Robust Visual Reasoning via Language Guided Neural Module Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural module networks (NMN) are a popular approach for solving multi-modal tasks such as visual question answering (VQA) and visual referring expression recognition (REF). A key limitation in prior implementations of NMNs is that the neural modules do not effectively capture the association between the visual input and the relevant neighbourhood context of the textual input. This limits their generalizability. For instance, NMNs fail to understand new concepts such as "yellow sphere to the left" even when it is a combination of known concepts from train data: "blue sphere", "yellow cube", and "metallic cube to the left". In this paper, we address this limitation by introducing a language-guided adaptive convolution layer (LG-Conv) into NMN, in which the filter weights of convolutions are explicitly multiplied with a spatially varying language-guided kernel. Our model allows the neural module to adaptively co-attend over potential objects of interest from the visual and textual inputs. Extensive experiments on VQA and REF tasks demonstrate the effectiveness of our approach. Additionally, we propose a new challenging out-of-distribution test split for REF task, which we call C3-Ref+, for explicitly evaluating the NMN's ability to generalize well to adversarial perturbations and unseen combinations of known concepts. Experiments on C3-Ref+ further demonstrate the generalization capabilities of our approach.

1 Introduction

Visual question answering (VQA) [6, 3] and visual referring expression recognition (REF) [32, 23] are fundamental language-to-vision matching tasks that have several downstream applications such as robot navigation [36], and image retrieval [43]. The high-level goal of these tasks is to perform joint reasoning over visual and textual queries. In the recent years, neural module networks (NMN) [5, 16, 28] attracted increasing attention due to their superior performance on these tasks [28, 22]. Briefly, NMN models learn to parse textual queries as executable programs composed of learnable *neural modules*. Each of these modules implements a single step of reasoning (e.g. `count`, `filter`, `compare`) and are dynamically assembled to perform multi-step reasoning over text.

Despite great success, the current NMN implementations require a large amount of training data and are less effective in generalizing to unseen but known language constructs [25, 8]. For example, NMNs fail to understand new concepts such as "*yellow sphere to the left*" that are constructed using a combinations of known concepts from train data such as "*blue sphere*", "*yellow cube*", and "*metallic cube to the left*". One of the main reasons for this is that the neural modules in existing works either use a shallow, indirect language guidance [34, 15] or pre-define the textual inputs in the module instantiation [22, 28], ignoring the rich correlations among the visual inputs and the relevant context from the textual inputs. For example, the neural module that filters based on the object size, "`filter_size(smallest)`", needs to localize a tiny sphere or a medium-sized sphere in the image depending on the object relationships in the expression (e.g. "*the smallest thing among the*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

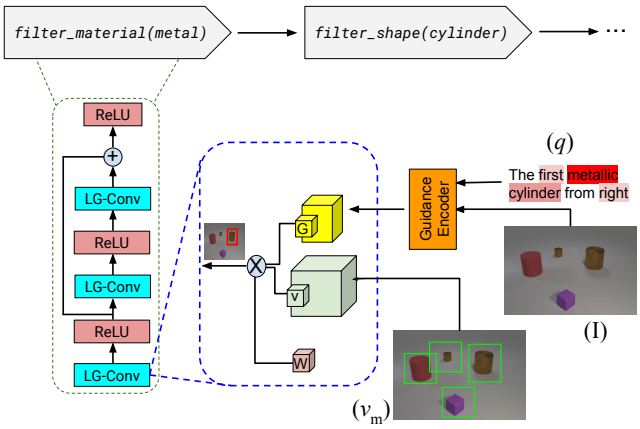


Figure 1: An example from the CLEVR-Ref+ dataset. Existing NMN implementations only provide the visual features (v_m) as inputs to the neural modules. In this work, we additionally condition each module on textual expression (q) by replacing the standard convolution layers with content adaptive convolution layers **LG-Conv** which modify the convolution by explicitly multiplying the filter weights (W) with a spatially varying language-guided kernel G. $\bigotimes$ denotes element-wise multiplication and $\bigoplus$ denotes summation.

