Structured Attentions for Visual Question Answering

Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, Yi Ma
ShanghaiTech University
{zhuchen, zhaoyp1, huangsy, tukw, mayi}@shanghaitech.edu.cn

Abstract

Visual attention, which assigns weights to image regions according to their relevance to a question, is considered as an indispensable part by most Visual Question Answering models. Although the questions may involve complex relations among multiple regions, few attention models can effectively encode such cross-region relations. In this paper, we demonstrate the importance of encoding such relations by showing the limited effective receptive field of ResNet on two datasets, and propose to model the visual attention as a multivariate distribution over a grid-structured Conditional Random Field on image regions. We demonstrate how to convert the iterative inference algorithms, Mean Field and Loopy Belief Propagation, as recurrent layers of an end-to-end neural network. We empirically evaluated our model on 3 datasets, in which it surpasses the best baseline model of the newly released CLEVR dataset [13] by 9.5%, and the best published model on the VQA dataset [3] by 1.25%. Source code is available at https://github.com/zhuchen03/vqa-sva.

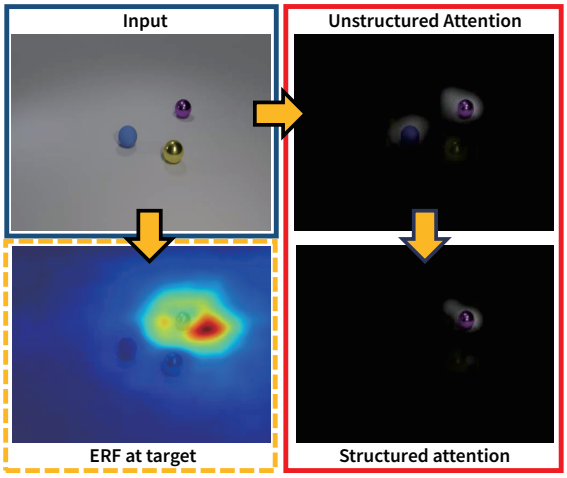


Figure 1: An example of the proposed model on CLEVR, demonstrating it is capable of inferring spatial relations despite the limited effective receptive field of the CNN. Question: *What is the color of the sphere on the right of the metal sphere?* Our model overcomes the unstructured attention's tendency to attend to isolated key words in the questions, attending to the right region and giving the correct answer *purple*.

1. Introduction

Visual Question Answering (VQA) is a comprehensive task inspecting intelligent systems' ability to recognize images and natural languages together. Advances in this area not only benefit real-world applications which require the synergistic reasoning of vision and language, such as querying events in surveillance videos [34] and searching specific goods in images [41], but also call for finer-grained understanding on the semantic structures of images.

Adoption of the visual attention mechanism like [38, 40, 24, 27] is a major source to boost performance of VQA. Visual attentions impose regularization on the models to find the most relevant image regions to the question. Still, experiments [6] point out that state-of-the-art models often fail to identify the related regions like humans do. We argue this problem comes from the fact that such attention models do not take into account the spatial relations between regions when predicting the attention. This is important because the effective receptive field (ERF) of deep CNNs only covers a small fraction of the image [25]. Even with memory mechanism [37, 27, 40, 38], it is difficult to infer the right attention corresponding to questions that involve the spatial relations between regions without overlapping ERFs.

In this paper, we propose a novel neural network to model the attention with a multivariate distribution which considers the arrangements of image regions. We adopt the most straightforward graph structure to model image region arrangements - a grid-structured Conditional Random Field (CRF) [29]. The framework of the proposed method is illustrated in Fig. 2. We show that attention can be formulated as the marginal distribution of each hidden variable in the CRF. Then, we implement the iterative approximate inference algorithms, Mean Field and Loopy Belief Propagation, as recurrent layers of neural networks, which iteratively refines the attention. An example of this process is shown in Fig. 1. We evaluate the proposed model on three repre-

sentative datasets, where our model is competitive with the rule-based model [2] on the SHAPES dataset, and surpasses the best baseline model of CLEVR by more than 9.5% and best published method [17] on the VQA dataset by 1.25%. We also demonstrate that the ERF on CLEVR and the VQA dataset is not large enough for previous methods to answer all questions involving object arrangements correctly.

Our work has the following contributions. First, we propose to model structured attentions with a CRF over image regions for Visual Question Answering, to address the problem of limited effective receptive fields of CNNs. Second, we demonstrate how to unfold both Mean Field and Loopy Belief Propagation algorithms for CRF as recurrent layers of neural networks, and perform comprehensive evaluation of the two different networks on three challenging datasets. Third, we give empirical evaluations of ERFs on CLEVR and the MSCOCO dataset.

2. Related Work

There have been many different directions for improving VQA performance, including predicting answer types [14], utilizing task-specific submodules [1, 2], and better multimodal feature pooling methods [8, 17]. Our focus is on structure-aware spatial attention applied on the visual feature. There are two major forms of spatial attention for VQA. One [31, 22, 11] is based on region proposals generated by Edge Box [43], and the other [38, 40, 24, 27, 8, 17] predicts attention on the individual feature vectors of convolutional feature maps. The most representative model SAN [40] adopted multiple attention layers to support multi-step reasoning. There are also methods [24, 27] that adopted both image attention and question attention to refine the image and question representation simultaneously. [38] proved its ability to recognize absolute and relative positions with two simple experiments, but the model itself does not consider the arrangement of regions and the success may be attributed to the power of CNNs.

Some of the methods considered the structures of images. [26] adopted the Bayesian framework based on the logic forms of segmentation results. The method was surpassed by some simple baseline models due to its demand for better semantic segmentation [15]. The DMN [37] adopted a bidirectional GRU that traverses the convolutional feature map in a snake-like fashion to encode the dependency of regions, which might not be the optimal choice for the 2D structure of images. [22] concatenated the 8D bounding box representation with the 4096D visual feature, in which the spatial information could be easily overweighed by the visual feature.

