Question-Led Semantic Structure Enhanced Attentions for VQA

Anonymous ACL submission

Abstract

The exploit of the semantic structure in the visual question answering (VQA) task is a trending topic where researchers are interested in leveraging internal semantics and bringing in external knowledge to tackle more complex questions. The prevailing approaches either encode the external knowledge separately from the local context, which magnificently increases the complexity of the ensemble system, or use graph neural networks to model the semantic structure in the context, which suffers from the limited reasoning capability due to the relatively shallow network. In this work, we propose a question-led structure extraction scheme using external knowledge and explore multiple training methods, including direct attention supervision, SGHMC-EM Bayesian multitask learning, and masking strategies, to aggregate the structural knowledge into deep models without changing the architectures. We conduct extensive experiments on two domain-specific but challenging subtasks of VrR-VG dataset and demonstrate that our proposed methods achieve significant improvements over strong baselines, showing the promising potentials of applicability.

1 Introduction

In recent years, visual question answering (VQA) attracts an increasing attention benefiting from the great success of the neural networks. Having made the remarkable achievements on the early benchmarks like (Agrawal et al., 2016; Lin et al., 2015; Johnson et al., 2016), researchers are now interested in more challenging tasks such as (Zellers et al., 2019; Hudson and Manning, 2019) where the external knowledge and the commonsense are additionally required in order to provide the correct answer to the question about an image. For example in Fig. 1, to correctly answer what materials are used, a system needs to be aware of the spatial relationships among individual objects and find the exact wall "behind the red flowers". It leads to a higher requirement for a neural model to make use of the structural information in inference.

Researchers have proposed to use graph neural networks (GNNs) to incorporate the visual context and the external knowledge. (Xu et al., 2017; Li et al., 2017b; Zellers et al., 2018) generate a scene graph to represent the visual context, where the nodes are the objects and the edges are either events or the attributes. The graph-based models are naturally friendly to the external knowledge from the large-scale knowledge graphs (Speer et al., 2017; Bollacker et al., 2008) because both share the same graph-structured format. However, due to the over-smoothing issue during the training process (Oono and Suzuki, 2019; Chen et al., 2020a), GNNs do not generally allow to build up layers or scale up in depth, hindering their reasoning capabilities to grow, which explains the fact that the state-of-the-art performances on the benchmarks are dominated by the conventional non-graph deep models (Li et al., 2019c,a; Chen et al., 2020b; Li et al., 2020b). In parallel, some other works (Li et al., 2017a; Su et al., 2018; Li et al., 2020a) make effort to encode the external knowledge separately and fuse with the local context through an additional memory network or graph network, whereas they either lose the semantic structure or add too much complexity to the overall system.

Q: What substance is the wall behind the red flowers made from ?

Figure 1: An Example of a complex question requiring the structure semantics and external knowledge.

An enormous effort has been made towards modeling the dependencies among the contextual objects and words within the conventional neural network. The attention mechanism is one of the most influential ones, which was first introduced by (Cho et al., 2014) to machine translation tasks and encourage the emergence of many variants, including general attention (Luong et al., 2015), the dot-product attention (Luong et al., 2015), the scaled dot-product attention (Vaswani et al., 2017), etc. Now the attention has been an indispensable part of the latest models for VQA tasks. We notice that the attention operations and the graph operations share a lot in common, and in particular the nature of the self-attention can be viewed as a fully-connected graph. Thus the attention layer can be potentially used to model the structure information. Meanwhile, despite the consistent gain brought by the attention mechanism, the attention weights are mostly learned in an unsupervised scheme and a prominent benefit is expected from further optimization. In this work, we focus on the scaled dot-product attention which is the core of the transformer block (Vaswani et al., 2017) and being widely used in the state-of-the-art models.

To this end, a question-initiated semantic structure extraction method is designed, following human thinking process, and aggregated into the attention layer in the transformer block through weak supervision and masking. The extracted semantic structure is further enhanced by the scene graph and the external knowledge such as word synonyms and object relevancy. Then we explore three novel strategies to improve the attention learning: (1) **indirectly** optimize the attention weights in the multi-task learning framework with Bayesian inference, by adding an auxiliary task of the attention object prediction to the model; (2) **directly** supervise the scaled dot-product attention weights with the enriched structural semantics in an explicit way; (3) selectively **mask out** the attention weights of the irrelevant objects during training based on the semantic structure. Compared to other works (Li et al., 2017a; Su et al., 2018; Kim et al., 2020; Huang et al., 2020; Zhu et al., 2020), our efforts do not change the backbone models. Our main contributions can be summarized as follows:

- We propose the direct and indirect attention supervision methods for the VQA task that are applicable to a more general situation.
- We introduce a debiased multitask training method with Bayesian inference to the VQA task for the first time that increases the model's stability.
- We propose a question-led semantic structure extraction schema, simulating human behaviors and boosting the model's interpretability.
- We apply our approaches to multiple state-of-the-art transformer-based models and show compelling results on two challenging subtasks of the VrR-VG dataset, demonstrating the encouraging potentials for future use.

2 Related Work

VQA Models With External Knowledge The traditional approaches to incorporate the external knowledge into the local context can be summarized into two categories. The works like (Li et al., 2017a; Su et al., 2018) first encode the local context and the external knowledge separately in their own representation space, and then perform a late fusion by projecting two spaces into a common hidden space through an additional neural network. The network usually contains a large number of parameters for decent performance and therefore bring more complexities to the base models.

