

Multimodal Unified Attention Networks for Vision-and-Language Interactions

Zhou Yu, *Member, IEEE*, Yuhao Cui, Jun Yu, *Member, IEEE*, Dacheng Tao, *Fellow, IEEE*, Qi Tian *Fellow, IEEE*

Abstract—Learning an effective attention mechanism for multimodal data is important in many vision-and-language tasks that require a synergic understanding of both the visual and textual contents. Existing state-of-the-art approaches use co-attention models to associate each visual object (*e.g.*, image region) with each textual object (*e.g.*, query word). Despite the success of these co-attention models, they only model inter-modal interactions while neglecting intra-modal interactions. Here we propose a general ‘unified attention’ model that simultaneously captures the intra- and inter-modal interactions of multimodal features and outputs their corresponding attended representations. By stacking such unified attention blocks in depth, we obtain the deep Multimodal Unified Attention Network (MUAN), which can seamlessly be applied to the visual question answering (VQA) and visual grounding tasks. We evaluate our MUAN models on two VQA datasets and three visual grounding datasets, and the results show that MUAN achieves top level performance on both tasks without bells and whistles.

Index Terms—Multimodal learning, visual question answering (VQA), visual grounding, unified attention, deep learning.

I. INTRODUCTION

Deep learning in computer vision and natural language processing has facilitated recent advances in artificial intelligence. Such advances drive research interest in multimodal learning tasks lying at the intersection of vision and language such as multimodal embedding learning [1][2][3], visual captioning [4][5], visual question answering (VQA) [6] and visual grounding [7], etc. In these tasks, learning a fine-grained semantic understanding of both visual and textual content is key to their performance.

The attention mechanism is a predominant focus of recent deep learning research. It aims to focus on certain data elements, and aggregate essential information to obtain a more discriminative local representation [8], [4]. This mechanism has improved the performance of a wide range of unimodal learning tasks (*e.g.*, vision [9], [10], [11], language [12], [13],

This work was supported in part by National Natural Science Foundation of China under Grant 61702143 and Grant 61836002, and in part by the Australian Research Council Projects under Grant FL-170100117. (Zhou Yu and Yuhao Cui contribute equally to this work. Jun Yu is the corresponding author.)

Z. Yu, Y. Cui and J. Yu are with Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, P. R. China (e-mail: yuz@hdu.edu.cn; yujun@hdu.edu.cn; cuiyh@hdu.edu.cn)

D. Tao is with the UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Australia (e-mail: dacheng.tao@sydney.edu.au).

Q. Tian is with the Noah’s Ark Lab, Huawei, P. R. China (e-mail: tian.qi1@huawei.com).

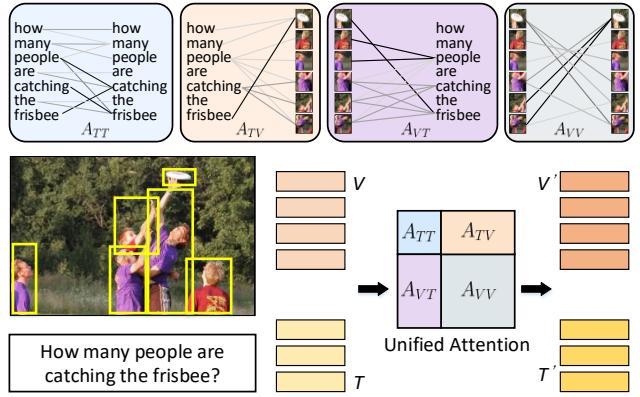


Fig. 1: Schematic of the proposed unified attention, which simultaneously models inter- and intra-modal interactions in a single framework. Given multimodal inputs V and T , A_{VV} , A_{TT} denote the intra-modal interactions within each modality, while A_{VT} and A_{TV} denote the inter-modal interactions across different modalities. V' and T' are the attended features for V and T respectively.

[14]) in conjunction with deep convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

For the multimodal learning tasks described above, attention learning considers the inputs from both the visual and textual modalities. Taking the VQA problem in Fig. 1 as an example, to correctly answer a question like ‘*How many people are catching the frisbee*’ for an image, the attention model should ideally learn to focus on particular image regions (*i.e.*, the person near the frisbee). Such visual attention based models have become an integral component in many multimodal tasks that require fine-grained visual understanding [4][15][16]. Beyond the visual attention models, recent studies have introduced co-attention models, which simultaneously learn the visual attention and textual attention to benefit from fine-grained representations for both modalities. Early approaches learned separate attention distributions for each modality in an iterative manner, neglecting the dense interaction between each question word and image region [17][18]. To address this problem, dense co-attention models have been proposed to capture complete interactions between word-region pairs, which are further extended to form deep co-attention models [19][20].

Despite the success of the co-attention models in multi-

modal learning tasks, these models only consider inter-modal interactions (*i.e.*, A_{TV} or A_{VT} in Fig. 1) while neglecting intra-modal ones (*i.e.*, A_{TT} and A_{VV}). On the other hand, modeling intra-modal interactions has been proved to be beneficial for many unimodal learning tasks [21][22][23][24]. We argue that intra-modal interactions within each modality provide complementary and important information to the inter-modal interactions.

Inspired by the famous self-attention model [21] in the NLP community, we naturally extend its idea for multimodal data and propose a *unified attention* accordingly. Our unified attention model characterizes the intra- and inter-modal interactions jointly in a unified framework which we call the unified attention (UA) block (see Fig. 1). The attention map learned from the UA block includes four relationships: the inter-modal interactions (A_{VT} and A_{TV}) to build co-attention across different modalities, and the intra-modal interactions (A_{VV} and A_{TT}) to build self-attention within each modality. The learned unified attention is further used to obtain the attended output features for multimodal inputs. By stacking such UA block in depth, we obtain the Multimodal Unified Attention Network (MUAN), which can be trained in an end-to-end manner to perform deep multimodal reasoning.

To evaluate the effectiveness of our proposed MUAN model, we apply it to for VQA and visual grounding. The quantitative and qualitative results on two VQA dataset VQA-v2 [25] and CLEVR [26], and three visual grounding datasets RefCOCO [27], RefCOCO+ [27] and RefCOCOg [28] show that MUAN achieves top level performance on both tasks without using any dataset specific model tuning.

In summary, we have made the following contributions in this study:

- We extend the self-attention model for single modality to a unified attention model, which can characterize intra- and inter-modal interactions of multimodal data. By stacking such unified attention model (*i.e.*, UA block) in depth, we obtain a neat multimodal unified attention network (MUAN), which can perform accurate multimodal reasoning.
- We modify the original self-attention model to a gated self-attention (GSA) model as the basic component for the UA block, which facilitates more accurate and robust attention learning and leads to more discriminative features for specific tasks.
- We apply MUAN to two multimodal learning tasks, namely VQA and visual grounding. The results on five benchmark datasets show the superiority of MUAN over existing state-of-the-art approaches.