There have been a number of approaches combining neural networks and CRF to predict structured outputs in both computer vision and natural language processing. [30] utilized neural networks with sigmoid activation to predict the unary potential for sequential labeling. [7] also used neural networks to predict the pairwise potential of labels. [4] learned fixed pairwise potentials for word recognition and image classification. [12] utilized CNNs to predict both the unary potentials and higher-order potentials for unconstrained word recognition. [42] proposed to unfold the Mean Field algorithm as recurrent layers for semantic segmentation. It modeled the pairwise potential with Gaussian kernels, which encourages similar features to take the same label. Besides the supervised structure in the output layers, [18] enforced the intermediate layers of neural networks to learn structured attentions for natural language tasks.

To the best of our knowledge, structured attention has not been explored for the complex task of VQA. We give the first empirical evaluations on unfolding both Mean Field and Loopy Belief Propagation as intermediate recurrent layers on the task of VQA, which can be seen as a further exploration of [18] in modeling 2D structures of visual data.

3. Attention Models and Methods

The general architecture of the proposed model is shown in Fig. 2. Here we define some notations used across the paper. We take the question feature $\boldsymbol{q} \in \mathbb{R}^{n_Q}$ from the last time step of a GRU such as [20], and the image feature map $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M] \in \mathbb{R}^{n_I \times M}$ from one of the convolution layers of a CNN such as [10]. Here n_Q, n_I are the dimensions of question and image feature vectors respectively, and M is the total number of image feature vectors which divides the image into M regions. We use $\mathrm{softmax}(\cdot)$ to denote the softmax activation function and $\sigma(\cdot)$ to denote the sigmoid activation function. The attention mechanism in VQA aims to produce a *context* $\mathbf{c}$ from $\mathbf{X}$ which represents the visual feature related to the question.

3.1. Unstructured Categorical Attention

In previous methods for VQA, visual attention is usually modeled as a single or multi-step soft-selection from $\mathbf{X}$. As shown in [39, 18], the soft-selection approach represents the selected region index by a categorical latent variable $z \in \{1, ..., M\}$ and defines $\mathbf{c}$ as an expectation of the selection:

$$\mathbf{c} = \mathbb{E}\left[\sum_i \mathbb{1}_{\{z=i\}} \cdot \mathbf{x}_i\right] = \sum_i p(z = i|\mathbf{X}, \boldsymbol{q})\mathbf{x}_i, \quad (1)$$

where $\mathbb{1}_{\{z=i\}}$ is an indicator function, and the distribution of z is parameterized by

$$p(z = i|\mathbf{X}, \boldsymbol{q}) = \mathrm{softmax}(\mathbf{U}g(\mathbf{x}_i, \boldsymbol{q})), \quad (2)$$

where $\mathbf{U} \in \mathbb{R}^{1 \times n_I}$ and $g(\cdot)$ is some multimodal feature pooling function such as [8, 17]. Noticing this model ignores the spatial arrangement of the feature vectors in $\mathbf{X}$ in

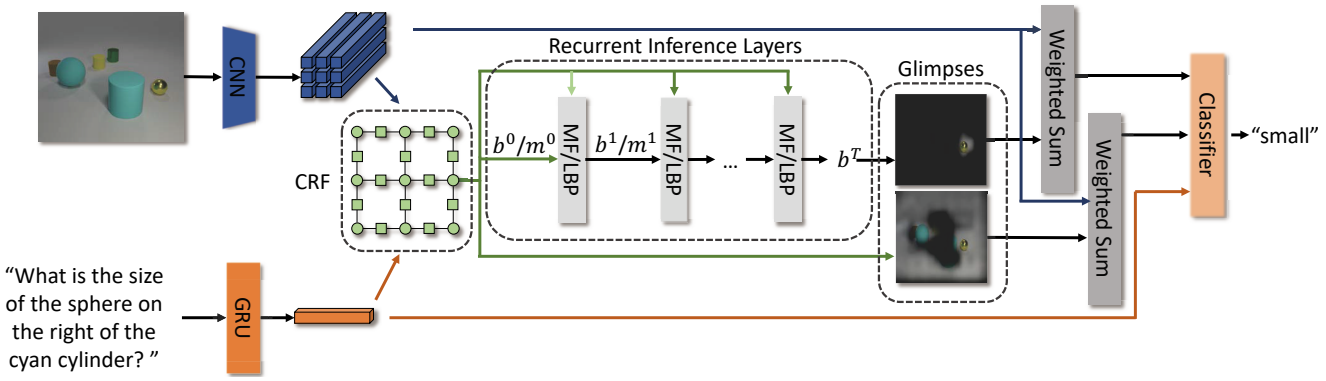


Figure 2: The whole picture of the proposed model. The inputs to the recurrent inference layers are the unary potential $\psi_i(z_i)$ and pairwise potential $\psi_{ij}(z_i, z_j)$, computed with Eq. 8. $\psi_i(z_i)$ can also be used as an additional glimpse, which usually detects the key nouns. In the inference layers, x^i represents $b^{(i)}$ for MF and $m^{(i)}$ for LBP. The recurrent inference layers generates a structured glimpse with MF or LBP. The 2 glimpses are used to weight-sum the visual feature vectors. The classifier use both of the attended visual features and the question feature to predict the answer. The demonstration is a real case.

each step, and the resulting hidden states in multi-step models [39, 38] are still unstructured, we have dropped the hidden states in the condition presented in [38] for notational convenience.

Since categorical distribution only requires the probabilities to be positive and sum to 1, the following normalized sigmoid attention is still a valid categorical attention:

$$p(z = i|\mathbf{X}, \boldsymbol{q}) = \frac{\sigma(\mathbf{U}g(\mathbf{x}_i, \boldsymbol{q}))}{\sum_j \sigma(\mathbf{U}g(\mathbf{x}_j, \boldsymbol{q}))}. \tag{3}$$

We can use such attention as a glimpse[1] of our model, which will be introduced in Section 3.4.