More recently, researchers leverage GNNs to model the structure within the context. (Li et al., 2019b) uses a graph attention network(GAT) to enocde the semantic, spatial and implicit relations among the visual objects. (Kim et al., 2020) constructs two symbolic graphs to separately encode the questions with dependency-tree structure and the objects with attribute-and-predicate-based structure. Similarly, (Huang et al., 2020; Zhu et al., 2020) proposes multiple independent graph convolutional networks(GCNs) to capture embed intra- and cross-modal relations using external knowledge. (Singh et al., 2019; Li et al., 2020a) perform an early fusion to merge the local context and the external knowledge into an entity graph, and leverage a graph neural network (GNN) to conduct encoding and reasoning. The shortcomings of this category mainly lie on the comparably weak reasoning capability of GNN restricted by its relatively shallow depth. Most recently, we have observed the boom of the Vision-and-Language(V+L) multimodal large-scale pre-trained models (Su et al., 2020; Li et al., 2019c,a; Chen et al., 2020b; Li et al., 2020b) and their great success in the VQA tasks.

In this work, we choose three state-of-the-art models as our competitive baselines. Since they are all transformer-based, the methods should be seamlessly applied to other Transformer-based models.

Attention Supervision & Masking (Liu et al., 2016; Gan et al., 2017; Qiao et al., 2017) achieve the attention supervision for VQA task by explicitly generating an attention map as an additional output of the model and optimizing it under the multitask learning framework. Inspired by the idea, we make adjustment for transformer-based models to predict the indices of the attention objects in the input, which is considered to be easier to learn because of its comparably smaller parameter space. (Patro et al., 2019) directly regularizes the attention weights in the model using the gradient information from Grad-CAM (Selvaraju et al., 2017) as the supervision signals at each training step. The attention is optimized iteratively under the adversarial learning framework. However, Grad-CAM can be only applied to a CNN-based model and makes it inapplicable to the latest state-of-the-art models. Therefore, we propose a direct weight supervision strategy for transformer-based models.

The masking is applied to incorporate the structural information into the transformer block for various tasks in the latest works. (Ahmad et al., 2020) uses the word distance as the reference to form the mask matrix to reflect word relations in a sentence for event extraction task; (Guo et al., 2020) proposes a new pre-trained model for programming language which uses the masked attention to represent the dependency among the programming variables; (Shao et al., 2020) use the masked attention to encode the parsing-tree-based structure into a sentence representation so that each work token only interact with its corresponding parents and not with nodes in different sub-trees. Considering the gain from the masking technique, we include it as one of our strategies.

3 Attention Enhancement Strategies

In this section, we provide the theories and the details of our **indirect**, **direct** and **masking** strategies, as shown in Fig. 2.

The original task of VQA is to predict the answer $\mathbf{A}$ given the question $\mathbf{Q}$ and the visual context $\mathbf{C}$. In this work, $\mathbf{C}$ is a list of object and whole image representations. Letting θ be the parameters of the base model and D_θ be the training samples for the original task, we learn θ by maximizing its log-likelihood as follows with a binary cross-entropy loss following the settings in (Yu et al., 2019).

$$\theta^*_{\text{orig}} = \arg\max_\theta \log p(D_\theta|\theta) \quad (1)$$

3.1 Indirect Strategy: Multitask Learning With Bayesian Inference

We add an auxiliary task of predicting the expected attention object(s) in the input with respect to each $(\mathbf{Q}, \mathbf{C})$ pair, and formulate it as a multi-label multi-class classification problem. Assuming the newly-added auxiliary-task-specific parameters is ϕ, we arrive at the objective function for our multitask learning:

$$\theta^* = \arg\max_\theta [\log p(D_\theta|\theta) + \log p(D_\phi|\theta, \phi)] \quad (2)$$

where D_ϕ is the training data for the auxiliary task.

Normally, multitask learning optimizes the parameters for the best overall performances of all downstream tasks on the cost of the performance drop on the individual task. However, in our indirect strategy, only the main task matters[1]. From this perspective, we take ϕ in Eq. 2 as a bias to the estimation of θ^*. According to Bayes' theorem, $p(\phi|\theta)$ is needed to remove ϕ from Eq. 2 which, however, is either unknown or require some strong assumptions on $p(\phi|\theta)$.

To diminish the bias and maximize the benefits of the auxiliary task to the original task, we claim that a better objective function is as follows:

$$\theta^* = \arg\max_\theta [\log p(D_\theta|\theta) + \log p(D_\phi|\theta, \widetilde{D_\phi})] \quad (3)$$

where $\widetilde{D_\phi}$ is another set of data used to estimate the posterior distribution $p(\phi|\theta, \widetilde{D_\phi})$. The underlying motivation is that, instead of making a strong assumption on the prior distribution of ϕ and relying on its sensitive initialization procedure, we introduce $\widetilde{D_\phi}$ and estimate the posterior distribution of ϕ from data.

To estimate $p(D_\phi|\theta, \widetilde{D_\phi})$ in Eq. 3, we apply Bayesian inference to the optimization procedure following

$$\begin{aligned} &\log p(D_\phi|\theta, \widetilde{D_\phi}) \\ &= \log \int_\phi p(D_\phi|\theta, \phi, \widetilde{D_\phi}) p(\phi|\theta, \widetilde{D_\phi})\, d\phi \quad (4) \\ &\geq \int_\phi p(\phi|\theta, \widetilde{D_\phi}) \log p(D_\phi|\theta, \phi)\, d\phi \quad (5) \\ &= E_{p(\phi|\theta, \widetilde{D_\phi})}[\log p(D_\phi|\theta, \phi)] \end{aligned}$$

[1]The auxiliary task is a proxy task to guide the attention weight during training, but is not considered in inference phase. This is what **indirect** means.

