures relations between all object pairs. It has two bilinear attention networks: the first captures bilinear interactions between relation vectors and question whereas the second captures interactions between visual features and the question.

In this work, we propose single model that we believe will work well on both kinds of datasets. The proposed model captures multi-modal interactions between visual, textual and relational information to excel on both natural image understanding and synthetic compositional reasoning tasks. We hypothesize that current natural VQA models lack compositional reasoning because they do not capture relations between visual elements which are necessary to answer questions that involve integrating information from multiple visual objects. We further hypothesize that, relation networks that do explicitly capture relations between visual objects [23, 7], use inadequate strategies to fuse those relations with textual features (e.g., simple concatenation [7]), which weakens their performance. The performance

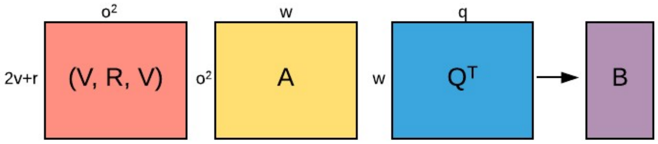

Figure 2. In this alternate approach, object pairs are combined with corresponding relation vectors. The model captures bilinear interactions between question and the combined vector.

is further limited because they discard object features after capturing relations.

In contrast, we predict answers using all pieces of information: object features, relation features, object-question interactions and relation-question interactions. Specifically, we use bilinear attention mechanism [16] which weights visual and relational features by their relevance to the question and captures interactions between every relevant object/relation with question to produce an answer. We propose two alternate strategies to capture such interactions: in the first strategy, we use separate bilinear attention maps: a) to combine visual objects with text and b) to combine relations with text. In the second strategy, we combine object pairs and their relations before applying bilinear attention mechanism to fuse with text. We evaluate both alternatives on VQA2.0 and CLEVR. If the models perform well on both kinds of datasets, we will further perform generalization tests on Task Driven Image Understanding Challenge (TDIUC) [15], Compositional VQA (C-VQA) [2], VQA under Changing Priors (VQA-CP) [1], CLEVR-Co Generalization Tests (CoGenT) and CLEVR-Humans [13].

If everything works out, we will have made the following contributions:

1) We develop a single framework that properly fuses visual, relational and linguistic information for VQA

2) The model reaches/beats state-of-the-art results on VQA2.0 and CLEVR thereby showcasing natural image understanding and compositional reasoning

3) The algorithm shows equal/improved performance on generalization tests, whereas current models don't even evaluate on them

2. Related Work

2.1. Natural VQA

Attention and fusion mechanisms have been the major focus of recent natural VQA algorithms. Traditional attention mechanisms either compute attention map for only visual input [3] or separate attention distributions for visual and textual modalities [18, 20], limiting the interactions between two modalities. Recent fusion mechanisms [8, 5] that capture bilinear interactions between visual and textual features, support multiple channels in only one of the modalities, so represent entire question as single vector instead of using separate vectors for each word. Bilinear Attention Network (BAN) [16] addresses all of these issues by predicting a bilinear attention map for fusing multiple textual channels (e.g., question words) with multiple visual channels (e.g., image regions), but still lacks compositional reasoning.

2.2. Compositional VQA

Modular networks have achieved upto 99% accuracy on CLEVR, but are not particularly well suited for natural VQA because they either require parsers [14, 4] or ground-truth functional programs [11, 19] during training. These networks are also limited by the choice of pre-defined modules used for reasoning. Memory, Attention and Composition (MAC) network [12] propose inferring reasoning steps from data itself by introducing structural priors and removing dependency on parsers, ground truth programs or pre-defined modules. While it achieved 98.9% accuracy on CLEVR, it yielded poor performance: 55% accuracy when we ran it on VQA2.0. We hypothesize that it is because MAC uses inadequate interaction between question and visual features, which is remedied in our work. Relational Networks [23] capture relations between object pairs and perform well on CLEVR, but use concatenation for fusion, which is rather simplistic.