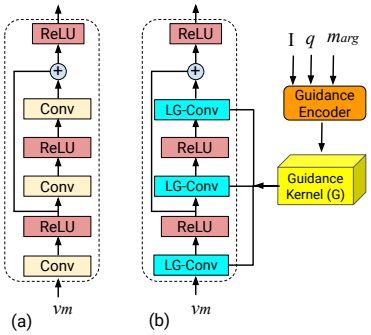


Figure 2: (a) Architecture of neural module (m) in existing NMNs [22] consuming a visual input v_m; (b) Our proposed architecture replacing Conv layers with content adaptive convolution layers guided by the input image I, input query q and parameterized textual input m_{arg}.

spheres" vs. "*the metallic sphere smaller than all the large cylinders*") and the different sizes of spheres and cylinders available in its visual input. We believe that explicitly conditioning the neural modules on the joint textual and visual context helps in inferring robust visiolinguistic relationships which further enhances the compositional reasoning skills.

In this work, we address the aforementioned issues by explicitly providing the relevant objects and relationships in the textual expression to neural modules. To do this, as shown in Figure 1, we replace the standard convolution operations in the neural modules with a novel language-guided adaptive convolution operation, which we call **LG-Conv**. More specifically, the filter weights W of LG-Conv are explicitly multiplied with a spatially varying language-guided kernel G, which allows the module to adaptively co-attend over potential objects of interest from the visual input and textual input by altering the convolution. Although content-adaptive convolutions [20, 11, 38] are used in several vision tasks, we are not aware of any prior works that does this filter adaptation using language as guidance. We propose two novel and effective methods namely, bi-salient attentional guidance (BiSAtt) network and co-salient attentional guidance guidance (CoSAtt) network to learn the guidance kernel G from textual and visual inputs.

We conduct extensive experiments on VQA and REF tasks using CLEVR [21], CLOSURE [8], CLEVR-Ref+ [28], and Cops-Ref [9] datasets. On the recently released VQA benchmark CLOSURE [8], our approach significantly outperforms all the previous works with 11.6% improvements in accuracy. On the REF benchmark CLEVR-Ref+ [28], we outperform competing approaches by as much as +9.8% accuracy on single-referent split (S-Ref) and +4.7% on full-referent split (F-Ref), suggesting the importance of language-guidance. Most significant gains with S-Ref, which consists of only 30% of the training data in CLEVR-Ref+, demonstrate the superior generalization of our model in learning from fewer training samples[1]. We further evaluate the generalization capabilities of our approach by collecting a new dataset consisting of unseen compositions and contrasting samples for CLEVR-Ref+ benchmark [13], and call our new dataset C3-Ref+. We significantly outperform the state-of-the-art models by as much as +24.7% absolute accuracy on C3-Ref+.

Our key contributions are summarized as follows:

1. We propose a novel language-guided adaptive convolution layer for NMN that guide modules in adaptively selecting informative visiolinguistic relationships and in attending to relevant objects of interest from the visual and textual inputs;

[1]S-Ref is a subset of full CLEVR-Ref+ dataset containing expressions that refer to only a single target object in the image (see Appendix).

2. We demonstrate the superiority of our approach by achieving new state-of-the-art results on multiple tasks and benchmarks;
3. We introduce a new benchmark to explicitly test the model's ability to generalize to adversarial perturbations and novel compositions of concepts unseen during training. We show that our model is more robust and generalizable compared to previous approaches.

To facilitate future research on generalizable module networks, we will publicly release our code, pre-trained models, and the proposed C3-Ref+ benchmark.