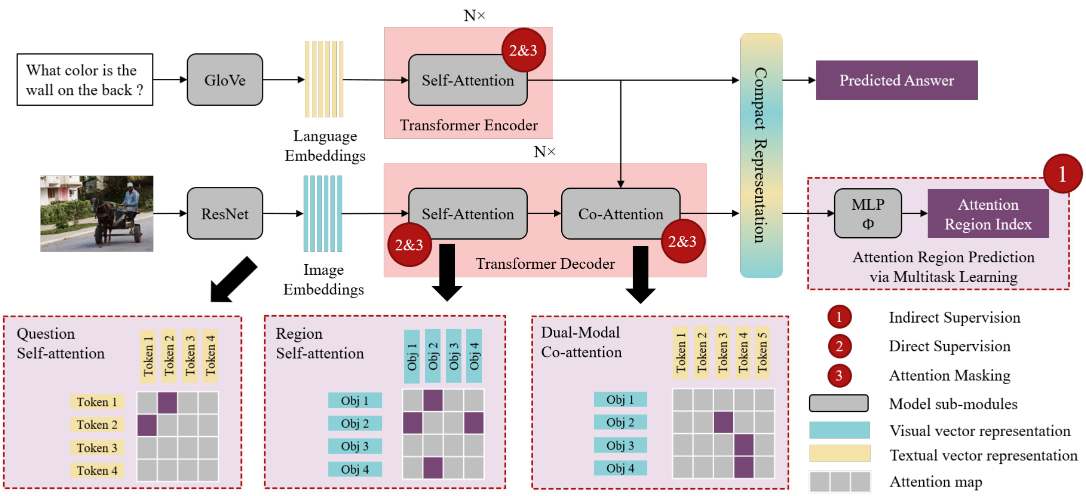

Figure 2: Illustration of three attention enhancing methods on MCAN. The boxes with red dashed line visualize the details of different attention structures. The grey cells in the attention map represent no link; the purple cells represent a valid link.

$$\approx \frac{1}{K}\sum_{k=1}^{K} \log p(D_\phi|\theta,\phi_k), \quad \phi_k \sim p(\phi|\theta,\widetilde{D_\phi}) \tag{6}$$

where Eq. 5 is drawn from Jensen's inequality and the independence between D_ϕ and $\widetilde{D_\phi}$ given ϕ. Eq. 4 is considered computationally intractable due to the huge parameter space for ϕ, but can be estimated by Monte Carlo sampling as shown in Eq. 6. In this work, we use stochastic gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014) to achieve sampling from $p(\phi|\theta,\widetilde{D_\phi})$. Specifically, we freeze θ during sampling and only update ϕ through maximizing $p(\widetilde{D_\phi}|\theta,\phi)$, following Eq. 7:

$$\phi^{i+1} = \phi^i + \mathcal{L}_{a1}\nabla\phi \tag{7}$$

$$\begin{aligned}\mathcal{L}_{a1}(\mathbf{x}) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log p(y_n|\theta,\phi^i,x)\\ + (1-y_n)\log(1-p(y_n|\theta,\phi^i,x))]\end{aligned} \tag{8}$$

where $\mathcal{L}_{a1}$ is the binary cross-entropy loss of the auxiliary task; N is the number of objects at input and output; $\{x, y_{1:N}\} \in \widetilde{D_\phi}$. The burn-in procedure of the sampling first ensures ϕ to be converged to a local optimal region and the actual samples $\{\phi_k\}$ are collected from the local optimal region.

Substituting Eq. 6 into Eq. 3 reaches Eq. 9, whose optimal solution can be approached by Expectation-Maximization (EM) algorithm.

$$\begin{aligned}\theta^* &= \arg\max_\theta \{\log p(D_\theta|\theta) + \log p(D_\phi|\theta,\widetilde{D_\phi})\}\\ &\approx \arg\max_\theta \{\log p(D_\theta|\theta) + \frac{1}{K}\sum_{k=1}^{K}\log p(D_\phi|\theta,\phi_k)\}\end{aligned} \tag{9}$$

We learn θ by iteratively sampling $\{\phi_k^t|\phi_k^t \sim p(\phi|\theta^t,\widetilde{D_\phi})\}$ at time t to estimate $E[\log p(D_\phi|\theta^t,\phi^t)]$ following Eq. 6, and then updating θ to be θ^{t+1} following Eq. 9 with the objective function as Eq. 10:

$$\min_\theta \ \mathcal{L}_{\text{orig}}(\theta : D_\theta) + \alpha\mathcal{L}_{a1}(\theta,\phi : D_\phi,\widetilde{D_\phi}) \tag{10}$$

Considering the efficiency of the algorithm in practice, we do not sample $\{\phi_k\}$ every training step for θ, because the burn-in process of SGHMC sampling takes time. Instead, we sample $\{\phi_k\}$ for every L steps.

3.2 Direct Strategy: Transformer Supervision

Scaled Dot-product Attention By definition in (Vaswani et al., 2017), the scaled dot-product attention can be expressed as

$$\begin{aligned}Q_V = A(Q,K,V) =& \text{softmax}(\frac{QK^\top}{\sqrt{d}})V\\ =& \text{softmax}(W_a)V\end{aligned} \tag{11}$$

where $A(\cdot)$ is the scaled dot-product attention function; $Q \in \mathbb{R}^{n_q\times d}$, $K \in \mathbb{R}^{n_k\times d}$ and $V \in \mathbb{R}^{n_v\times d}$ ($n_v = n_k$) are the matrices that contain n_q query vectors, n_k key vectors and n_k value vectors; W_a can be regarded as the unnormalized affinity matrix with W_a^{ij} representing the affinity score between the i-th query and the j-th key. Assuming $K = V$, normalizing the affinity matrix along the row axis with a SoftMax function makes the output a co-attention map. As a result, $Q_V \in \mathbb{R}^{n_q\times d}$ becomes an attended query matrix containing n_q attended query vectors with respect to n_k key vectors or n_v value vectors since $V = K$. For simplicity, we call $A(Q,K,K)$ co-attention module for Q and K

in the rest of the paper and call $A(Q, Q, Q)$ self-attention module for Q.