II. RELATED WORK

We briefly review existing studies on VQA and visual grounding, and establish a connection between these two tasks by attention learning.

Visual Question Answering (VQA). VQA aims to answer a question in natural language with respect to a given image, so requires multimodal reasoning over multimodal inputs. Since Antol *et al.* presented a large-scale VQA benchmark

dataset with free-form questions [6], multimodal fusion and attention learning have become two major research focuses for VQA. For multimodal fusion, early methods used simple concatenation or element-wise multiplication between multimodal features [29][6]. Fukui *et al.* [16], Kim *et al.* [30], Yu *et al.* [18] and Ben *et al.* [31] proposed different approximated bilinear pooling methods to effectively integrate the multimodal features with second-order feature interactions. For attention learning, question-guided visual attention on image regions has become the de-facto component in many VQA approaches [15][32]. Chen *et al.* proposed a question-guided attention map that projects the question embeddings to the visual space and formulates a configurable convolutional kernel to search the image attention region [32]. Yang *et al.* proposed a stacked attention network to learn the attention iteratively [15]. Some approaches introduce off-the-shelf object detectors [33] or object proposals [34] as the candidates of the attention regions and then use the question to identify the relevant ones. Taken further, co-attention models that consider both textual and visual attentions have been proposed [17][18]. Lu *et al.* proposed a co-attention learning framework to alternately learn the image attention and question attention [17]. Yu *et al.* reduced the co-attention method into two steps, self-attention for a question embedding and the question-conditioned attention for a visual embedding [35]. The learned co-attentions by these approaches are coarse, in that they neglect the interaction between question words and image regions. To address this issue, Nguyen *et al.* [20] and Kim *et al.* [19] introduced dense co-attention models that established the complete interaction between each question word and each image region.

Visual Grounding. Visual grounding (*a.k.a.*, referring expression comprehension) aims to localize an object in an image referred to in query text. Most previous approaches follow a two-stage pipeline [7][36][16]: 1) use an off-the-shelf object detector, such as Edgebox [37] or Faster R-CNN [38] to generate a set of region proposals along with the proposal features for the input image; and 2) compute a matching score between each proposal feature and query feature and adopt the proposal (or its refined bounding box [39]) with the highest score as the referent. From the attention learning point of view, visual grounding represents a task of learning query-guided attention on the image region proposals. The aforementioned two-stage approaches are analogous to the visual attention models in VQA. Yu *et al.* [40], Zhang *et al.* [41] and Deng *et al.* [42] also modeled the attention on question words along with visual attention, providing a connection to the co-attention model in VQA.

Joint Modeling of Self- and Co-Attention. Although extensive studies on self-attention and co-attention have been made by existing multimodal learning methods, the two kinds of attentions are usually considered solely. To the best of our knowledge, only a few attempts have modeled intra- and inter-modal interactions jointly. Li *et al.* introduced a videoQA approach which used self-attention to learn intra-modal interactions of video and question modalities respectively, and then fed them through a co-attention block to model inter-modal interactions [43]. Gao *et al.* presented a dynamic fusion framework for VQA with modeling intra- and

inter-modal attention blocks. [44]. Yu *et al* applied a modular co-attention network for VQA which stacked multiple self-attention and guided-attention blocks in depth to perform deep visual reasoning. In summary, all these methods models the self-attention and co-attention in two sequential stages, which is sub-optimal and may result in serious information lose. This inspires us to design a general unified attention framework to simultaneously model the two attentions in one stage.

III. MULTIMODAL UNIFIED ATTENTION

In this section, we introduce the multimodal *unified attention*, which is the basic component of our Multimodal Unified Attention Network (MUAN). Taking the multimodal input features X from the image modality and Y from the text modality, the unified attention outputs their corresponding attended features. In contrast to existing visual attention methods, which model unidirectional inter-modal interactions (*i.e.*, $X \rightarrow Y$) [16][30], or the co-attention methods, which model bidirectional inter-modal interactions (*i.e.*, $X \leftrightarrow Y$) [19][20], our unified attention models the intra-modal and inter-modal interactions simultaneously (*i.e.*, $X \rightarrow X$, $Y \rightarrow Y$ and $X \leftrightarrow Y$) in a general framework.

Inspired by the *self-attention* model which has achieved remarkable performance in natural language processing [21][45][22], we design a unified attention model for multimodal data. Furthermore, to obtain more accurate attention map in the unified attention learning, we introduce a bilinear pooling based gating model to reweight the importance of input features, which can to some extent eliminate the irrelevant or noisy features.

A. Gated Self-Attention

The self-attention model proposed in [21] takes a group of input features $X = [x_1; \dots; x_m] \in \mathbb{R}^{m \times d_x}$ and outputs a group of attended features $F = [f_1, \dots, f_m] \in \mathbb{R}^{m \times d}$, where m is the number of samples, d_x and d are the dimensionalities of input and output features, respectively. To achieve this goal, X is first fed into three independent fully-connected layers.

$$Q = \text{FC}_q(X), \quad K = \text{FC}_k(X), \quad V = \text{FC}_v(X) \quad (1)$$

where $Q, K, V \in \mathbb{R}^{m \times d}$ are three feature matrices of the same shape, corresponding to the queries, keys, and values, respectively.

Given a query $q \in Q$ and all keys K , we calculate the dot-products of q with K , divide each by a scaling factor \sqrt{d} and apply the softmax function to obtain the attention weights on the values. In practice, the attention function can be computed on all queries Q simultaneously, and in doing so we obtain the output features F as follows:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \quad (2)$$

$$F = AV \quad (3)$$

where $A \in \mathbb{R}^{m \times m}$ is the attention map containing the attention weights for all query-key pairs, and the output features F are the weighted summation of the values V determined by A .