2 Related Work

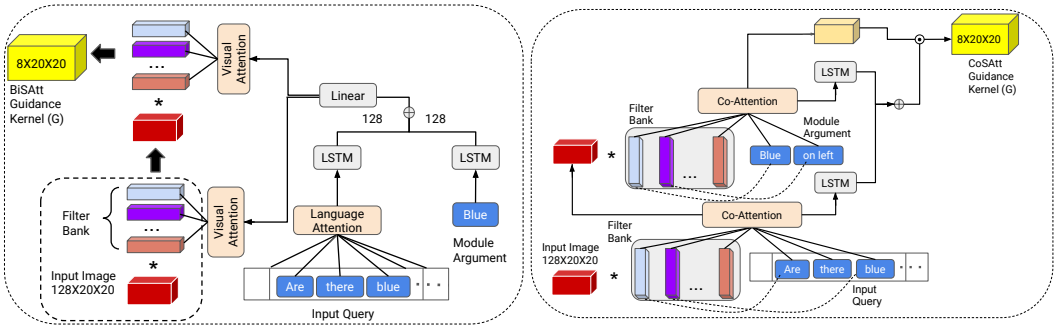
Multi-modal Grounding. Early approaches for tackling grounding tasks such as VQA and REF used recurrent networks with CNNs and attention-based models to jointly understand visual and language inputs [12, 41, 29]. [3] proposed bottom-up and top-down approach to learn attention over image regions obtained from a pretrained object detector. However, these models are shown to be heavily driven by annotation artifacts in the training data [1]. Balanced and synthetic datasets such as CLEVR [21], CLOSURE [8], CLEVR-Ref+ [28] are proposed by explicitly controlling the bias and language priors. Transformers [31, 27, 40], using pretrain-then-transfer approach, have shown superior performance on these datasets. However, these models fail to learn robust visio-linguistic representations and are shown to exploit the imbalanced distribution in the train and test splits [2, 10].

Neural Module Networks. Neural module networks (NMNs) leverage specialized modules to compute basic reasoning tasks. These modules can be assembled to perform complex and compositional reasoning [5, 21, 22, 15]. [4] proposed dynamic NMNs that learns and adapts the structure of the execution layouts to the question. Recently, [22] proposed homogeneous (IEP) and generic neural modules, unlike fixed and hand-crafted neural module, in which the semantics of each neural module is learnt during training. IEP model achieves promising performance on CLEVR dataset. [28] proposed IEP-Ref by extending IEP model to CLEVR-Ref+ dataset and outperformed all the prior works. Although, compositional by design, these models lack robust generalization abilities and fail to ground novel combinations of known linguistic constructs [8]. Few implementations of NMN such as FiLM [34] and N2NMN [15] condition the neural modules on textual guidance. However, the visiolinguistic context in these modules is rather shallow as they cannot jointly co-attend over potential objects of interest directly from the visual input and textual inputs. The major difference between our work and the prior works of NMN is that we explicitly condition the neural modules on the language-guidance by directly altering and adapting the convolution operation.

Adaptive Convolutions. There is an extensive literature on content-adaptive convolutions, in which standard 2D convolutions are generalized to high-dimensional convolutions [7, 19, 35, 14]. [19] proposes to use bilateral filtering layers inside CNN architectures. [20] proposed dynamic filter networks in which input-specific custom filter weights are predicted by a different network branch. [38] proposed pixel-adaptive convolutions which modifies the filters in a position-specific fashion. While these prior works study adaptive convolution techniques in CNN representation tasks, it has not been explored before for guiding NMNs for multi-modal language-vision tasks. Closest to our work is the FiLM [34] model that learn parameters for scaling up or down the CNN activations by conditioning on the textual input. Our model differs from FiLM primarily in using language guidance that depends on learnable, local pixel and textual features.

3 Approach

Problem Setup and Notation. Given an image I and a natural language query q as input, our goal is to develop a NMN model that selects an answer $a \in A$ to the query from a fixed set A of possible answers. We generalize this notation for both VQA and REF tasks; q, a denote question and a natural language answer respectively in VQA, whereas they represent a referring expression and a bounding box of the target object respectively in REF. We represent input image I as an ordered sequence of a set of image regions $R = (r_0, r_1, ..., r_N)$ and the query q as the set of words $(w_1, w_2, ..., w_L)$ where w_i is the i-th word, N is the number of image regions extracted from input image I, and L is the total number of the tokens in the input query.