We add the regularization directly to the affinity matrix W_a and formulate it as a multi-label multi-class classification problem of predicting whether the row vectors in Q and K are pair-wise associated, leading to the final loss function of our direct attention supervision strategy as follows:

$$\min_{\theta} \mathcal{L}_{\text{orig}}(\theta : D_\theta) + \beta \mathcal{L}_{a2}(s(W_a), W_a^{\text{gt}}) \quad (12)$$

where $s(\cdot)$ is a sigmoid function; W_a^{gt} is the ground-truth attention map based on the structural semantics that will be discussed in the next section; $\mathcal{L}_{a2}$ is mean squared error(MSE) loss; β is a weight to be tuned. Note that KL divergence between $s(W_a)$ and W_a^{gt} also has been tried for $\mathcal{L}_a$, but only provides weaker results.

3.3 Masking Strategy: Masked Attention

Following (Ahmad et al., 2020; Guo et al., 2020; Shao et al., 2020), we give the definition of the masked attention as in Eq. 13 on the basis of Eq. 11:

$$Q_V = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right)V \quad (13)$$

where $M^{ij} = 0$ if the i-th query and the j-th key are linked and $M^{ij} = -\infty$ if the i-th query and the j-th key are considered irrelevant. The top-down semantic structure maneuvers the back-propagation procedure through the mask M to only optimize the weights where there are valid interactions between two nodes.

4 Semantic Structure Extraction

In this section, we discuss how we form our question-led intra- and inter-structures for attention supervision.

We divide the structural semantics in VQA into three types: word-to-word(W2W), region-to-region(R2R) and word-to-region(W2R). Different from the independent generic intra-modality structures in (Li et al., 2019b; Kim et al., 2020; Teney et al., 2017), we look for question-led semantics with the help of the external knowledge from the language models in Spacy (Honnibal and Montani, 2017) and the commonsense from ConceptNet (Speer et al., 2017), following the human behavior in answering a VQA question. Our goal is to impose the structural semantics into the attention modules.

4.1 Question-led Semantic Structure

W2W We first detect the keywords in the questions based on the dependency and constituency parsing results, including the noun words in the noun phrases and their corresponding adjectival modifiers. Then we build a fully connected subgraph among all the keywords and generate an adjacent matrix W_{at}^{gt} for the question self-attention.

W2R Led by the question keywords, we search for essential regions in the image according to the conceptual relations. Assuming the availability of the object names or the caption of each candidate region[2], each value in W2R matrix is determined by the pair-wise affinity function $f(W, R)$ between a question keyword and a region description. In the cases where the keyword or the description consists of multiple words, we use the maximum word-level score for the whole phrase. The affinity score is measured from four perspectives, including string matching, the Euclidean distance in word vector space[3], the relevancy score supported by ConceptNet, a customized mapping function. The score from each perspective is normalized to the scale of 1 and the maximum score in the four perspectives will be taken as the final word-wise affinity. Thresholding method is then adopted to generate the final W2R adjacent matrix W_{av}^{gt} i.e.

$$W_{av}^{\text{gt}}(i,j) = \begin{cases} 1, & f(W_i, R_j) \geq \sigma \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

where σ is a hyperparameter; W_i is the i-th word in the question; R_j is the j-th region in the image. More details are included in Append. A.

R2R Similar to W2R, we build the R2R adjacent matrix W_{ax}^{gt} for region self-attention based on their conceptual relations from the four perspectives. To keep W_{ax}^{gt} question-led, we only consider the candidate relations centered on the essential regions detected in W2R. Different from (Teney et al., 2017; Li et al., 2019b; Huang et al., 2020; Zhu et al., 2020), we do not consider spatial relation. Based on the results in (Yang et al., 2019) and our empirical observations, we claim the automatically extracted 2D/3D spatial relations are too noisy to be valuable attention groundings.

In our work, W_{at}^{gt}, W_{av}^{gt} and W_{ax}^{gt} are used as the attention groundings in **direct supervision** and as the mask in **attention augmentation**.

[2] The object name or region caption can be either obtained by hand annotation or inferred by pre-trained models.

[3] We use the "en_core_web_md" model provided by Spacy.

4.2 Answer-Led Semantic Structure

In **indirect supervision**, a set of ground-truth attention regions are required for the multi-class multi-label auxiliary task. To guarantee the quality of the weak labels, we look for the essential regions that are best described or most related to the answers. Specifically, we follow the same procedure used for W2W structure of measuring the conceptual closeness between the answer words and the region descriptions, and selected the top ranked region(s) from all the candidates as the weak labels for the auxiliary task.

5 Experimental Setting

5.1 Dataset

We use Visual-relevance Relationships (VrR-VG) dataset (Liang et al., 2019) which is a subset of Visual Genome (VG) dataset (Krishna et al., 2017) for the experiments. According to (Liang et al., 2019), VrR-VG discards those highly predictable and biased question-answer pairs in VG dataset and therefore becomes a more challenging task. Moreover, the annotated scene graph for each image allows higher-quality labels for our attention supervision. We find that a large percentage of question-answer pairs in VrR-VG do not require the fine-grained relations in the context and are not expected to benefit from our extracted semantic structure. Examples are provided in Append. B. To better verify our proposed methods, we further distill two subsets from VrR-VG and have "What-Color" and "What-There" questions through simple string matching methods.