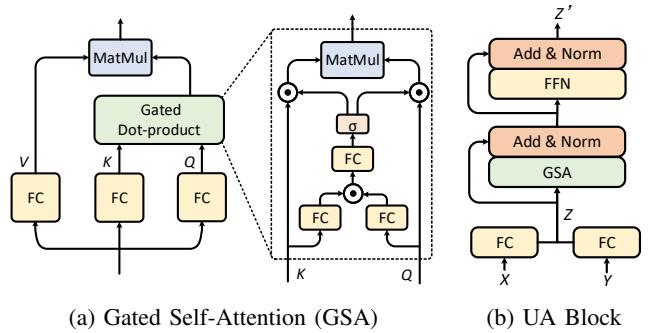


Fig. 2: Flowcharts of the Gated Self-Attention (GSA) model and unified attention (UA) block for multimodal data

Learning an accurate attention map A is crucial for self-attention learning. The scaled dot-product attention in Eq.(2) models the relationship between feature pairs. However, the importance of each individual features is not explicitly considered during attention learning. Consequently, irrelevant or noisy features may have a negative impact on the attention map, resulting in inaccurate output features. To address this problem, we introduce a novel *gating model* into Eq.(2) to improve the quality of the learned attention. Inspired by the bilinear pooling models which have been in fine-grained visual recognition [46] and multi-modal fusion [30], we design a gating model based on low-rank bilinear pooling to reweight the features of Q and K before their scaled dot-products:

$$M = \sigma(\text{FC}_q^g(Q) \odot \text{FC}_k^g(K)) \quad (4)$$

where $\text{FC}_q^g, \text{FC}_k^g \in \mathbb{R}^{d \times d_g}$, $\text{FC}^g \in \mathbb{R}^{d_g \times 2}$ are three independent fully-connected layers, and d_g is the dimensionality of the projected space. \odot denotes the element-wise product function and $\sigma(\cdot)$ the sigmoid function. $M \in \mathbb{R}^{m \times 2}$ corresponds to the two masks $M_q \in \mathbb{R}^m$ and $M_k \in \mathbb{R}^m$ for the features Q and V , respectively.

The learned two masks M_q and M_k are tiled to $\tilde{M}_q, \tilde{M}_k \in \mathbb{R}^{m \times d}$ and then used to formulate a gated self-attention (GSA) model as follows:

$$A^g = \text{softmax}\left(\frac{(Q \odot \tilde{M}_q)(K \odot \tilde{M}_k)^T}{\sqrt{d}}\right) \quad (5)$$

$$F = A^g V \quad (6)$$

Fig. 2a illustrates the flowchart of our gated self-attention model. Similar to [21], the multi-head strategy is introduced in our model to attain more diverse attention.

B. Unified Attention Block

Based on the gated self-attention model above, we introduce the multimodal unified attention block, which simultaneously models intra- and inter-modal interactions.

Given a group of textual features (*e.g.*, question words) $X \in \mathbb{R}^{m \times d_x}$ and a group of visual features (*e.g.*, image regions) $Y \in \mathbb{R}^{n \times d_y}$, we first learn two fully-connected layers FC_x and FC_y to embed X and Y into a d_z -dimensional common space, and then concatenate the two groups of embedded features on rows to form a unified feature matrix Z :

$$Z = [\text{FC}_x(X); \text{FC}_y(Y)] \quad (7)$$

where $Z = [z_1, \dots, z_s] \in \mathbb{R}^{s \times d_z}$ with $s = m + n^1$.

The UA block (see Fig. 2b) consists of a gated self-attention (GSA) module and a feed-forward network (FFN) module. Taking the unified feature matrix Z as input, the GSA module learns the pairwise interactions between the sample pairs $\langle z_i, z_j \rangle$ within Z . Since z_i and z_j may come from different (or the same) modalities, the intra- and inter-modal relationships are represented at the same time. Compared to existing co-attention models, which only model the inter-modal relationships [19][20], the intra-modal relationships (e.g., word-to-word or region-to-region) are also important for understanding the intrinsic structure within each modality, thus facilitating more accurate visual reasoning. The FFN module takes the output features of the GSA module as input, and then performs transformation through two consecutive fully-connected layers ($\text{FC}(4d)$ -ReLU-Drop(0.1)- $\text{FC}(d)$). To simplify optimization, shortcut connection [38] and layer normalization [47] are applied after the GSA and FFN modules. It is worth noting that the final output features Z' of the UA block are of the same shape as the input features Z , making it possible to stack multiple UA blocks in depth².

IV. MULTIMODAL UNIFIED ATTENTION NETWORKS

In this section, we describe the MUAN architectures for VQA and visual grounding (see Fig. 3). The core component of both models is the deep MUAN- L model, which consists of L UA blocks stacked in depth to perform deep multimodal reasoning and attentional feature transformation. The proposed VQA model and the visual grounding model are very similar to each other, except for the input feature representations and the loss functions used during model training. We therefore highlight these two parts in each model.

A. Architecture for VQA

Image and Question Representations. The inputs for VQA consist of an images and a question, and the goal is to predict an answer to the question. Our model first extracts representations for the image and the question and then feeds the multimodal features into the MUAN model to output their corresponding output features with unified attention learning. Finally, one of the attended feature is fed to a multi-label classifier to predict the correct answer.

The input question is first tokenized into a sequence of words, and then trimmed (or zero padded) to a maximum length of m . Similar to [22], we add a dummy token [ans] at the beginning of the question, and the attended feature of this token will be used to predict the answer. These words are firstly represented as one-hot vectors and then transformed to 300-D word embeddings using the pre-trained GloVe model [48]. Finally, the word embeddings are fed into a one-layer LSTM network [49] with d_x hidden units, resulting in the final question feature $X \in \mathbb{R}^{(m+1) \times d_x}$. The input image is represented as a group of d_y -dimensional visual features

¹In our implementation, we let $d_x = d_z = d$ and omit $\text{FC}_x(\cdot)$ for simplicity, and rewrite Eq.(7) as $Z = [X; \text{FC}_y(Y)]$

²For multiple UA blocks stacked in depth, only the first block needs to handle multimodal inputs. Eq.(7) is omitted in the other blocks.

extracted from a pre-trained CNN model [38] or a pre-trained object detector [50]. This results in the image feature $Y \in \mathbb{R}^{n \times d_y}$, where n is the number of extracted features.

Note that we mask the zero-padded features during attention learning to make their attention weights all zero.

MUAN- L . The multimodal features X and Y are fed into a deep MUAN- L model consisting of L UA blocks [$\text{UA}^{(1)}, \text{UA}^{(2)}, \dots, \text{UA}^{(L)}$]. For $\text{UA}^{(1)}$, X and Y are integrated by Eq.(7) to obtain the initialized unified features $Z^{(0)}$, which are further fed to the remaining UA blocks in a recursive manner.

$$Z^{(l+1)} = \text{UA}^{(l+1)}(Z^{(l)}) \quad (8)$$

where $l \in [0, L - 1]$. Note that the final output features $Z^{(L)}$ are the same shape as the input features $Z^{(0)}$, and each paired $\langle z_i^{(0)}, z_i^{(L)} \rangle$ has a one-to-one correspondence.

Answer Prediction. Using the attended features $Z^{(L)}$ from MUAN- L , we project the first feature $z_1^{(L)}$ (the [ans] token) into a vector $p \in \mathbb{R}^k$, where k corresponds to the size of the answer vocabulary.