(a) (b)

Figure 3: (a) Bi-Salient Attentional Guidance Encoder (**BiSAtt**); (b) Co-Salient Attentional Guidance Encoder (**CoSAtt**). In BiSAtt Encoder architecture, we first encode text inputs and then use it to learn a set of adaptive weights to linearly combine the basis filters which produces the convolution filters applied on input image. CoSAtt Encoder jointly attends over the input image and text inputs (early fusion) to identify co-salient regions and relationships in visual and language features that are contextually associated with each other.

Similar to [22], we use a two-stage model for generating answer: (1) Program Generation Model $p(z|q;\theta_p)$: where the query is parsed to z representing the reasoning steps required to answer the query, and (2) Program Execution Model $p(a|z,I;\theta_e)$: where the predicted program z is used to assemble a input-specific neural network that is composed from a set of neural modules m and is executed to produce a distribution over answers.

As shown in Figure 2(a), the neural modules in current implementations of NMN [21, 28] typically use a standard Residual Convolution Block (RCB), consisting of convolution layers and ReLU activations. Formally, a module (f_n) of arity n receives n feature maps ($\mathbf{F_i}$) of shape $128 \times 20 \times 20$ and outputs a same-sized tensor $f_o = f_n(\mathbf{F_1}, \mathbf{F_2}, ..., \mathbf{F_n})$.

$$\begin{aligned} h_m &= \text{ReLU}\left(\text{conv}_1\left(F_i\right)\right) \\ f_o &= \text{ReLU}\left(\text{conv}_2\left(h_m\right) \oplus h_m\right) \end{aligned} \tag{1}$$

As we can see, these modules [21, 22] are not explicitly conditioned on the input expression q, and therefore fails to extract robust visiolinguistic relationships. In contrast, as shown in Figure 2(b), we explicitly condition the neural modules on q, in addition to visual inputs, by replacing the standard convolution operations in RCB block with a novel adaptive and language-guided convolution operation, which we call LG-Conv. Also, we parameterize the module arguments, i.e., for example, we treat "`filter_material`" module as parameterized by argument (m_{arg}) "*rubber*" instead of as a standalone module "`filter_material[rubber]`". As a result of this parametrization, the number of a distinct set of modules used in the parameterized model reduce by 75%. We condition our LG-Conv layer on both query q and the module argument m_{arg} (See Figure 2). In the following, we describe the LG-Conv operation and detail its formal specification.

3.1 Language-Guided Convolution

Our high-level goal is to empower the neural modules to learn adapting visiolingustic features from both visual and language inputs. We achieve this by introducing novel LG-Conv layers which allows the module to adaptively co-attend over potential objects of interest from the visual and textual inputs by altering the convolution [20, 11, 38]. Formally, a standard *conv* layer in the RCB block performing a spatial convolution operation over the n image pixels $P = (p_1, p_2, ...p_n)$ is given as:

$$\mathbf{p_i'} = \sum_{\mathbf{j} \in \mathbf{\Omega(i)}} \mathbf{W}[\mathbf{c_i} - \mathbf{c_j}]\mathbf{p_j} + \mathbf{b} \tag{2}$$

where $\mathbf{W}$ denotes the filter weights, $\mathbf{c_i}$ denote the coordinates of the pixels in the image, $\mathbf{b}$ denotes biases, and $\mathbf{\Omega(i)}$ defines a convolution window. This convolution operation, with spatially shared

weights, is agnostic to pixel features and independent of language features. As shown in Figure 1, we modify this to depend on both pixel features and language features using a spatially varying guidance kernel G as follows:

$$\mathbf{p_i'} = \sum_{\mathbf{j} \in \mathbf{\Omega(i)}} \mathbf{G(g_i, g_j) W [c_i - c_j] p_j + b} \tag{3}$$