3.2. Structured Multivariate Attention

To consider the arrangement of $\mathbf{X}$, we adopt a structured multivariate attention model similar to [18], in which we consider the distribution $\mathbf{z} \sim p(\mathbf{z}|\mathbf{X}, \boldsymbol{q})$ as a vector of binary latent variables $\mathbf{z} = [z_1, ..., z_M]^T$ with $z_i = 1$ if $\mathbf{x}_i$ is related to the question and $z_i = 0$ otherwise. Multiple regions can now be selected at the same time. We define the context as the expectation of the sum over all related regions, which can be derived as a sum of $\mathbf{x}_i$ weighted by the marginal probability $p(z_i = 1|\mathbf{X}, \boldsymbol{q})$:

$$\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{X},\boldsymbol{q})}[\mathbf{X}\mathbf{z}] = \sum_i p(z_i = 1|\mathbf{X}, \boldsymbol{q})\mathbf{x}_i. \tag{4}$$

Let $S = \sum_i p(z_i = 1|\mathbf{X}, \boldsymbol{q})$. Since $0 \leq S \leq M$ and M is relatively large, to reduce covariate shift, we normalize the expectation to get the context $\mathbf{c}$:

$$\mathbf{c} = \frac{1}{S}\sum_i p(z_i = 1|\mathbf{X}, \boldsymbol{q})\mathbf{x}_i, \tag{5}$$

[1]Glimpses refer to multiple attentions, same as in [8, 17].

We model the distribution $p(\mathbf{z}|\mathbf{X}, \boldsymbol{q})$ in the most straightforward form, a grid-structured Conditional Random Field, which represents the joint probability $p(\mathbf{z}|\mathbf{X}, \boldsymbol{q})$ with a grid-structured factor graph that considers the pairwise joint distribution of a region's 4-neighbourhood, as shown in Fig. 2. Let $\mathcal{N} = \{(i, j)|i < j, j \in N_i\}$, where N_i is the set of i's neighbors on the graph. The grid-structured CRF assumes

$$p(\mathbf{z}|\mathbf{X}, \boldsymbol{q}) = \frac{1}{Z}\prod_{(i,j)\in\mathcal{N}} \psi_{ij}(z_i, z_j)\prod_i \psi_i(z_i), \tag{6}$$

where the unary potential $\psi_i(z_i) \geq 0$ measures the likelihood of region i taking the value $z_i \in \{0, 1\}$, and the pairwise potential $\psi_{ij}(z_i, z_j) \geq 0$ measures the likelihood of regions (i, j) taking values z_i, z_j respectively.

3.3. Recurrent Inference Layers

The inference problem in such a gird-structured factor graph, which aims to calculate the marginal probability $p(z_i|\mathbf{X}, \boldsymbol{q}; \theta)$, is known to be NP-hard [32]. Still, there are approximate inference algorithms to solve the problem, such as Mean Field (MF) and Loopy Belief Propagation (LBP). These algorithms take potential functions $\psi_i(z_i)$ and $\psi_{ij}(z_i, z_j)$ as inputs and update $p(z_i|\mathbf{X}, \boldsymbol{q}; \theta)$ iteratively through message passing. We train neural networks to predict optimal $\psi_i(z_i)$ and $\psi_{ij}(z_i, z_j)$ and then run the algorithms for a fixed number of T steps. The iterative algorithms are implemented as recurrent inference layers in the neural network.

3.3.1 Potential Functions

In VQA, the potential functions should depend on both the image and the question. We use low-rank bilinear pooling [17], a parsimonious bilinear model, to capturing the interaction between 2 feature vectors $\mathbf{x}, \mathbf{y}$:

$$g(\mathbf{x}, \mathbf{y}; \mathbf{P}_x, \mathbf{P}_y) = \tanh(\mathbf{P}_x\mathbf{x}) \odot \tanh(\mathbf{P}_y\mathbf{y}), \tag{7}$$

Algorithm 1 MF Recurrent Layer in VQA

Input: $\psi_i(z_i)$, $\ln \psi_{ij}(z_i, z_j)$
  initialize $b_i^{(0)}(z_i) = \psi_i(z_i)$
  for t=1:T **do**
    for $i = 1 : M$, z_i in $\{0, 1\}$ **do**
      s←0
      for j in N_i, z_j in $\{0, 1\}$ **do**
        s←s+$b_j^{(t-1)}(z_j) \ln \psi_{ij}(z_i, z_j)$
      $b_i^{(t)}(z_i) \leftarrow \psi_i(z_i) \exp(s)$
      normalize($b_i^{(t)}$)
  return $b_i^{(T)}(z_i)$

Algorithm 2 LBP Recurrent Layer in VQA

Input: $\psi_i(z_i)$, $\psi_{ij}(z_i, z_j)$
  initialize $m_{ij}^{(0)}(z_j) = 0.5$, $m_{ij}^{(t)}(z_j) = 0$ for $t > 0$
  for t=1:T **do**
    for $j = 1 : M$, i in N_j **do**
      for z_i in $\{0, 1\}$ **do**
        s(z_i) ←1
        for k in $N_i \setminus \{j\}$ **do**
          s(z_i) ← s(z_i) $\cdot\, m_{ki}^{(t-1)}(z_i)$
      for (z_i, z_j) in $\{0, 1\} \times \{0, 1\}$ **do**
        $m_{ij}^{(t)}(z_j) \leftarrow m_{ij}^{(t)}(z_j) + \psi_{ij}(z_i, z_j)\psi_i(z_i)s(z_i)$
      normalize($m_{ij}^{(t)}$)
    for $i = 1 : M$ **do**
      $b_i(z_i) \leftarrow \psi_i(z_i)$
      for k in N_i, z_i in $\{0, 1\}$ **do**
        $b_i(z_i) \leftarrow b_i(z_i) \cdot m_{ki}^{(T)}(z_i)$
      normalize(b_i)
  return $b_i(z_i)$

where $\mathbf{P}_x, \mathbf{P}_y$ are learnt projection matrices projecting $\mathbf{x}, \mathbf{y}$ to the same dimension, and $\odot$ represents Hadamard product. Based on this, we model $\psi_i(z_i)$ and $\psi_{ij}(z_i, z_j)$ as following:

$$\psi_i(z_i = 1) = \sigma\left(\mathbf{U} g\left(\mathbf{x}_i, \boldsymbol{q}; \mathbf{U}_x, \mathbf{U}_q\right)\right), \tag{8}$$