3. Proposed Model

We use proposals from bottom-up attention mechanism [3] as our visual objects and discover relations between them using a relation network [23]. We present two alternatives to combine visual and relational features with lingustic features: in the first approach, we use one bilinear attention map for fusing objects with question features and a separate map for fusing relations with question features. In the second approach, we combine relational features with corresponding object pairs before fusing with question features.

Let us denote visual features by $\mathbf{V} \in \mathbb{R}^{o \times v}$, where o is the number of visual objects from bottom-up attention and v is the dimensionality of visual feature. Let us denote question features by $\mathbf{Q} \in \mathbb{R}^{w \times q}$, where w is the number of words in the question and q is the dimensionality of word embedding. We use a relation network to capture relations between all object pairs. The relation between object pair (V_i, V_j) is given the relation vector:

$$R_{ij} = g_\theta(V_i, V_j)$$

where, g_θ is a Multi Layer Perceptron (MLP).

We present the following alternatives to fuse visual, relational and linguistic features:

3.1. Approach 1: Separate fusions for visual features and relation vectors

We capture bilinear interactions between visual objects with question using bilinear attention mechanism formulated in [16]:

$$B_v = BAN(V, Q) = (VS)^T \mathbf{A_V}(QU_v) \tag{1}$$

where,

$$\mathbf{A_v} := softmax(((\mathbf{1}.p_v^T) \circ VS)QU) \in \mathbb{R}^{o \times w}$$

is the bilinear attention map between visual features and question features and,

$$S \in \mathbb{R}^{v \times d} \text{ and } U_v \in \mathbb{R}^{q \times d}$$

are learnable weights and p_v is used for pooling.

Next, we fuse relation vectors and question features using the same fusion mechanism formulated above. Let us denote the matrix of relation vectors between all object pairs as: $R \in \mathbb{R}^{o^2 \times r}$, where o^2 is the total number of relations to be computed and r is the length of relation vector. Then, relation vectors and question features are again fused using another bilinear map $\mathbf{A_r}$:

$$B_r = (RZ)^T \mathbf{A_r}(QU_r) \tag{2}$$

where, Z and U_r are learnable weights

Finally, B_v, B_r and sum of all relations ΣR are used to predict the answer:

$$Answer = MLP(B_v, B_r, \Sigma R)$$

Fig 1 illustrates the first approach.

3.2. Approach 2: Combining visual and relation vectors before fusing with question features

In this approach, we first apply linear projection to object features and relation vectors and concatenate each relation vector with corresponding object pairs before fusing the combined feature with question.

The combined feature for object pair (i, j) is given by:

$$C_{ij} = (e_V(V_i)||e_R(R_{ij})||e_V(V_j))$$

where, || denotes concatenation, e_V projects visual objects, e_R embeds relation vectors.

Then we fuse C with Q using bilinear attention mechanism:

$$B = BAN(C, Q) \tag{3}$$

Finally, B and sum of all relations $\sum R$ are used to predict the answer:

$$Answer = MLP(B, \Sigma R)$$

Fig 2 illustrates the second approach.

3.3. Hard Attention to select relevant relations

In both approaches, if we consider relations between all object pairs, then, we have to handle a large number of relations: o^2, among which only a few are likely to be relevant for a given question. So, instead we find relations between only those object pairs that are relevant to the question, by thresholding relation scores obtained from one of the following mechanisms:

1. Following [21], we can compute an adjacency matrix J which indicates the degree of relevance of an object pair to the question, by taking the outer product between linear projections of concatenated visual and question features. The adjacency (relevance) between objects V_i and V_j is given by:

$$J_{ij} = h(V_i||Q)^T h(V_j||Q) \tag{4}$$

where || denotes concatenation and h is a linear layer. J is thresholded to obtain only the most useful object pairs.