The spatial convolution W is adapted at each pixel in the visual input using the guidance kernel G. Similar to [38], we represent G using a fixed parametric Gaussian: $G(g_i, g_j) = \exp(-\frac{1}{2}(g_i - g_j)^T(g_i - g_j))$, where g represents guidance features that we learn using the following two methods: (a) **Bi-Salient Attentional Guidance (BiSAtt Encoder)**: We generate spatial guidance features using the architecture shown in Figure 3a. The input image of dimensions $128 \times 20 \times 20$, the input query q, and the module's parametrized text argument are used in producing the guidance features. Specifically, in BiSAtt architecture, we add visual attention layers over I to generate spatial guidance from non-spatial q. (b) **Co-Salient Attentional Guidance (CoSAtt Encoder)**: Here, we apply a joint attention over I, q, and the module argument to identify co-salient regions and relationships in visual and language features that are contextually associated with each other. The architecture is shown in Figure 3b. In comparison to BiSAtt, as we show in our experiments, CoSAtt improves the relevance and interaction between objects in the image and the query. For efficient implementation, we use the same learned guidance across all the LG-Conv layers in a RCB block.

As our parametrized model require only a few number of modules, the total number of parameters in our NMN is significantly less compared to the state-of-the-art models, even though the network parameters in our parameterized module slightly increase due to the additional conv and LSTM units in the guidance encoder (see Appendix).

Program Generator. We implement program generator using an attention-based sequence to sequence (seq2seq) model with an encoder-decoder structure [39, 21] to map the input query q into an executable program z. Both the encoder and decoder have two hidden layers with a 256-dim hidden vector. Similar to [21], we convert the decoded sequence of program functions to syntax trees (in an in-order traversal) in which each node contains a RCB module.

Execution Engine. The execution engine assembles a neural network using the predicted program z by mapping function f at each node in syntax tree to its corresponding neural module. The parent modules in the syntax tree takes the outputs from the child modules. Since we use a homogeneous architecture for designing our modules, the output generated from all modules is of same shape $128 \times 20 \times 20$. We flatten the final feature map before passing it to a multi-layer perceptron classifier, producing a distribution over all possible answers.

Training. During training, we find the optimal module parameters by maximizing the likelihood of the data. We optimize $p(z|q; \theta_p)$ using a policy gradient method.

$$\nabla \mathbf{J}(\theta_\mathbf{p}) = \mathbb{E}[\nabla \log \mathbf{p(z|q;} \theta_\mathbf{p}) \cdot \mathbf{r}] \tag{4}$$

where r is the reward and the expectation is taken with respect to rollouts of the policy. In order to enforce the network for generating the most accurate predictions, we then train the execution engine directly by maximizing $\log p(a|z, I, q; \theta_e)$ with respect to θ_e.

$$\mathbb{E}[\nabla \log \mathbf{p(z|q;} \theta_\mathbf{p}) \cdot \log \mathbf{p(a|z, I, q;} \theta_\mathbf{e})] \tag{5}$$

4 Experiments

In this section, we start by discussing the datasets and the baselines considered in evaluating our approach. We show the superiority of our approach by performing quantitative and qualitative comparisons between our method and the baselines. We then demonstrate the importance of the proposed language guided adaptive convolution through ablation studies. Finally, we present our new benchmark, C3-Ref+, and show that our approach is robust to adversarial inputs and generalizes well to novel compositions.

Model	CLEVR-Dev	CLEVR-Test	CLOSURE	Cops-Ref (Full)
IEP-Ref [28]	$98.7^{\pm 0.3}$	$97.1^{\pm 0.2}$	$59.8^{\pm 0.4}$	NA
FiLM [34]	96.2	96.9	58.9	24.6
MAC [18]	99.1	98.2	71.6	26.7
Vector NMN [8]	98.8	97.6	71.0	NA
NS-VQA [42]	**99.2**	**99.4**	76.4	NA
LCGN [17]	NA	NA	NA	32.7
ViLBERT [30]	95.3	93.0	51.2	29.7
Visual BERT [30]	96.0	92.8	50.6	27.0
Ours (with BiSAtt)	$98.9^{\pm 0.2}$	$99.2^{\pm 0.1}$	$86.1^{\pm 0.1}$	**35.5**
Ours (with CoSAtt)	$98.9^{\pm 0.1}$	$99.2^{\pm 0.1}$	$\mathbf{88.0^{\pm 0.2}}$	**35.5**

Table 1: Performance of our approach and baselines on CLEVR, CLOSURE and Cops-Ref benchmarks.