$$\psi_i(z_i = 0) = 1 - \psi_i(z_i = 1), \tag{9}$$

$$\psi_{ij}(z_i, z_j) = h\left(\mathbf{v}_{z_i z_j} g\left(\mathbf{y}_{ij}, \boldsymbol{q}; \mathbf{V}_y, \mathbf{V}_q\right)\right), \tag{10}$$

where $\mathbf{U}_x \in \mathbb{R}^{n_c \times n_I}, \mathbf{U}_q \in \mathbb{R}^{n_c \times n_Q}$, $\mathbf{U} \in \mathbb{R}^{1 \times n_c}, \mathbf{V}_y \in \mathbb{R}^{n_c \times 2n_I}, \mathbf{V}_q \in \mathbb{R}^{n_c \times n_Q}$ are learnt projection matrices, $\mathbf{v}_{z_i z_j}$ is a row vector of $\mathbf{V} \in \mathbb{R}^{4 \times n_c}$ indexed by z_i and z_j, $\mathbf{y}_{ij} = [\mathbf{x}_i^T, \mathbf{x}_j^T]^T$, n_c is the common projection dimension, $h(\cdot)$ is a certain activation function which will differ in the 2 inference algorithms.

3.3.2 Mean Field Layers

The Mean Field algorithm approximates the distribution $p(\mathbf{z}|\mathbf{X}, \boldsymbol{q})$ in Eq. 6 with a fully factorized distribution $q(\mathbf{z})$:

$$q(\mathbf{z}) = \prod_i b_i(z_i),$$

where $b_i(z_i)$ are variational parameters corresponding to the marginal probabilities $p(z_i|\mathbf{X}, \boldsymbol{q})$. The variational parameters are optimized by iteratively minimizing the mean-field free energy

$$\begin{aligned} F_{MF}(b_i) = &- \sum_{(i,j)\in\mathcal{N}} \sum_{z_i, z_j} b_i(z_i) b_j(z_j) \ln \psi_{ij}(z_i, z_j) \\ &+ \sum_i \sum_{z_i} b_i(z_i)[\ln b_i(z_i) - \ln \psi_i(z_i)] \end{aligned} \tag{11}$$

subject to the constraint $\sum_{z_i} b_i(z_i) = 1$, which is shown to be equivalent to minimizing the KL divergence between $p(\mathbf{z}|\mathbf{X}, \boldsymbol{q})$ and $q(\mathbf{z})$ [35]. Specifically, MF initializes $b_i^{(0)}(z_i) = \psi_i(z_i)$ and updates $b_i^{(t)}(z_i)$ as:

$$b_i^{(t)}(z_i) = \alpha \psi_i(z_i) \cdot \exp\left(\sum_{j \in N_i} \sum_{z_j} b_j^{(t-1)}(z_j) \ln \psi_{ij}(z_i, z_j)\right), \tag{12}$$

where α is the normalizing constant. Since Eq. 12 only involves $\ln \psi_{ij}(z_i, z_j)$, we adopt a log model for pairwise potential,

$$\ln \psi_{ij}(z_i, z_j) = \tanh(\mathbf{v}_{z_i, z_j} g(\mathbf{y}_{ij}, \boldsymbol{q}; \mathbf{V}_y, \mathbf{V}_q)). \tag{13}$$

MF can be unfolded as a recurrent layer of neural networks without parameters, where in each step t the inputs are $\psi_i(z_i), \psi_{ij}(z_i, z_j)$ and $b_i^{(t-1)}(z_i)$, as demonstrated by Algorithm 1.

3.3.3 Loopy Belief Propagation Layers

Loopy Belief Propagation can be applied similarly by initializing all messages to a fixed value and updating the messages iteratively in a fixed or random order [29]. In our grid-structured graph, since each factor is connected to only 2 variables, the variable-to-factor and factor-to-variable messages can be merged as a single message m_{ij}, representing the message from z_i to z_j. We initialize $m_{ij}^{(0)}(z_j) = 0.5$, and iteratively update the messages based on the formula

$$m_{ij}^{(t)}(z_j) = \alpha \sum_{z_i} \psi_{ij}(z_i, z_j) \psi_i(z_i) \cdot \prod_{k \in N_i \setminus \{j\}} m_{ki}^{(t-1)}(z_i), \tag{14}$$

where α is the normalizing constant as above. After a fixed number of steps T, the variable z_i gathers all the messages from its neighbourhood to get the marginal probability:

$$b_i(z_i) = \beta \psi_i(z_i) \prod_{k \in N_i} m_{ki}^{(T)}(z_i), \tag{15}$$

where β is the normalizing constant.

Similar to MF, LBP can also be unfolded into parameterless recurrent layer, where at each time step t the input is $\psi_i(z_i), \psi_{ij}(z_i, z_j)$ and $m_{ij}^{(t-1)}(z_j)$, and after T steps

$m^{(T)}(z_i)$ is used to compute marginal probability $b_i(z_i)$, as shown in Algorithm 2.

3.4. The overall structure

As shown in Fig. 2, the overall structure of the proposed model is an end-to-end classification neural network. Firstly, $\psi_i(z_i)$ and $\psi_{ij}(z_i, z_j)$ are computed using the extracted features $\mathbf{X}$ and $\boldsymbol{q}$. Then, the recurrent inference layers run for T steps to get the structured marginal probability $p(z_i = 1|\mathbf{X}, \boldsymbol{q}) = b_i^{(T)}(z_i = 1)$, which is then used to compute the structured context $\hat{\mathbf{c}}$ with Eq. 5. We also compute a unstructured context $\tilde{\mathbf{c}}$ with Eq. 1 by replacing $\mathbf{U}g(\mathbf{x}_i, \boldsymbol{q})$ in Eq. 3 with $\psi_i(z_i = 1)$. In the classifier, the contexts are both pooled with the question to get

$$\hat{\mathbf{s}} = g(\hat{\mathbf{c}}, \boldsymbol{q}; \hat{\mathbf{W}}_c, \hat{\mathbf{W}}_q), \tag{16}$$

$$\tilde{\mathbf{s}} = g(\tilde{\mathbf{c}}, \boldsymbol{q}; \tilde{\mathbf{W}}_c, \tilde{\mathbf{W}}_q), \tag{17}$$

where $\hat{\mathbf{W}}_c, \tilde{\mathbf{W}}_c \in \mathbb{R}^{n_c \times n_I}$, $\hat{\mathbf{W}}_q, \tilde{\mathbf{W}}_q \in \mathbb{R}^{n_c \times n_Q}$. The answer is predicted with

$$a = \arg\max_{k \in \Omega_K} \text{softmax}(\mathbf{w}_k[\hat{\mathbf{s}}^T, \tilde{\mathbf{s}}^T]^T), \tag{18}$$

where $\mathbf{w}_k$ is the k-th row of $\mathbf{W} \in \mathbb{R}^{K \times 2n_c}$, K is the number of answers, Ω_K is the answer space with up to K answers.