	What-Color	What-There
# of questions in train	50726	33736
# of questions in val	17234	11120
# of questions in test	17465	11398
# of answers (classes)	248	1049

Table 1: Statistics of two subsets.

5.2 Baselines & Metrics

We take three Transformer-based models, MCAN (Yu et al., 2019), MMnasNet (Yu et al., 2020) and LXMERT (Tan and Bansal, 2019), as our strong baselines, to demonstrate both the effectiveness and applicability of our proposed methods. We also run the experiments with MFB (Yu et al., 2017) under the baseline setting which is commonly compared in VQA task. All the models and methods are evaluated by the QA accuracy[4].

6 Evaluation

6.1 Annotation Results

	What-Color	What-There
# of question keywords	1.53	1.72
# of attention regions	1.16	1.20
# of relevant regions	5.26	5.02
MFB	45.53 (52.78)	23.69 (26.48)
MCAN	45.45 (53.43)	24.87 (29.13)
MMnasNet	45.35 (53.40)	24.21 (29.30)
LXMERT	46.55 (52.68)	26.22 (28.22)

Table 2: Validation of the extracted attentions in training set. The upper side shows the average amount of the items per question. The lower side compares the accuracy on the dev set with the baseline and attention-region-only inputs. The numbers in the brackets correspond to the latter.

It is difficult to directly examine the quality of our weak attention annotations. In this work, we evaluate it from two aspects. The upper side of Table 2 includes the average number[5] of attention words and objects for each question. Those close-to-1 mean values show that our question-led structure scheme is capable of finding the concrete and specific question keywords and the essential visual regions in both subsets. Additionally, we also conduct experiments where we train the baseline models using only our extracted attention regions as the visual input. The corresponding performance on the dev set is included in the brackets in the lower side of Table 2. The significant improvement over the baseline performance (outside the brackets) validates the quality and feasibility of our question-led semantic structure extraction scheme.

6.2 VQA Results

Effect of Training Methods The small difference among baseline results in Table 2 reveals the limited profit from model architectures on two challenging subsets. Comparably, Table 3 shows the arresting benefits from the integrated structure knowledge in most cases, especially with the masking techniques bringing a maximum of 19% rise in QA accuracy. Direct supervision on attention weights can also bring significant improvement over the baseline in the cases where the masking technique

[4]The implementation details are included in Append. C.

[5]average = # found / (# of question - # of empty annotation)

	MCAN			MMnasNet			LXMERT		
	Color	**There**	**Combined**	**Color**	**There**	**Combined**	**Color**	**There**	**Combined**
Baseline	45.45	24.87	33.19	45.35	24.21	32.78	46.55	26.22	34.72
+ Indirect	45.62	25.07	33.57	45.38	24.97	33.17	46.30	26.68	34.90
+ Direct	**48.30**	25.22	34.56	**48.94**	25.06	35.07	46.65	26.33	34.63
+ Masking	45.82	**28.27**	**38.44**	45.57	**28.85**	**39.19**	**53.04**	**29.67**	**38.37**

Table 3: The effect of our training methods. The numbers are the maximum accuracy (%) on the dev set out of multiple runs.

is not as effective. The finding may imply potential complementation between two methods and encourage users to try the other if one does not show promising results in future applications. Our indirect supervision strategy only provides a modest improvement, which matches our observations in the works (Qiao et al., 2017; Zhang et al., 2019) where only a tiny gain is earned from similar indirect supervision methods.

Another interesting finding is that both direct attention supervision and masking should not influence the learning of the dense representation for each modality. Different from MCAN and MMnasNet that use the pre-trained GloVe (Pennington et al., 2014) embeddings for the textual input, LXMERT uses Transformer blocks to learn the textual representations simultaneously before multimodal fusion. We experiment with LXMERT by adding the supervision and the masking to the Transformer blocks before and at the fusion module. A significant performance drop is observed if we supervise or mask the attention weights before the fusion module, which indicates that our semantic structures capture the high-level relations and should only be used to guide the attention learning in the deeper layers.

	Color	**There**	**Combined**
MMnasNet	45.35	24.21	32.78
+ R2R Direct	45.39	24.37	33.15
+ W2W Direct	45.35	24.75	33.16
+ W2R Direct	**48.94**	**25.06**	**35.07**
+ Full Direct	46.20	24.68	34.55
+ Masked R2R	45.51	24.53	33.02
+ Masked W2W	45.23	24.54	33.10
+ Masked W2R	45.42	27.17	36.14
+ Masked Full	**45.57**	**28.85**	**39.19**

Table 4: Accuracy(%) on the dev set with different semantic structures. Reported numbers are the maximum out of multiple runs.

Effect of Semantic Structures An ablation study is conducted on the different types of semantic structures in Sec. 4.1. Table 4 reveals that the inter-modal semantics(W2R) play a more important role than the inner-modal semantics(W2W, R2R) in all conditions with MMnasNet model. Similar results are also found with MCAN and LXMERT models, which conform to our intuition and can be instructive to the future effort on increasing the information exchange for VQA tasks.

Effect of Bayesian Inference Table 5 compares the performance of the indirect strategy on the test set with and without Bayesian inference. We repeat our experiments with the MCAN model and $K = 50$ for 10 times and compute the mean and standard deviation of the QA accuracy. We notice that our debiased multitask learning increases the model stability with a much smaller variation in test performance, which attributes to the fact that we use the posterior distribution of ϕ during the optimization rather than its likelihood, while still enjoying the performance growth from the indirect supervision.

	Color	**There**
MCAN	44.54 ± 0.23	23.92 ± 0.14
+ Multitask	44.80 ± 0.28	24.45 ± 0.30
+ Multitask + B.I.	44.70 ± 0.16	24.23 ± 0.16

Table 5: Accuracy(%) on the test set. "B.I" stands for Bayesian inference.