For the datasets that have multiple answers to each question, we follow the strategy in [51] and use the binary cross-entropy (BCE) loss to train an k -way classifier with respect to the ground-truth label $y \in \mathbb{R}^k$:

$$\mathcal{L} = \sum_{i=1}^k (y_i \log(\sigma(p_i)) + (1 - y_i) \log(1 - \sigma(p_i))) \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid activation function.

For the datasets that have exactly one answer to each question, we use the softmax cross-entropy loss to train the model with respect to the one-hot ground-truth label $y \in \{0, 1\}^k$:

$$\mathcal{L} = -y^T \log \text{softmax}(p) \quad (10)$$

B. Architecture for Visual Grounding

The inputs for visual grounding consist of an image and a query. Similar to the VQA architecture above, we extract the query features $X \in \mathbb{R}^{m \times d_x}$ using GloVe embeddings followed by a LSTM network, and extract the region-based proposal features $Y \in \mathbb{R}^{n \times d_y}$ for the image using an pre-trained object detector. Note that we do not use the dummy token for visual grounding which is specially designed for VQA.

The multimodal input features are integrated and transformed by MUAN- L to output their attended representations. On top of the attended feature for each region proposal, we append two fully-connected layers to project each attended feature $z^{(L)} \in Z^{(L)}$ into a score $s \in \mathbb{R}$ and a 4-D vector $t \in \mathbb{R}^4$ to regress the refined bounding box coordinates for the proposal, respectively.

$$s = \text{FC}(z^{(L)}); \quad t = \text{FC}(z^{(L)}) \quad (11)$$

Accordingly, a ranking loss L_{rank} and a regression loss L_{reg} are designed to optimize the model in a multitask learning manner. Following the strategy in [39], KL-divergence is used as the ranking loss:

$$\mathcal{L}_{\text{rank}} = \frac{1}{n} \sum_{i=1}^n s_i^* \log\left(\frac{s_i^*}{s_i}\right) \quad (12)$$

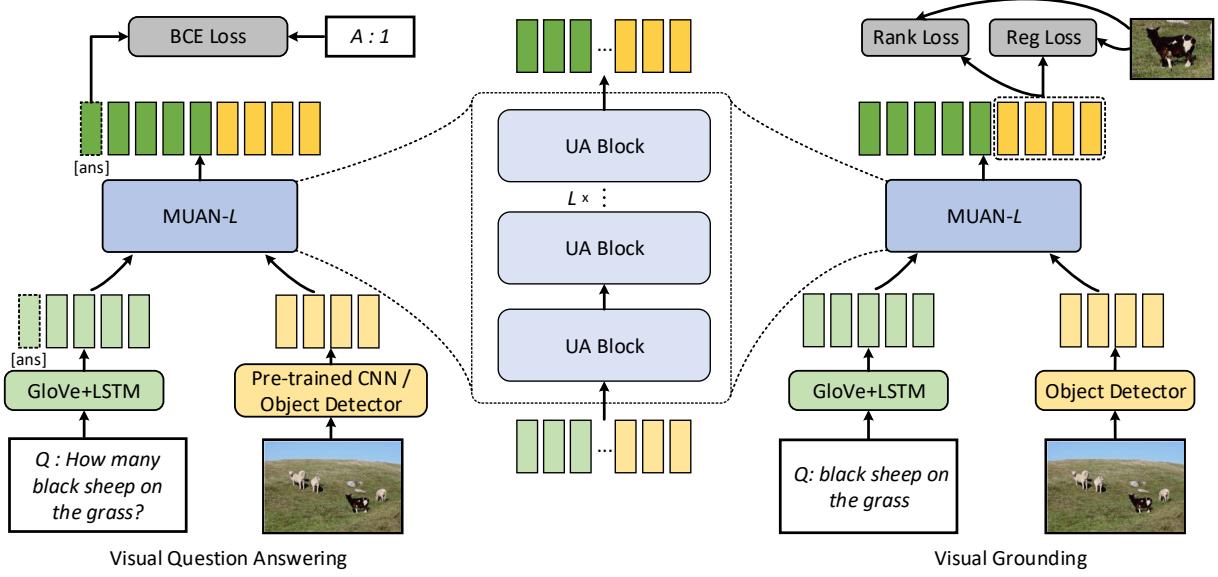


Fig. 3: Architectures of the Multimodal Unified Attention Networks (MUAN) for visual question answer (left) and visual grounding (right), respectively. Both architectures contain the a MUAN- L model which consists of L stacked UA blocks to output the features with unified attention learning. For VQA, we add a dummy token [ans] at the beginning of the question, and use its attended feature to predict the answer. For visual grounding, the attended features of the region proposals are used to predict their ranking scores and refined bounding boxes.

where $S = [s_1, s_2, \dots, s_n] \in \mathbb{R}^n$ are the predicted scores for n proposals. The ground-truth label $S^* = [s_1^*, s_2^*, \dots, s_n^*] \in \mathbb{R}^n$ is obtained by calculating the IoU scores of all proposals w.r.t. the unique ground-truth bounding box and assign the IoU score of the i -th proposal to s_i^* if the IoU score is larger than a threshold η and 0 otherwise. Softmax normalizations are respectively applied to S and S^* to make them form a score distribution.

The smoothed ℓ_1 loss [52] is used as the regression loss to penalize the differences between the refined bounding box and the ground-truth bounding box:

$$\mathcal{L}_{\text{reg}} = \frac{1}{n} \sum_{i=1}^n \text{smooth}_{\ell_1}(t_i^*, t_i) \quad (13)$$

where $t_i \in \mathbb{R}^4$ and $t_i^* \in \mathbb{R}^4$ correspond to the coordinates of the predicted bounding box and the ground-truth bounding box for i -th proposal, respectively.

By combining the two terms, we obtain the overall loss function L_{all} as follows:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{reg}} \quad (14)$$

where λ is a hyper-parameter to balance the two terms.

V. EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of the MUAN models in VQA and visual grounding tasks. We conduct extensive ablation experiments to explore the effect of different hyper-parameters in MUAN. Finally, we compare the best MUAN models to current state-of-the-art methods on five benchmark datasets (two VQA datasets and three visual grounding datasets).

A. Datasets

VQA-v2 is a commonly-used benchmark dataset for open-ended VQA [25]. It contains human annotated question-answer pairs for MS-COCO images [53]. The dataset is split into three subsets: train (80k images with 444k questions); val (40k images with 214k questions); and test (80k images with 448k questions). The test subset is further split into test-dev and test-std sets that are evaluated online with limited attempts. For each questions, multiple answer are provided by different annotators. To evaluate the performance of a model with respect to such multi-label answers, an accuracy-based evaluation metric is defined as follows which is robust to inter-human variability in phrasing the answer a [6]:

$$\text{Accuracy}(a) = \min \left\{ \frac{\text{count}(a)}{3}, 1 \right\} \quad (15)$$

where $\text{count}(a)$ is a function that count the answer a voted by different annotators.