4. Experiments

4.1. Datasets

The SHAPES dataset [2] is a synthetic dataset consisting of images containing 3 basic shapes in 3 different colors with a resolution of 30×30, and queries about the arrangements of the basic shapes, as shown in Fig. 3. The answer is "yes" when the image satisfies the query, and "no" otherwise. There are 3 different lengths of queries. The original dataset [1] has 14592 and 1024 image/question pairs for the training and test sets. All the queries in the test set do not appear in the training set.

The CLEVR dataset [13] is a much more complex but unbiased synthetic dataset aiming at testing visual abilities such as counting, comparing and logical reasoning. It consists of 100,000 images of simple 3D objects with random shapes, sizes, materials, colors and positions with a resolution of 320×480, and nearly a million natural language questions, 853,554 of which are unique. The questions can be categorized into 5 general types: *exist*, *count*, *compare integer*, *query attribute*, and *compare attribute*. There are 699,989 training questions, 149,991 validation questions and 149,988 test questions. The vocabulary sizes for the questions and answers are 82 and 28 respectively.

The VQA real-image dataset [3] is a comprehensive dataset which requires knowledge beyond the dataset to answer all the questions correctly. It has about 204,721 images from MSCOCO [23] each with 3 natural language questions, and each question has 10 answers collected from online workers. It consists of 3 splits: `train`, `val` and `test`, each of which has 248,349, 121,512 and 244,302 questions respectively. `test-dev` is a subset of `test`, which has 60,864 questions. We keep a collection of 2,000 most frequent answers from the union of `train` and `val`, and ignore questions with no answers from this collection, which leaves us 334,554 samples for training. With the same preprocessing procedure as [8, 17], we also get 837,298 training samples from Visual Genome [21] for augmentation.

4.2. Model Configuration and Training

For extracting image features, we use a 2-layer LeNet trained with the whole network on SHAPES as in [1], and ImageNet-pretrained ResNets [10] on CLEVR and the VQA dataset. For sentence embedding, we use single-layer GRU. On the SHAPES dataset, we set $n_c = 128, n_Q = 128, n_I = 50, M = 9$. On CLEVR, following [13], we resize the input images to 224×224, and use feature maps at the `res4b22` layer of ResNet-101 ($n_I = 1024, M = 196$) and the `res5c` layer of ResNet-152 ($n_I = 2048, M = 49$) in different experiments. We also set $n_Q = 2048$, comparable to the 2-layer LSTM with 1024 units per layer used in [13]. On the VQA dataset, we fix $n_c = 1200$. Images are resized to 448×448 and we use the feature at the `res5c` layer of ResNet-152 ($n_I = 2048, M = 196$), the same as [24, 8, 17]. We set $n_Q = 2400$ since we use the pre-trained skip-thought vector [20] provided by [17] as initialization.

For training, we implement our network with MXNet [5]. In all 3 tasks, we use the Adam optimizer [19] with the default setting except for the learning rate, which is picked using grid search. We adopt Bayesian dropout [9] for GRU's as in [17], and apply dropout [33] before every other fully connected layer. On SHAPES and CLEVR, we find setting both dropout probability to 0.2 to be optimal, while on the VQA dataset, setting a small Bayesian dropout 0.25 for the GRU and a large dropout of 0.5 for the other parts achieved the optimal results. In addition, we use answer sampling by default on the VQA dataset as in [8, 17].

4.3. Visualization of ERF

To visualize the ERF, we need to compute the influence of a pixel $I_{ij} \in \mathbb{R}^3$ on the entry y^n_{rc}, the feature at (r, c) of the n-th channel in a certain conv layer, represented by $\|\partial y^n_{rc}/\partial I_{ij}\|$. As in [25], we assume a loss function l^n_{rc} which is related only to channel y^n_{rc}, i.e., $\partial l^n_{rc}/\partial y^n_{rc} = 1$ and $\partial l^n_{rc}/\partial y^n_{ij} = 0$ for $i \neq r$ or $c \neq j$, so that $\partial y^n_{rc}/\partial I_{ij} = \partial l^n_{rc}/\partial I_{ij}$, since $\partial l^n_{rc}/\partial I_{ij}$ can be computed efficiently by DL frameworks. Finally, we draw the heat map of the total effect of a subset Ω_C of channels:

$$E_{ij} = \left\| \sum_{n \in \Omega_C} \left(\frac{\partial l^n_{rc}}{\partial I_{ij}} \right)^2 \right\|. \tag{19}$$

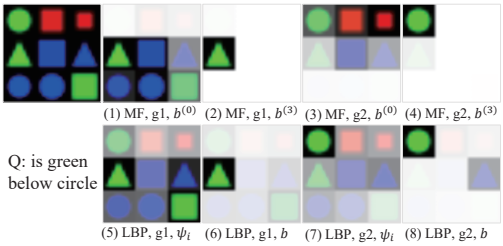


Figure 3: Visualization of MF-G2 and LBP-G2 on the test set of `large`. $b^{(0)}$ or ψ_i represents the initial attention, $b^{(3)}$ or b represents the refined attention after MF or LBP. The query looks for a green object under a circle.