2. We can use a separate relation network (RN) to predict relevance score between each object pair

$$J_{ij} = RN(V_i||Q, V_j||Q)$$

4. Setup

4.1. Visual Features

We use bottom-up object proposals from pre-trained FasterRCNN [9] as our visual features. For natural images, we use FasterRCNN trained for object recognition, attribute recognition and bounding box regression on Visual Genome [17]. For CLEVR, we train a separate FasterRCNN for multi-class classification and bounding box regression. Each CLEVR object is associated with one of 8 colors (red, green, blue, purple, gray, brown, cyan, yellow), 3 shapes (sphere, cylinder, cube), 2 sizes (small, large) and 2 materials (shiny, matte). We label each CLEVR object as: *color_shape_size_material*, getting a total of 96 labels for multi-class classification. CLEVR dataset does not provide ground truth bounding boxes for regression, so, we estimate the bounding boxes by projecting the 3D objects from the provided annotations. We do not use attribute classification for CLEVR because attributes are encoded into class labels. The feature length is 2048.

4.1.1 Spatial Information

To encode position and size of an object, we append a flattened grid of (x, y) coordinates to visual features of the object. Specifically, we divide each bounding box into 16X16 grid of coordinates, with coordinates taken relative to whole image. The coordinates are normalized to [0, 1] by dividing by image width and height and are concatenated with bottom-up features. Final visual feature is a 2560 dimensional vector.

4.2. Question Features

We first represent each question word with a 300-dimensional GloVe embedding [22] and encode the question using a Gated Recurrent Unit (GRU) [6], which is also trained when training for VQA.

5. Evaluation

5.1. Datasets

We evaluate our model on VQA2.0 [10] to test natural image understanding and CLEVR [13] to test compositional reasoning. If the model performs well on both datasets, we will further test it on Task Driven Image Understanding Challenge (TDIUC) [15] to test generalization to multiple task types, Visual Question Answering under Changing Priors (VQACP) [1] to evaluate how well it handles unknown answer distributions and Compositional VQA (CVQA) [2] to see how well it generalizes to unseen concept combinations in natural images. We will also evaluate the model on CLEVR-CoGenT, which provides two splits “A” and “B” with disjoint shape+color combinations. By training only on split “A” and testing on “B”, we can evaluate generalization to unseen concept combinations for synthetic images. By further fine-tuning the model on split “B” and testing on split “A”, we evaluate how well the model remembers old concept compositions after learning new concept combinations.

5.2. Metrics

VQA2.0 provides multiple answers per question, so we use the standard “10-choose-3 VQA metric” which scores answer by their frequency:

$$acc(ans) = min(\frac{\text{number of times “ans” is chosen}}{3}, 1)$$

For CLEVR, each question has single answer, so each answer is given a weight of 1.0 and this metric becomes equivalent to traditional accuracy. For TDIUC, we use the unnormalized and normalized arithmetic mean per type (AMPT and NAMPT) and harmonic mean per type (HMPT and NHMPT) metrics as proposed in the paper.

5.3. Baseline

We run Bilinear Attention Networks and Relation Networks on all the datasets to establish a rather strong baseline.

Task	**Due Date**
Implement Approach #1	10/07/2018
Results for #1 on VQA2/CLEVR	10/12/2018
Implement Approach #2	10/16/2018
Results for #2 on VQA2/CLEVR	10/21/2018
Finetuning + Hacks	11/04/2018
Results on remaining datasets	11/10/2018
CVPR Paper ready	11/14/2018

Table 1. Milestones (October 01 to November 14)

6. Timeline

Timeline is shown in Table 6.

References

[1] 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.

[2] A. Agrawal, A. Kembhavi, D. Batra, and D. Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*, 2017.

[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 vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[4] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[5] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3, 2017.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] M. T. Desta, L. Chen, and T. Kornuta. Object-based reasoning in vqa. *arXiv preprint arXiv:1801.09718*, 2018.

[8] 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. *arXiv preprint arXiv:1606.01847*, 2016.

[9] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[10] 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 *CVPR*, volume 1, page 3, 2017.

[11] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, 3, 2017.

[12] D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[13] 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 *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017.

[14] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 3008–3017, 2017.

[15] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1983–1991. IEEE, 2017.

[16] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018.

[17] 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.

[18] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[19] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4942–4950, 2018.

[20] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016.

[21] W. Norcliffe-Brown, E. Vafeais, and S. Parisot. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*, 2018.

[22] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[23] A. Santoro, D. Raposo, D. G. 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*, pages 4967–4976, 2017.