CLEVR is a synthesized dataset containing 100k images and 853k questions [26]. Each image contains 3D-rendered objects and is associated with a number of questions that test various aspects of visual reasoning including attribute identification, object counting, and logical operations. The whole dataset is split into three subsets: train (70k images with 700k questions), val (15k images with 150k questions) and test (15k images with 15k questions). Each question is associated with exactly one answer and standard accuracy metric is used to evaluate model performance.

RefCOCO, RefCOCO+, and RefCOCOg are three datasets to evaluate visual grounding performance. All three datasets are collected from MS-COCO images [53], but the queries are

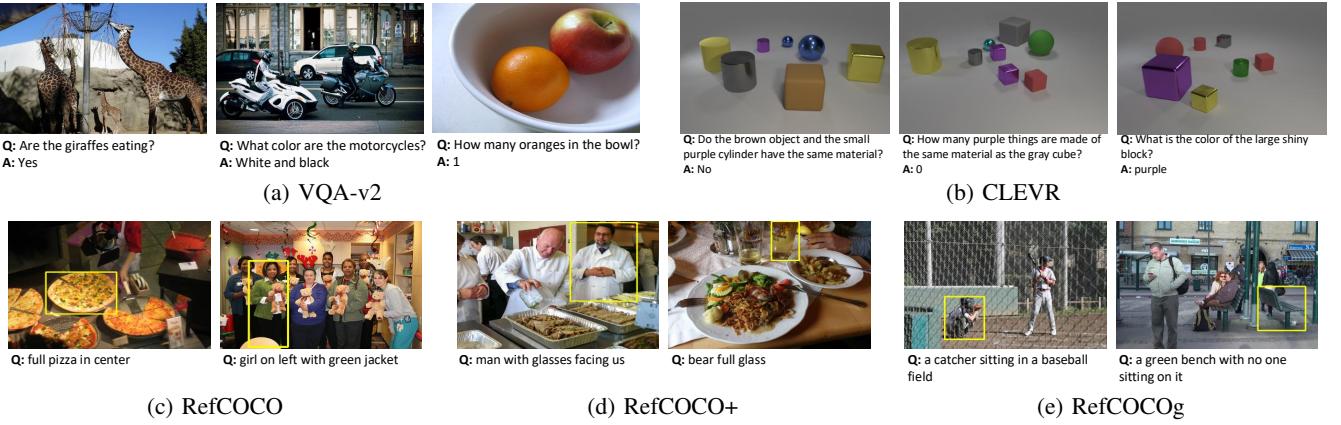


Fig. 4: Typical examples from VQA-v2, CLEVR, RefCOCO, RefCOCO+, and RefCOCOg.

different in three respects: 1) RefCOCO [27] and RefCOCO+ [27] contains short queries (3.6 words on average) while RefCOCOg [28] contains relatively long queries (8.4 words on average); 2) RefCOCO and RefCOCO+ contain 3.9 same-type objects on average, while in RefCOCOg this number is 1.6; and 3) RefCOCO+ does not contain any location word, while the counterparts do not have this constraint. RefCOCO and RefCOCO+ are split into four subsets: train (120k queries), val (11k queries), testA (6k queries about people), and testB (5k queries about objects). RefCOCOg is split into three subsets: train (81k queries), val (5k queries), and test (10k queries). For all the three datasets, accuracy is adopted as the evaluation metric, which is defined as the percentage in which the predicted bounding box overlaps with the ground-truth bounding box by $\text{IoU} > 0.5$.

Fig. 4 shows some typical examples from these datasets.

B. Experimental Setup

Universal Setup. We use the following hyper-parameters as the default settings for MUAN unless otherwise noted. In each UA block, the latent dimensionality d is 768 and the number of heads h is 8, so the dimensionality of each head is $d/h = 96$. The latent dimensionality in the gating model d_g is 96. The number of UA blocks L ranges from 2 to 12.

All the models are optimized using the Adam solver [54] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The models (except those for CLEVR) are trained up to 13 epochs with a batch size 64 and a base learning rate α set to $1.5e^{-2}/\sqrt{dL}$. Similar to [19], the learning rate is warmed-up for 3 epochs and decays by 1/5 every 2 epochs after 10 epochs. We report the best results evaluated on the validation set. For CLEVR, a smaller base learning rate $\alpha = 3.5e^{-3}/\sqrt{dL}$ is used to train up to 20 epochs and decay by 1/5 at the 16th and 18th epochs, respectively.

VQA Setup. For VQA-v2, we follow the strategy in [51] and extract the pool15 feature for each object from a Faster R-CNN model (with a ResNet-101 backbone) [55] pre-trained on the Visual Genome dataset [56], resulting in the input visual features $Y \in \mathbb{R}^{n \times 2048}$, where $n \in [10, 100]$ is the number of extracted objects with a confidence threshold. The maximum number of question words $m = 14$, and the size of the answer

vocabulary $k = 3129$, which corresponds to answers appearing more than 8 times in the training set. For CLEVR, we follow the strategy in [57] and extract the res4b22 features from a ResNet-101 model pre-trained on ImageNet [38], resulting in the image features $Y \in \mathbb{R}^{196 \times 1024}$. The maximum number of question words $m = 43$, and the size of the answer vocabulary $k = 28$.

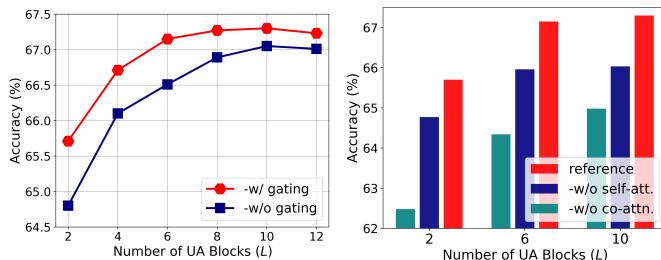
Visual Grounding Setup. We use the same settings for the three evaluated datasets. To detect proposals and extract their visual features for each image, we use two pre-trained proposal detectors as previous works did: 1) a Faster R-CNN model [55] pre-trained on the Visual Genome dataset [39]; and 2) a Mask R-CNN model [58] pre-trained on MS-COCO dataset [40]. During the training data preparation for the proposal detectors, we exclude the images in the training, validation and testing sets of RefCOCO, RefCOCO+ and RefCOCOg to avoid contamination of the used visual grounding datasets. Each of the obtained proposal visual features is further concatenated with a spatial feature containing the bounding-box coordinates of the proposal³. This results in the image features $Y \in \mathbb{R}^{100 \times 4096}$. The maximum number of question words m is 15 and the loss weight λ is 0.5.