	Query Length	3	4	5	All
	% of test set	12.5	62.5	25	-
`small`	NMN[2]	**91.4**	**95.6**	**92.6**	**94.3**
	SIG-G2	57.0	70.5	66.8	67.9
	MF-G2	53.1	71.4	66.0	67.8
	LBP-G2	63.3	72.2	62.5	68.7
`medium`	NMN	**99.2**	92.1	**85.2**	91.3
	SIG-G2	68.8	79.6	73.8	76.8
	MF-G2	98.0	**99.6**	71.5	**92.4**
	LBP-G2	87.1	99.5	71.9	91.0
`large`	NMN	**99.7**	94.2	**91.2**	94.1
	SIG-G2	93.2	95.6	72.5	89.5
	MF-G2	**99.7**	99.9	79.2	**94.7**
	LBP-G2	95.1	**100**	78.9	94.1

Table 1: Accuracy on SHAPES.

4.4. Results and Analysis

We perform experiments on the 3 datasets, in which we fix the general structure except for the visual attention models, to quantify the role of structured attention in our model and the best configuration for it. We will use the following abbreviations to distinguish the models we implemented:

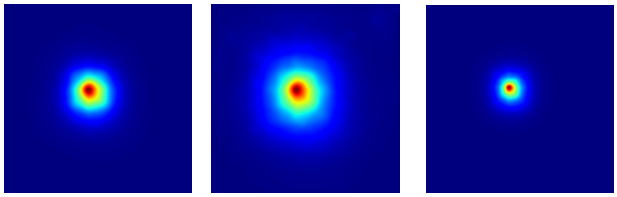
- SM/SIG: 1-glimpse softmax or sigmoid attention.
- MF/LBP: 1-glimpse multivariate attention with a MF or LBP recurrent layer and a default $T = 3$.
- *MODEL*-G2: 2-glimpses with a default $T = 3$, where *MODEL*=SIG,MF or LBP is the attention model.
- MF-SIG/LBP-SIG: 2-glimpse model by concatenating a MF or LBP attention with a SIG attention, as mentioned in Section 3.4.
- *MODEL*-Tn: *MODEL* with $T = n$ inference steps.

4.4.1 On SHAPES

In this part, we will look into the influence of the volume of the data on our model's generalization, and compare the performance of SIG-G2, MF-G2 and LBP-G2 models. We find it is difficult for all 3 models to generalize with the same amount of data as [1]. This may because [1] uses a parser to understand the queries with guaranteed correctness on this dataset, while we have to train the GRU to understand the queries from scratch. The parser-based method may not perform well in more general tasks such as the VQA dataset, since they found using fewer modules on the VQA dataset turned out to be better, but our RNN-based approach should generalize better with enough training data. So we generate more data with the same answer distributions for each query as [1] to train and test our model, and re-trained their model using the released code. We name the original dataset `small`, and the newly generated datasets with 2 and 3 times as much data in both training and test sets as `medium` and `large` respectively. We find with more training data, our model becomes competitive with [1] and the MF-G2 model surpasses it on both `medium` and `large`, as shown in Table 1. Our models are extremely good at handling length-4 queries, which looks for object arrangements in 4-neighborhood, as demonstrated in Fig. 3. The high accuracy also implies the model is capable of set theory reasoning, since it achieved high test accuracy with length-3 queries which contain self-conflict queries such as *is red green*, which aims to find an object that is both red and green. For complex queries, such as *is red below below green*, which aims to find a red object below another object that is below a green object, it is not as competitive as [1], probably because the GRU in our model has not generalized to higher order logic.

4.4.2 On CLEVR

In this part, we study the role of different visual features and different kinds of attentions on the performance, and test our best models on the test set, as shown in Table 2. Our best model surpasses the best baseline model in [13] by more than 9.5% on the test set. Both MF and LBP outperform SM and SIG, demonstrating the effectiveness of our method. The maximum margin of MF/LBP vs. SIG on overall accuracy is 2.62% and 1.33% with the ResNet-152 and ResNet-101 features respectively. The most significant improvement of MF/LBP over other models is on Compare Attribute, which involves comparing specific attributes of objects specified by spatial relations with other objects. This also proves that our model alleviates the problem of previous methods of ignoring arrangements of regions. Further, we show the receptive fields of the 2 selected layers on the test set in Fig. 4. In each image, we choose the feature vector closest to the center, where there is a higher chance for objects to appear. Still, both of them only occupy a small portion of the image, indicating the importance of considering the structure of regions. Overall, the performance with `res4b22` features is better than that with `res5c` features. From the ERF point of view, the ERF of `res5c` has more



`res5c` on CLEVR `res5c` on CLEVR `res5c` on MSCOCO

Figure 4: The average ERF [25] of 32 channels chosen at regular intervals, on 15000 images from the CLEVR test set and 52500 images from the MSCOCO test set with resolution of 224×224 and 448×448 respectively. The ERF images are smoothed with $\sigma = 4$ Gaussian kernels.

	All	Exist	Count	CI	QA	CA
`res5c`	CLEVR validation					
SM	68.80	73.20	53.16	76.52	81.58	56.77
SIG	70.52	73.90	53.89	76.52	82.46	63.06
MF	73.14	76.46	**56.89**	77.43	83.72	**68.76**
LBP	72.30	76.32	54.92	77.50	83.35	67.54
MF-SIG	73.19	76.53	56.22	**78.56**	**84.23**	68.34
LBP-SIG	**73.33**	**77.50**	56.39	77.97	84.09	68.70
`res4b22`	CLEVR validation					
SM	75.63	77.69	57.79	78.63	87.76	71.83
SIG	75.32	76.54	58.93	78.12	87.94	69.38
MF	76.65	77.90	58.87	**80.48**	88.10	74.34
LBP	76.21	78.97	57.52	80.14	87.90	73.43
MF-SIG	77.4	**79.8**	61.0	79.3	88.0	75.1
LBP-SIG	**77.97**	79.7	**61.39**	80.17	**88.54**	**76.31**
`res4b22`	CLEVR test					
MCB[13]	51.4	63.4	42.1	63.3	49.0	60.0
SAN[13]	68.5	71.1	52.2	73.5	85.2	52.2
MF-SIG	77.57	**80.05**	60.69	80.08	88.16	75.27
LBP-SIG	**78.04**	79.63	**61.27**	**80.69**	**88.59**	**76.28**

Table 2: Accuracy on CLEVR. CI, QA, CA stand for Count Integer, Query Attribute and Compare Attribute respectively. The top half uses ResNet-152 features and the bottom half uses ResNet-101 features. Our best model uses the same visual feature as [13].

than twice the area as the ERF of `res4b22`. As a result, the feature vector of `res5c` may require more than twice the number of parameters to represent same amount of information in this region as `res4b22`, but its has only twice.