6.3 Supervised Attention Results

To validate the effect of the supervision in inference, we visualize the visualizes the attention weights of the last transformer layer in the encoder and decoder of MCAN. Fig. 3 is on a "What-There" question from the dev set[6]. We find that the indirect supervision does not make a significant difference to the attention weight against the baseline, which can partially explain its limited contribution to the accuracy.

Comparatively, the direct supervision guides the textual self-attention module to focus more on the structure-aware keywords in the question. For the example in Figure 3, the baseline focuses on "green"

[6] Visualization for a "What-Color" question is available in Append. D.

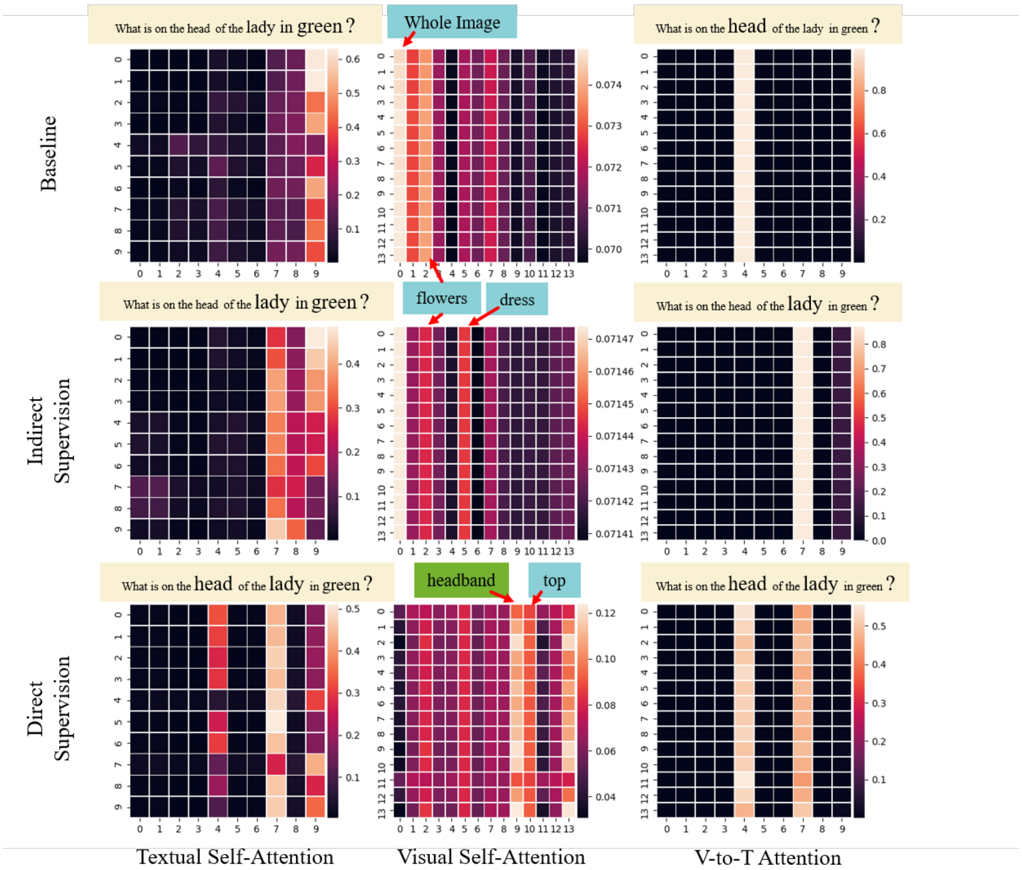

Figure 3: Visualization of the horizontally-normalized attention weights of the last transformer layer in the encoder and decoder of MCAN. For textual and visual self-attentions, they are normalized along the horizontal axis. "V-to-T" means Visual-to-Textual attention. The brighter the cell, the higher weight it carries. The larger font of the text in yellow box, the higher weight it carries. The green box is the true attention object, the blue boxes are candidate objects. The direct supervision is the model trained with full attention supervision; the indirect supervision is the model trained with SGHMC-EM multitask learning. The sparsity in text-related attention results from the small number of keyword annotations per question as it is shown in Table 2.

and the indirect-supervised model focuses on "lady in green". Only the direct-supervised model realizes the true keywords of "head", "lady" and "green", and consequently leads the visual self-attention module to find the attention object "headband". As a result of imposing the semantic structure into the attention, the direct supervision also helps the visual self-attention module concentrate more on individual objects rather than the global context[7], which explains the greater improvements on the end-goal performance from our direct supervision strategy.

What's more, our extracted semantic structure boost the interpretability of the model through attention weights. With our supervision methods, the textual self-attention finds candidate keywords in questions and the V-to-T attention further filters out those less related to the visual context. On the contrary, the baseline model leaves it unclear about why the textual attentions shift from "green" to "head".

7 Conclusion

In this work, we develop three strategies to enhance attention training with the question-led semantic structure without any changes to the backbone models. Both direct supervision and masking techniques lead to notable improvements with structural knowledge, but the magnitude may be subject to data and model. The debiased multitask learning is beneficial to increase a model's stability during inference. Our further ablation study reveals that the cross-modal semantics performs a more critical role in the VQA task. We value our work as a systematic study on boosting attention with the semantic structure for VQA tasks and may inspire future work towards this direction.

[7]The first item(column) in visual self-attention is the vector representation of the whole image.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2020. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. *arXiv preprint arXiv:2010.03009*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. 2020a. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3438–3445.