4.1 Datasets and Baselines

We evaluate our approach on both VQA and REF benchmarks. We use CLEVR [21] as the VQA benchmark, consisting of synthetically generated image and question pairs. Specifically, it consists of 100K images and 860K questions. We train our model on CLEVR train split and evaluate the performance on its val and test splits. In addition, using the model trained on CLEVR, we evaluate the performance on CLOSURE benchmark [8], consisting of novel compositions of objects and relations not seen in CLEVR train split. We then report results on CLEVR-Ref+ [28], a synthetic benchmark for referring expressions. It contains nearly 0.8M referring expressions of which 32% of expressions refer to only a single object (*Single-referent*) and 68% refer to more than one object (*Multi-referent*). We refer to the full dataset as F-Ref and the single-referent subset as S-Ref. Additionally, we report performance on Cops-Ref [9], a REF dataset for evaluating compositional reasoning skills [2].

We compare the performance of our approach against several NMN baselines such as FiLM (Feature-wise Linear Modulation), MAC [18], IEP-Ref [28] VectorNMN [8], and non-NMN baselines such as NS-VQA [42], LCGN [17], ViLBERT [30], and VisualBERT [26].

Model	S-D	S-T	F-D	F-T
IEP-Ref	$49.8^{\pm 0.1}$	$51.5^{\pm 0.6}$	$80.5^{\pm 0.2}$	$78.2^{\pm 0.3}$
FiLM	44.9	46.7	76.5	75.7
MAC	46.3	49.2	81.3	77.4
Vector NMN	48.3	53.5	83.2	77.1
NS-VQA	51.5	52.9	82.5	79.6
LCGN	46.8	48.0	77.0	74.8
ViLBERT	42.4	44.3	69.3	68.7
Visual BERT	41.7	43.2	69.8	63.2
Ours (with BiSAtt)	$61.1^{\pm 0.3}$	$59.7^{\pm 0.2}$	$87.2^{\pm 0.3}$	$83.5^{\pm 0.2}$
Ours (with CoSAtt)	$\mathbf{62.3^{\pm 0.1}}$	$\mathbf{63.3^{\pm 0.1}}$	$\mathbf{89.1^{\pm 0.2}}$	$\mathbf{84.3^{\pm 0.3}}$

Table 2: Performance of our language-guided NMN models and state-of-the-art models on S-Ref Dev (S-D), S-Ref Test (S-T), F-Ref Dev (F-D) and F-Ref Test (F-T).

Model	CLS	S-T	F-T
Ours	**88.0**	**63.3**	**84.3**
C1 Ours-L+G	62.1	51.6	77.8
C2 Ours-L-G	61.5	52.1	75.2
C3 Ours+L+(G w/o I)	80.2	57.7	79.9
C4 Ours+L+(G w/o q)	78.8	54.1	76.2
C5 Ours+L+(G w/o m_{arg})	82.1	61.7	80.9

Table 3: **Ablations.** Performance of our model with and without LG-Conv layer (L) and CoSAtt encoder (G) on CLOSURE (CLS), S-Ref Test, and F-Ref Test.

4.2 Implementation Details

Similar to [21], we use 18K ground-truth programs to train the program generator (PG). We train PG and the execution engine using Adam [24] with learning rates 0.0005 and 0.0001, respectively. Our PG is trained for a maximum of 32K iterations, while EE is trained for a maximum of 450K iterations. We employ early stopping based on validation set accuracy. While reporting accuracies on S-Ref test split, we use the model trained on S-Ref train split. We repeat the experiment 5 times on each benchmark and report the mean/variance on each of them (see Appendix).

[2] In evaluating our model on Cops-Ref, as the ground-truth programs are not available for this benchmark, we use FiLM as our baseline model instead of IEP-Ref and apply the proposed language-guidance kernel to the neural modules in FiLM.