4.4.3 On the VQA dataset

Since we have found MF-SIG and LBP-SIG are the best on CLEVR, in this part, we mainly compare the two models with different T. Notice now the total number of glimpses is the same as MCB [8] and MLB [17], and both of them use `res5c` features and better feature pooling methods. The optimal choice in these experiments is MF-SIG-T3, which is 0.92% higher in overall accuracy than the previous best method [17], and outperforms previous methods on all 3 general categories of questions. We then use external data from Visual Genome to train MF-SIG-T3 and MF-T3, in

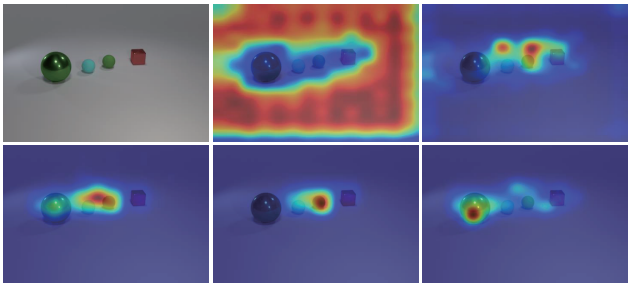
Model	All	Y/N	No.	Other
MCB[8]	64.7	82.5	37.6	55.6
MLB[17]	65.08	84.14	38.21	54.87
MF-SIG-T1	65.90	84.22	39.51	56.22
MF-SIG-T2	65.89	84.21	39.57	56.20
MF-SIG-T3	**66.00**	**84.33**	**39.34**	**56.37**
MF-SIG-T4	65.81	84.22	38.96	56.16
LBP-SIG-T1	65.93	84.31	39.27	56.26
LBP-SIG-T2	65.90	84.23	39.70	56.16
LBP-SIG-T3	65.81	84.05	39.76	56.12
LBP-SIG-T4	65.73	84.08	38.87	56.13
MCB+VG[8]	65.4	82.3	37.2	57.4
MLB+VG[17]	65.84	83.87	37.87	56.76
MF+VG	**67.17**	**84.77**	**39.71**	**58.34**
MF-SIG+VG	**67.19**	**84.71**	**40.58**	**58.24**
On `test-dev2017` of VQA2.0				
MF-SIG+VG	64.73	81.29	42.99	55.55

Table 3: Results of the Open Ended task on `test-dev`.

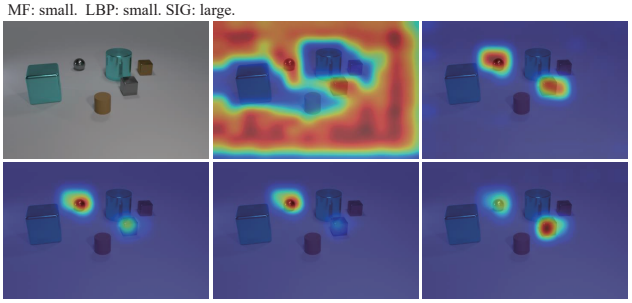
	Open Ended				MC
Single Model	All	Y/N	No.	Other	All
SMem[38]	58.24	80.8	37.53	43.48	-
SAN[40]	58.85	79.11	36.41	46.42	-
D-NMN[1]	59.4	81.1	38.6	45.5	-
ACK[36]	59.44	81.07	37.12	45.83	-
FDA[11]	59.54	81.34	35.67	46.10	64.18
QRU[22]	60.76	-	-	-	65.43
HYBRID[14]	60.06	80.34	37.82	47.56	-
DMN+[37]	60.36	80.43	36.82	48.33	-
MRN[16]	61.84	82.39	38.23	49.41	66.33
HieCoAtt[24]	62.06	79.95	38.22	51.95	66.07
RAU[28]	63.2	81.7	38.2	52.8	67.3
MLB[17]	65.07	84.02	37.90	54.77	68.89
MF-SIG-T3	**65.88**	**84.42**	**38.94**	**55.89**	**70.33**
Ensemble Model					
MCB[8]	66.47	83.24	39.47	58.00	70.10
MLB[17]	66.89	84.61	39.07	57.79	70.29
Ours	**68.14**	**85.41**	**40.99**	**59.27**	**72.08**
Ours `test2017`	65.84	81.85	43.64	57.07	-

Table 4: Results of the Open Ended and Multiple Choice tasks on `test`. We compare the accuracy of single models (without augmentation) and ensemble models with published methods.

which MF-SIG surpassed MLB under the same condition by 1.35%. The accuracy boost of our model is higher than MCB and MLB, showing that our model has higher capacity. The LBP models, which performs better than MF layers on CLEVR, turns out to be worse on this dataset, and $T = 1$ is the optimal choice for LBP. We also find the single MF attention model, which should not be as powerful as MF-SIG, achieved 67.17% accuracy with augmentation. These might be caused by the bias of the current VQA dataset [3], where there are questions with fixed answers across all involved images. We also show the results on `test`, as shown in Table 4. Our model is the best among published methods without external data. With an ensemble of 3 MF-T3 and 4 MF-SIG-T3 models, we achieve 68.18% accuracy on `test`, 1.25% higher than best published ensemble model on the



Q: What is the size of the object that is both right of the cyan sphere and left of the tiny red metallic object?
MF: small. LBP: small. SIG: large.