C. Ablation Studies

We run a number of ablation experiments on VQA-v2 to explore the effectiveness of MUAN.

First, we explore the effectiveness of the gating mechanism for the UA block with respect to different number of block L . In Fig. 5a, we report the overall accuracies of the MUAN- L models (L ranges from 2 to 12) with the gating mechanism (*i.e.*, Eq.(5)) or without the gating mechanism (*i.e.*, Eq.(2)) for the UA block. From the results, we can see that MUAN with the gating model steadily outperforms counterpart without the gating model. Furthermore, increasing L consistently improves the accuracies of both models, which finally saturate at $L = 10$. We think the saturation is caused by over-fitting. To train a deeper model we may require more training data [22].

³For each proposal, we first extract a 5-D spatial feature $[x_{tl}/W, y_{tl}/H, x_{br}/W, y_{br}/H, wh/WH]$ proposed in [59], and then linearly transform it to a 2048-D feature with a fully-connected layer to match the dimensionality of a 2048-D proposal visual feature.



(a) Effect of gating mechanism (b) Effect of self- and co-attention

Fig. 5: Ablation of the MUAN models with the number of UA blocks L ranges from 2 to 12. All results are evaluated on the val split of VQA-v2. (a) Results of MUAN- L variants with or without the gating mechanism. (b) Results of reference MUAN model along with the variants without modeling self-attention (A_{TT} and A_{VV}) or co-attention (A_{TV} and A_{VT}).

TABLE I: Ablation of the MUAN-10 models with different hyper-parameters. All results are reported on the val split of VQA-v2. Unlisted hyper-parameter values are identical to those of the reference model.

	d	h	d/h	d_g	Acc. (%)	#Param ($\times 10^6$)
ref.	768	8	96	96	67.28	83.0
(A)				32	67.16	82.9
				64	67.27	82.9
				128	67.17	83.1
(B)			6	128	67.11	83.1
			12	64	67.23	82.9
			16	48	67.25	82.9
(C)	256		32		66.30	14.5
	512		64		66.92	40.6
	1024		128		67.30	141.6

Next, we conduct the ablation studies to explore the effects of self-attention and co-attention in MUAN. By masking the values in the self-attention part (*i.e.*, A_{TT} and A_{VV}) or the co-attention part (*i.e.*, A_{TT} and A_{VV}) to $-\infty$, we obtain two degraded variants of MUAN. We compare the two MUAN variants to its reference model in Fig. 5b with $L \in \{2, 6, 10\}$. The results show that: 1) both the self-attention and co-attention in MUAN contribute to the performance of VQA; and 2) co-attention plays a more important role than self-attention in MUAN, especially when the model is relatively shallow.

Finally, we investigate MUAN-10 model performance with different hyper-parameters for the UA block in Table I. In row (A), we vary the dimensionality d_g in the gating model. The results suggest that the reference model results in a 0.12 point improvement over the worst counterpart. Further, the model sizes of these variants are almost identical, indicating that the computational cost of the gating model can be more or less ignored. In row (B), we vary the number of parallel heads h with a fixed output dimensionality d , keeping the computational cost constant. The results suggest that $h = 8$ is the best choice for MUAN. Too few or too many heads reduces the quality of learned attention. In row (C), we fix the number of heads to $h = 8$ and vary the dimensionality d , resulting in much smaller and larger models with the model

TABLE II: Accuracies (%) of the *single-model* on the test-dev and test-std splits of VQA-v2 to compare with the state-of-the-art methods. All models use the same bottom-up attention visual features [50] and are trained on the train+val+vg splits, where vg indicates the augmented training samples from Visual Genome [56].

Method	Test-dev				Test-std
	All	Y/N	Num	Other	
Bottom-Up [51]	65.32	81.82	44.21	56.05	65.67
Counter [60]	68.09	83.14	51.62	58.97	68.41
MFH+CoAtt [35]	68.76	84.27	49.56	59.89	-
BAN [19]	69.52	85.31	50.93	60.26	-
BAN+Counter [19]	70.04	85.42	54.04	60.52	70.35
DFAF [44]	70.22	86.09	53.32	60.49	70.34
MCAN [61]	70.63	86.82	53.26	60.72	70.90
MUAN (ours)	70.82	86.77	54.40	60.89	71.10

complexity proportional to $O(d^2)$. From the results, we can see that d is a key hyper-parameter to the performance. Too small d may restrict the model capacity, leading to inferior performance. The model with $d = 1024$ slightly surpasses the reference model at the expense of much higher computational complexity and greater risk of over-fitting.

The hyper-parameters in the reference model is a trade-off between efficiency and efficacy. Therefore, we adopt the reference MUAN-10 model (abbreviated to MUAN for simplicity) in all the following experiments.

D. Results on VQA-v2

Taking the ablation studies into account, we compare our best MUAN model to the state-of-the-art methods on VQA-v2 in Table II. With the same bottom-up-attention visual features [50], MUAN significantly outperforms current state-of-the-art methods BAN [19] by 1.3 points in terms of overall accuracy on the test-dev split. Furthermore, for the Num-type questions, which verify object counting performance, BAN+Counter [19] reports the best result by utilizing an elaborate object counting module [60]. In contrast, MUAN achieves slightly higher accuracy than BAN+Counter, and in doing so does not use the auxiliary bounding-box coordinates of each object [60]. This suggests that MUAN can perform accurate object counting based on the visual features alone. As far as we know, MUAN is the first single model that achieves 71%+ accuracy on the test-std split with the standard bottom-up-attention visual features provided by [50].

E. Results on CLEVR

We also conduct experiments to compare MUAN with existing state-of-the-art approaches, and human performance on CLEVR, which is a synthesized dataset for evaluating compositional visual reasoning. Compared to VQA-v2, CLEVR requires a model not only to focus on query-specific objects, but only to reason the relations among the related objects, which is much more challenging. In the meantime, since the image contents are completely synthesized by the algorithm, it is possible for a model to fully understand the semantic,

TABLE III: Overall accuracies (%) on the test split of CLEVR to compare with the state-of-the-art methods. (*) denotes use of extra program labels. (\dagger) denotes use of data augmentation.

Method	Human [26]	Q-type Prior [26]	LSTM [26]	CNN+LSTM [26]	N2NNM* [62]	RN \dagger [63]	PG+EE* [64]	FiLM [65]	MAC [57]	MUAN (ours)
Accuracy	92.6	41.8	46.8	52.3	83.7	95.5	96.9	97.7	98.9	98.7

TABLE IV: Accuracies (%) on RefCOCO, RefCOCO+ and RefCOCOg to compare with the state-of-the-art methods. All methods use the detected proposals rather than the ground-truth bounding-boxes. COCO [53] and Genome [56] denote two datasets for training the proposal detectors. SSD [66], FRCN [55] and MRCN [58] denote the used detection models with VGG-16 [67] or ResNet-101 [38] backbones.