Tianqi Chen, Emily Fox, and Carlos Guestrin. 2014. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. 2017. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1811–1820.

Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Jian Yin, Daxin Jiang, et al. 2020. Graphcodebert: Pre-training code representations with data flow. *arXiv preprint arXiv:2009.08366*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7166–7176.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.

Eun-Sol Kim, Woo Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. 2020. Hypergraph attention networks for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14581–14590.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training.

Guohao Li, Hang Su, and Wenwu Zhu. 2017a. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. *arXiv preprint arXiv:1712.00733*.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020a. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1227–1235.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019b. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10313–10322.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019c. Visualbert: A simple and performant baseline for vision and language.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017b. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1261–1270.

Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the*

IEEE International Conference on Computer Vision, pages 10403–10412.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2016. Attention correctness in neural image captioning. *arXiv preprint arXiv:1605.09553*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Kenta Oono and Taiji Suzuki. 2019. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*.

Badri N Patro, Vinay P Namboodiri, et al. 2019. Explanation vs attention: A two-player game to obtain attention for vqa. *arXiv preprint arXiv:1911.08618*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2017. Exploring human-like attention supervision in visual question answering. *arXiv preprint arXiv:1709.06308*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Bo Shao, Yeyun Gong, Weizhen Qi, Guihong Cao, Jianshu Ji, and Xiaola Lin. 2020. Graph-based transformer with cross-candidate verification for semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8807–8814.

Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. 2019. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4602–4612.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. 2018. Learning visual knowledge memory networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7736–7745.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Damien Teney, Lingqiao Liu, and Anton van Den Hengel. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Computer Vision and Pattern Recognition (CVPR)*.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. 2019. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *International Conference on Computer Vision*.

Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. 2020. Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3743–3752.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6281–6290.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840.

Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2019. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 349–357. IEEE.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visualquestion answering. *arXiv preprint arXiv:2006.09073*.

A Conceptual Relation Measuring

The conceptual relations are used in creating W2R, R2R graphs and the attention region annotations. It is determined by the conceptual closeness between two words or phrases, which is estimated by the word-wise affinity. Given two phrases $P = [p_1, ..., p_m]$ and $Q = [q_1, ..., q_n]$ where p_m and q_n are the tokens in the phrases, the affinity S between P and Q are defined as

$$S_{PQ} = \max(f(p_1, q_1), f(p_1, q_2), ..., f(p_2, q_1), ..., f(p_m, q_n)) \tag{15}$$

where $f(\cdot, \cdot)$ is a word-wise affinity function. As aforementioned, the word-wise affinity is measured from the perspectives of string matching, word vector distance, ConceptNet relevancy score, a customized mapping function, i.e.

$$f(a, b) = \max(g_{str}(a, b), g_{word}(a, b), g_{net}(a, b), g_{map}(a, b)) \tag{16}$$

where $g_{str}(\cdot, \cdot)$ and $g_{map}(\cdot, \cdot)$ are binary functions returning 0 or 1 depending on whether two strings are matched and whether a certain mapping is defined. $g_{map}(\cdot, \cdot)$ is customized to handle some dataset-specific biases, e.g. in VrR-VG dataset, "computer" is much more frequently described as "CPU" than in reality. $g_{net}(a, b)$ represents the ConceptNet relevancy function that computes the relevancy score between a and b; we normalize the score as follows

$$g_{net}(a, b) = \min(1, \frac{\text{relevancy score}}{1.3}) \tag{17}$$

where 1.3 is our empirical scale value. $g_{word}(a, b)$ is defined in word vector space, i.e.

$$g_{word}(a, b) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{18}$$

where $\vec{a}$ and $\vec{b}$ represent a's and b's word vectors.

B Examples of VrR-VG QA Pairs

When examining the data samples, we find that the majority of QA pairs in VrR-VG dataset do not rely on the detailed semantic relations among the objects. Below are some samples whose answers are unlikely to be learned from the semantic structures.

- Q: "Where are the trees?" A: "Right."
- Q: "What time of day is it?" A: "Mid morning."
- Q: "How are the cars?" A: "Parked."
- Q: "How is the weather?" A: "Fair."

Comparatively, our extracted "what-color" and "what-there" subset are more representative of the important role that the semantic structures are playing. For example,

- Q: "what color is the building the L&M ad is on?" A: "Tan."
- Q: "What color is the car parked closest to the people?" A: "Silver."
- Q: "What is on the man's face wearing the red shirt?" A: "Glasses."
- Q: "What is behind the boy" A: "Buildings."

C Implementation Details

Dataset We use a pre-trained wide ResNet model[8] to create the vector representation for both image and regions in the image. We further clean up the data samples in the subsets, removing the images that contains over 60 regions and the questions whose answer frequency is lower than 5. We randomly split the resulting samples into train, dev and test sets with the ratio of $3:1:1$.

[8]Precisely, we use "wide_resnet101_2" model provided by Pytorch

Hyperparameters During annotation generation, we take the whole image as the attention if no attention object is detected; and to allow some degree of fuzziness in natural expressions, we set the thresholds for the Spacy similarity score and the ConceptNet "RelatedTo" score to be 0.7 and 1.3 respectively.

During training, We leverages the OpenVQA's[9] implementations of MFB, MCAN and MMnasNet models and the original implementation[10] of LXMERT for the baselines. Specifically, we choose the small version of MCAN and MMnasNet with fewer layers of transformer blocks. The maximum length of the questions token is 24. we use $K = 30$ in Eq. 9 and sample $\{\phi_k\}$ every 20 training steps (i.e. $L = 20$) when performing EM algorithm. Following (Zhang et al., 2019), we adopt dynamic weights for α and β in Eq. 10 and Eq. 12, i.e.

$$\alpha(e) = 20 * (1 + \cos \frac{e}{\mathrm{E}}) \tag{19}$$

$$\beta(e) = 100000 * (1 + \cos \frac{e}{\mathrm{E}}) \tag{20}$$

where e is the current epoch and E is the total number of training epochs. When constructing semantic graphs, we use $\sigma = 0.7$ in Eq. 14.