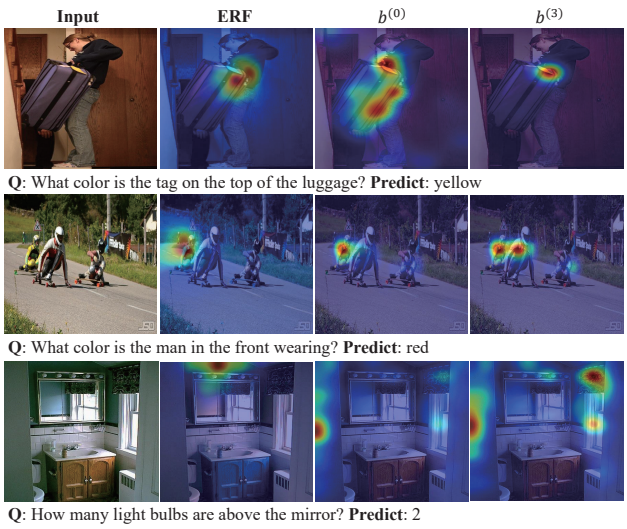
Q: the tiny gray object that is made of the same material as the gray cube is what shape?
MF: sphere. LBP: sphere. SIG: cube.

Figure 5: Two instances of different attentions on CLEVR, where the SIG model gives wrong answers but MF-SIG and LBP-SIG both give the correct answer. For each instance, from left to right, the first row to the second row, the images are: input image, $b^{(0)}$ of MF-SIG, $b^{(3)}$ of MF-SIG, $\psi_i(z_i)$ of LBP-SIG, b of LBP-SIG, attention of SIG. Notations are the same as in Fig. 3. Best viewed in color.

Open Ended task. By the date of submission, we rank the second on the leaderboard of Open Ended task and the first on that of the Multiple Choice task. The champion on Open Ended has an accuracy of 69.94% but the method is not published. We have also recorded our model's performance on the `test-dev2017` and `test2017` of VQA2.0 in Table 3 and 4. Accuracy on `test2017` is achieved with 8 snapshots from 4 models with different learning rates.

4.4.4 Qualitative Results

We demonstrate some attention maps on CLEVR and the VQA dataset to analyze the behavior of the proposed models. Fig. 5 shows 2 instances where the SIG model failed but both MF and LBP succeeded. We find the MF-SIG model has learned interesting patterns where its attention often covers the background surrounding the target initially, but converges to the target after iterative inference. This phenomenon almost never happens with the LBP-SIG model, which usually has better initializations that contained the target region. The shortcoming of the unstructured SIG model is also exposed in the 2 instances, where it tends to get stuck with the key nouns of the question. Fig. 6 demonstrates 3 instances of the MF-SIG model together with the effective receptive field. The model gives 2 correct answers for the first 2 instances and 1 wrong answer for the last instance. In the first instance, the ERF at the target should be enough to encode the relations. The initial attention involves some extra areas due to the key word "luggage", but it manages to converge to the most relevant region. In the second instance, the initial attention is wrong, as we can see the ERF at the initial attention does not overlap with the target, but with the help of MF, the final attention captures the relation "in the front" and gives an acceptable answer. In the third instance, the ERF at the target region is very weak on the keyword "bulb", which means the feature vector does not encode this concept, probably due to the size of the bulb. The model fails to attend to the right region and gives a popular answer "2" (3rd most popular answer on the VQA dataset) according to the type of the question.



Q: What color is the tag on the top of the luggage? **Predict**: yellow

Q: What color is the man in the front wearing? **Predict**: red

Q: How many light bulbs are above the mirror? **Predict**: 2

Figure 6: Some instances in the VQA dataset. The ERFs locate at the target region in row 1 and 3, and at at initial attention in row 2. Best viewed in color.

5. Conclusion

In this paper, we propose a novel structured visual attention mechanism for VQA, which models attention with binary latent variables and a grid-structured CRF over these variables. Inference in the CRF is implemented as recurrent layers in neural networks. Experimental results demonstrate that the proposed method is capable of capturing the semantic structure of the image in accordance with the question, which alleviates the problem of unstructured attention that captures only the key nouns in the questions. As a result, our method achieves state-of-the-art accuracy on three challenging datasets. Although structured visual attention does not solve all problems in VQA, we argue that it should be an indispensable module for VQA in the future.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *NAACL*, 2016.

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, 2016.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual question answering. In *ICCV*, 2015.

[4] L.-C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun. Learning deep structured models. In *ICML*, 2015.

[5] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[6] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *EMNLP*, 2016.

[7] T.-M.-T. Do and T. Artieres. Neural conditional random fields. In *AISTATS*, 2010.

[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. *EMNLP*, 2016.

[9] Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *NIPS*, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] I. Ilievski, S. Yan, and J. Feng. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485*, 2016.

[12] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. *ICLR*, 2015.

[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. *CVPR*, 2017.

[14] K. Kafle and C. Kanan. Answer-type prediction for visual question answering. In *CVPR*, 2016.

[15] K. Kafle and C. Kanan. Visual question answering: Datasets, algorithms, and future challenges. *arXiv preprint arXiv:1610.01465*, 2016.

[16] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual QA. In *NIPS*, 2016.

[17] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *ICLR*, 2017.

[18] Y. Kim, C. Denton, L. Hoang, and A. M. Rush. Structured attention networks. *ICLR*, 2017.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[20] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, 2015.

[21] 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. *IJCV*, 2016.

[22] R. Li and J. Jia. Visual question answering with question representation update (QRU). In *NIPS*, 2016.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[24] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.

[25] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NIPS*, 2016.

[26] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.

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

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

[29] S. Nowozin, C. H. Lampert, et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365, 2011.

[30] J. Peng, L. Bo, and J. Xu. Conditional neural fields. In *NIPS*, 2009.

[31] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.

[32] S. E. Shimony. Finding maps for belief networks is np-hard. *Artificial Intelligence*, 68(2):399–410, 1994.

[33] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[34] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 2014.

[35] Y. Weiss. Comparing the mean field method and belief propagation for approximate inference in mrfs. *Advanced Mean Field MethodsTheory and Practice*, pages 229–240, 2001.

[36] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016.

[37] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *ICML*, 1603, 2016.

[38] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.

[39] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[40] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[41] T. Yeh, J. J. Lee, and T. Darrell. Photo-based question answering. In *ACMMM*, 2008.

[42] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

[43] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.