Method	Proposal Generator			RefCOCO			RefCOCO+			RefCOCOg	
	Dataset	Detector	Backbone	TestA	TestB	Val	TestA	TestB	Val	Test	Val
Attr [68]	COCO	FRCN	VGG-16	72.0	57.3	-	58.0	46.2	-	-	-
CMN [69]	COCO	FRCN	VGG-16	71.0	65.8	-	54.3	47.8	-	-	-
VC [70]	COCO	FRCN	VGG-16	73.3	67.4	-	58.4	53.2	-	-	-
Spe.+Lis.+Rein.+MMI [36]	COCO	SSD	VGG-16	73.7	65.0	69.5	60.7	48.8	55.7	59.6	60.2
Spe.+Lis.+Rein.+MMI [36]	COCO	SSD	VGG-16	73.1	64.9	69.0	60.0	49.6	54.9	59.2	59.3
DDPN [39]	Genome	FRCN	VGG-16	76.9	67.5	73.4	67.0	50.2	60.1	-	-
DDPN [39]	Genome	FRCN	ResNet-101	80.1	72.4	76.8	70.5	54.1	64.8	67.0	66.7
MAttNet [40]	COCO	FRCN	ResNet-101	80.4	69.3	76.4	70.3	56.0	64.9	67.0	66.7
MAttNet [40]	COCO	MRCN	ResNet-101	81.1	70.0	76.7	71.6	56.0	65.3	67.3	66.6
MUAN (ours)	COCO	MRCN	ResNet-101	82.8	78.6	81.4	70.5	62.9	68.9	71.5	71.0
MUAN (ours)	Genome	FRCN	ResNet-101	86.5	78.7	82.8	79.5	64.3	73.2	74.3	74.2

resulting in relatively higher performance of existing state-of-the-arts compared to those on VQA-v2.

From the results shown in Table III, we can see that MUAN is at least comparable to the state-of-the-art, even if the model is not specifically designed for this dataset. While some prior approaches used extra supervisory program labels [64][62] or augmented dataset [63] to guide training, MUAN is able to learn to infer the correct answers directly from the image and question features.

F. Results on RefCOCO, RefCOCO+, and RefCOCOg

We report the comparative results on RefCOCO, RefCOCO+, and RefCOCOg in Table IV. We use the common evaluation criterion accuracy, which is defined as the percentage of predicted bounding box overlaps with the groundtruth of $IoU > 0.5$. From the results, we can see that: 1) with the standard proposal features extracted from the detector pre-trained on MSCOCO, MUAN reports a remarkable improvement over MAttNet, the state-of-the-art visual grounding model; 2) with the powerful proposal features extracted from the detector pre-trained on Visual Genome, MUAN reports $\sim 9\%$ improvement over a strong baseline DDPN [39], which uses the same visual features. These results reveal the fact that MUAN outperforms existing state-of-the-arts steadily regardless of the used proposal features. Compared with existing approaches, MUAN additionally models the intra-modal interactions within each modality, which provide contextual information to facilitate visual grounding performance.

G. Qualitative Analysis

In Fig. 6, we show one VQA example and visualize four attention maps (obtained by Eq.(5)) from the 1st, 3rd, 6th and 9th UA blocks, respectively. Since only the feature of

the [ans] token is used to predict the answer, we focus on its related attention weights (*i.e.*, the first row of each attention map). In the 1st attention map, the word ‘many’ obtains the largest weight while the other words and visual objects are almost abandoned. This suggests that the 1st block acts as a question-type classifier. In the 3rd attention map, the word ‘street’ is highlighted, which is a contextual word to understand the question. The key word ‘buses’ is highlighted in the 6th attention map, and the two buses (*i.e.*, the 22th and 31th objects) are highlighted in the 9th attention map. This visual reasoning process explains the information of the highlighted words and objects is gradually *aggregated* into the [ans] feature. For the 9th UA block, we split its attention map into four parts (*i.e.*, ATT , AVT , ATV and AVV). In ATT , the largest values reflect the relationships between the key word and its context, providing a structured and fine-grained understanding of the question semantics (*i.e.*, bus is on the street). In ATV , some words on the rows attend to the key objects, suggesting that these words aggregate the information from the key objects to improve their representations. Similar observations can be observed from AVV and AVT .

In Fig. 7, we demonstrate one visual grounding example and visualize the prediction and the learned unified attention. In the first image, we can see that MUAN accurately localize the most relevant object proposal, and then output the refined bounding boxes as the final prediction. We visualize the learned textual and visual attentions of the 1st, 3rd, 6th and 9th UA blocks, respectively. By performing columnwise max-pooling over the unified attention map, we obtain the attention weights for the words and objects. For better visualization effect, we only visualize three representative objects with the largest attention weights. From the results, we can see that: 1) the keywords are highlighted only in the 1st block, indicating