[9] https://github.com/MILVLG/openvqa

[10] https://github.com/airsplay/lxmert

D More Examples of Attention Inference

(a) For Fig. 3: "What is on the head of the lady in green?"

(b) For Fig. 5: "What color is the cat's eye?"

Figure 4: Original images for the attention visualization of "What-Color" and "What-There" questions.

Similar to the example of the "What-There" question, We find the model trained with the direct supervision constantly focus on more specific visual regions. Specifically, the visual self-attention weights from the baseline and the indirect supervision are almost equally distributed, while that from the direct supervision provides a more meaningful distribution with peaks and troughs. It illustrates the efficiency of our direct supervision. Contrarily, no difference as significant as in "What-There" questions is observed among the V-to-T attention results. A plausible explanation is that since the structure of the color question is simpler than the "What-There" question overall, the model tends to just simply pay attention to the most important keyword in the question.

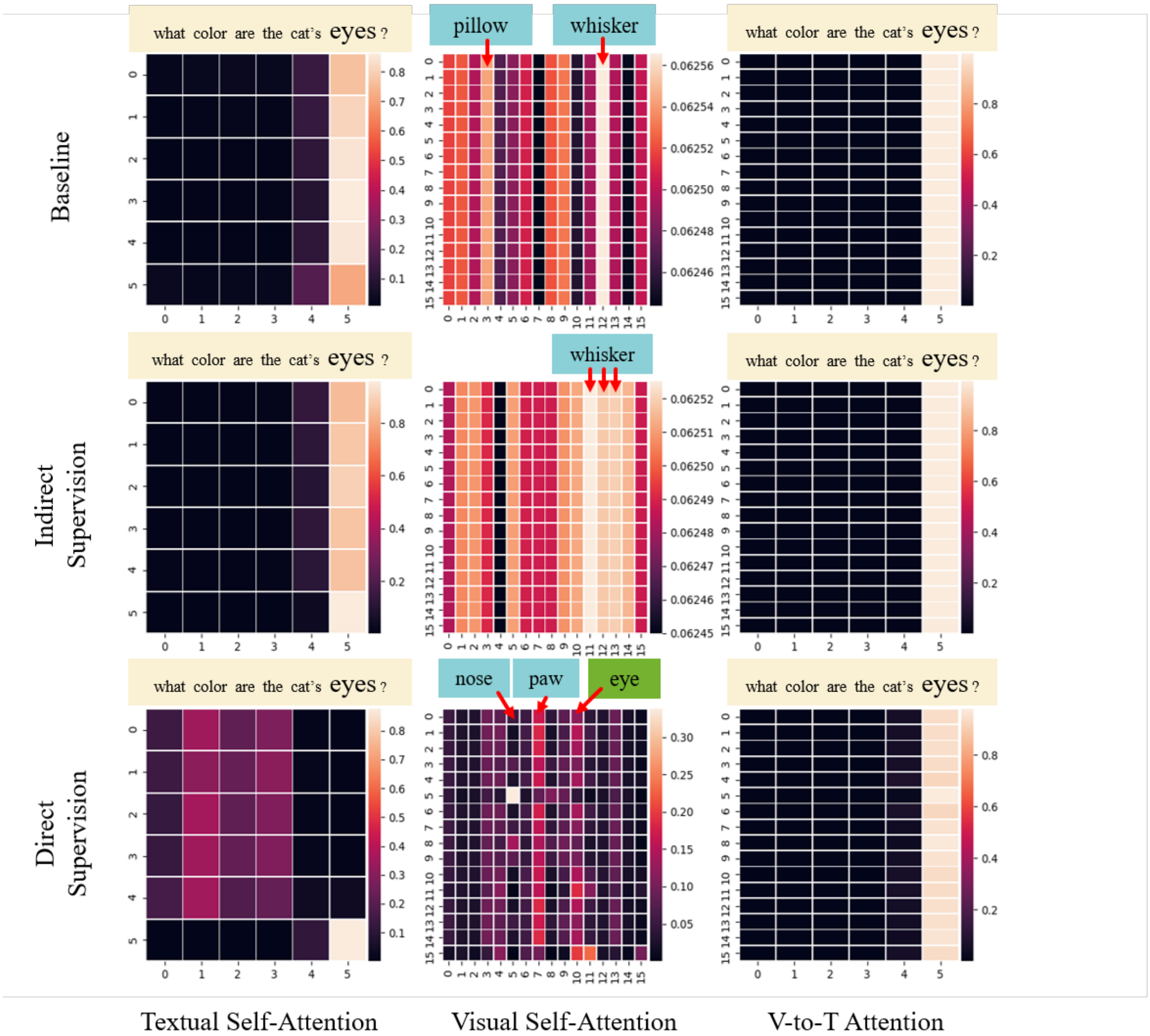

Figure 5: Visualization of the horizontally-normalized attention weights of the last transformer layer in the encoder and decoder of MCAN. For textual and visual self-attentions, they are normalized along the horizontal axis. "V-to-T" means Visual-to-Textual attention. The brighter the cell, the higher weight it carries. The larger font of the text in yellow box, the higher weight it carries. The green box is the true attention object, the blue boxes are candidate regions. Multiple objects with the same name may exist in the image. The direct supervision is the model trained with full attention supervision; the indirect supervision is the model trained with SGHMC-EM multitask learning.