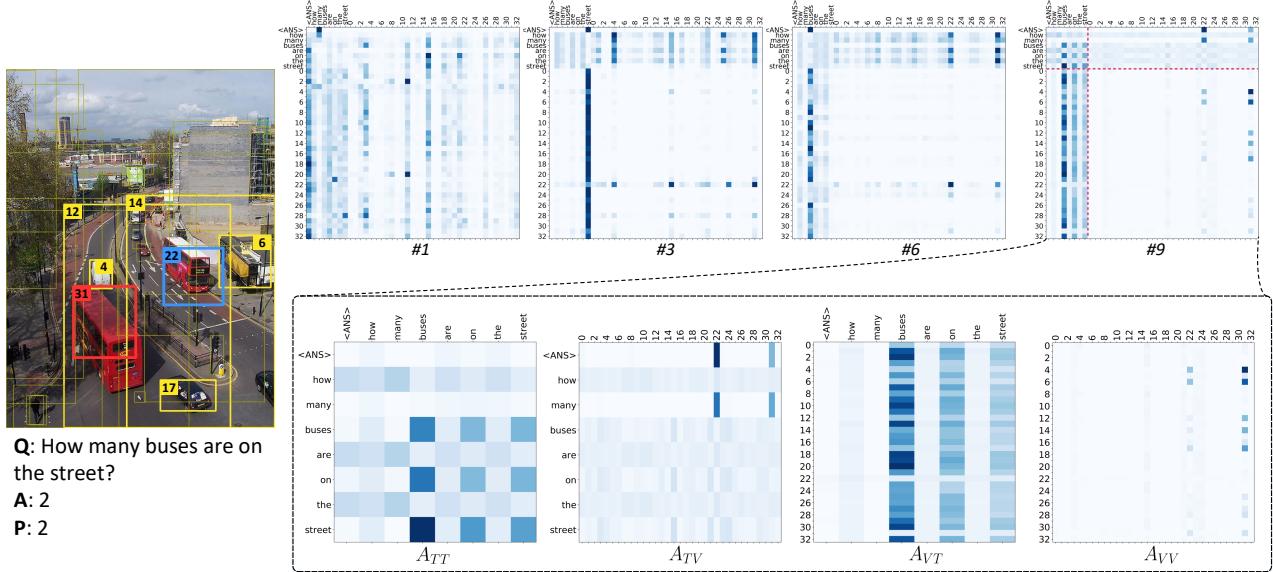


Fig. 6: Visualizations of the learned unified attention maps (Eq.(5)) for VQA. The attention maps come from the 1st, 3rd, 6th and 9th UA block, respectively. The index within [0-32] on the axes of the attention maps corresponds to the object in the image (33 objects in total). For better visualization effect, we highlight the objects in the image that are related to the answer. Furthermore, we split the last attention map into four parts (*i.e.*, A_{TT} , A_{VT} , A_{TV} and A_{VV}) to carry out detailed analysis.



Fig. 7: Visualizations of the prediction and the learned visual attention for visual grounding. The groundtruth (red), top-ranked proposal (blue) and refined prediction (yellow) are shown in the first image. Next four images illustrate the learned visual attentions from the 1st, 3rd, 6th and 9th UA blocks, respectively. The visual attention is represented by three representative objects with the largest attention values. The brightness of objects and darkness of words represent their importance in the attention weights.

that this information has been successfully transferred to the attended visual features in the following blocks; and 2) the learned visual attention in the 1st block is meaningless. After receiving the textual information, the visual attention tends to focus on the contextual objects in the 3rd and 6th blocks (*i.e.*, the hat and the baby), and finally focuses on the correct target object (*i.e.*, the woman) in the 9th block.

VI. CONCLUSION AND FUTURE WORK

In this work, we present a novel unified attention model that captures intra- and inter-modal interactions simultaneously for multimodal data. By stacking such unified attention blocks in depth, we obtain a Multimodal Unified Attention Network (MUAN), that is suitable for both VQA and visual grounding tasks. Our approach is simple and highly effective. We verify the effectiveness of MUAN on five datasets, and the experimental results show that our approach achieves top level performance on all the benchmarks without using any dataset specific model tuning.

Since MUAN is a general framework that can be applied to many multimodal learning tasks, there remains significant room for improvement, for example by introducing multitask learning with sharing the same backbone model or introducing weakly-supervised model pre-training with large-scale multimodal data in the wild.

REFERENCES

- [1] C. Zheng, L. Pan, and P. Wu, “Multimodal deep network embedding with integrated structure and attribute information,” *IEEE transactions on neural networks and learning systems*, 2019.
- [2] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, “Discriminative coupled dictionary hashing for fast cross-media retrieval,” in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2014, pp. 395–404.
- [3] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, “Shared predictive cross-modal deep quantization,” *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5292–5303, 2018.
- [4] 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 *International Conference on Machine Learning (ICML)*, vol. 14, 2015, pp. 77–81.

- [5] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: multi-modal stochastic rnns for video captioning," *IEEE transactions on neural networks and learning systems*, 2018.
- [6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.
- [7] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 817–834.
- [8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [9] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *NIPS*, 2014, pp. 2204–2212.
- [10] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *International Conference on Machine Learning (ICML)*, 2015, pp. 1462–1471.
- [11] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3640–3649.
- [12] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.
- [13] T. Dozat and C. D. Manning, "Deep biaffine attention for neural dependency parsing," in *International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: <https://nlp.stanford.edu/pubs/dozat2017deep.pdf>
- [14] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.
- [15] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 21–29.
- [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," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [17] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NIPS*, 2016, pp. 289–297.
- [18] 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)*, pp. 1839–1848, 2017.
- [19] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *NIPS*, 2018.
- [20] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.
- [22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [23] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.
- [24] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [25] 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," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [26] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence 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, pp. 2901–2910.
- [27] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 787–798.
- [28] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *IEEE International Conference on Computer Vision (ICCV)*, 2016, pp. 11–20.
- [29] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.
- [30] 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 *International Conference on Learning Representation (ICLR)*, 2017.
- [31] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multi-modal tucker fusion for visual question answering," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [32] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abcnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.
- [33] I. Ilievski, S. Yan, and J. Feng, "A focused dynamic attention model for visual question answering," *arXiv preprint arXiv:1604.01485*, 2016.
- [34] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4613–4621.
- [35] 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.
- [36] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speaker-listener-reinforcer model for referring expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7282–7290.
- [37] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 391–405.
- [38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [39] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, and D. Tao, "Rethinking diversified and discriminative proposal generation for visual grounding," *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.
- [40] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.
- [41] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4252–4261.
- [42] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, "Visual grounding via accumulated attention," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7746–7755.
- [43] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *AAAI*, 2019.
- [44] G. Peng, H. Li, H. You, Z. Jiang, P. Lu, S. Hoi, and X. Wang, "Dynamic fusion with intra-and inter-modality attention flow for visual question answering," *arXiv preprint arXiv:1812.05252*, 2018.
- [45] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," URL <https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language-understanding-paper.pdf>, 2018.
- [46] Y. Li, N. Wang, J. Liu, and X. Hou, "Factorized bilinear models for image recognition," *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [48] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 14, 2014, pp. 1532–1543.
- [49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [50] 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," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [51] D. Teney, P. Anderson, X. He, and A. v. d. Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," *arXiv preprint arXiv:1708.02711*, 2017.

- [52] R. Girshick, “Fast r-cnn,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
- [53] 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 *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.
- [54] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [55] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015, pp. 91–99.
- [56] 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,” *arXiv preprint arXiv:1602.07332*, 2016.
- [57] D. A. Hudson and C. D. Manning, “Compositional attention networks for machine reasoning,” *arXiv preprint arXiv:1803.03067*, 2018.
- [58] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2961–2969.
- [59] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 69–85.
- [60] Y. Zhang, J. Hare, and A. Prügel-Bennett, “Learning to count objects in natural images for visual question answering,” *International Conference on Learning Representation (ICLR)*, 2018.
- [61] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6281–6290.
- [62] 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.
- [63] 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*, 2017, pp. 4967–4976.
- [64] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, “Inferring and executing programs for visual reasoning,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2989–2998.
- [65] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [66] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.
- [67] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [68] J. Liu, L. Wang, and M.-H. Yang, “Referring expression generation and comprehension via attributes,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4856–4864.
- [69] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, “Modeling relationships in referential expressions with compositional modular networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1115–1124.
- [70] H. Zhang, Y. Niu, and S.-F. Chang, “Grounding referring expressions in images by variational context,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.