4.3 Evaluation

Table 1 and Table 2 show results in comparison with the baselines. We find that our model outperforms all prior work on CLOSURE, Cops-Ref, and CLEVR-Ref+ benchmarks, while showing on-par performance on CLEVR test split. This demonstrates the effectiveness of the proposed language guided convolutions in capturing visiolinguistic relations and contextual dependencies from the longer CLEVR-like expressions. In particular, we achieve +11.6% in accuracy on CLOSURE test split compared to the best prior model Vector-NMN, indicating that our model generalizes well to unseen compositions. The multi-modal transformer based approaches ViLBERT and VisualBERT performed poorly on both CLOSURE and CLEVR-Ref+, probably due to the mismatched image distribution in pre-training (with conceptual captions [37]) and fine-tuning. Our model improves the accuracy on CLEVR-Ref+ test splits by 9.8% on S-Ref and 4.7% on F-Ref, compared with the current state-of-the-art method IEP-Ref. Significant gains on S-Test also suggest the superior generalization skills of our model in learning from fewer training samples. Relatively more improvements with CoSAtt encoder compared to BiSAtt encoder shows that early fusion of image and text features facilitate in generating more robust guidance kernel.

To gain better insight into the relative contribution of the design choices we made, we perform experiment with the following five ablated models:
C1: Conv vs. LG-Conv (L). We investigate the contribution of the proposed content adaptive convolution layer in the RCB block by replacing LG-Conv layer with standard convolution. In this setting, we use guidance (G) from CoSAtt encoder for directly scaling up or down the CNN activations in the RCB block.
C2: Conditioning on CoSAtt Guidance (G). In this ablation, we use LG-Conv layers but skip the CoSAtt encoder to verify the importance of module level conditioning on the interaction between image and text features. We instead only pass the module argument (m_{arg}) as guidance to the LG-Conv layer.
C3: CoSAtt w/o. Image (I) We encode guidance using only input query q and the module argument to test the importance of conditioning on image I in the CoSAtt encoder.
C4: CoSAtt w/o. Query (q) We encode guidance using only input image to test the importance of conditioning on input query q in the CoSAtt encoder.
C5: CoSAtt w/o. Module Arg (m_{inp}) In this variant, we keep q and I, but skip m_{arg} in the CoSAtt encoder.

The results are shown in Table 3. As we can see, all the above five variants underperform, confirming the importance of our proposed content-adaptive convolutions and guidance kernel. Results show that module argument in the CoSAtt guidance has less significant effect compared to other components, suggesting that our model is able to infer the semantic context of the module. Figure 4 illustrates the qualitative differences of `filter_color(red)` module trained using IEP-Ref and our approach. With IEP-Ref, the model selects all red objects from the image, ignoring the context in the expression. On the other hand, our approach correctly locates objects based on their contextual relevance.

4.4 The C3-Ref+ Dataset

Following [13, 33, 8], we construct a new benchmark, C3-Ref+, to critically examine the generalization capabilities of NMNs in grounding out-of-domain (o.o.d) referring expressions - a fundamental expressive power of human intelligence. Specifically, C3-Ref+ consists of two kinds of samples constructed using S-Ref split of CLEVR-Ref+ dataset: (a) **Novel Compositions**, consisting of samples that evaluate the model on combinations of objects and their spatial relationships not seen in S-Ref train split; and (b) **Contrast Sets**, consisting of samples that help in exposing model brittleness by probing a model's decision boundary local to examples in the S-Ref test set. Figure 5 shows an example.

In constructing novel compositions of samples, we first extract all the combinations of simple and complex noun phrases (e.g. "*The metallic object*", "*The metallic object to the left*") from the S-Ref train split and then manually construct new expressions using the unseen pairs of these phrases. To construct a contrast sample, we manually perturb the semantics of various parts of the referring expressions in the S-Ref test split such that the ground-truth referent object changes. For example, we modify the expression *first one of the tiny <u>rubber</u> thing from <u>left</u>* to *first one of the tiny <u>metallic</u> thing from <u>right</u>*. We verify and validate the correctness of the collected samples and their ground-truth annotations using three human annotators. The annotations that are not consistent among the

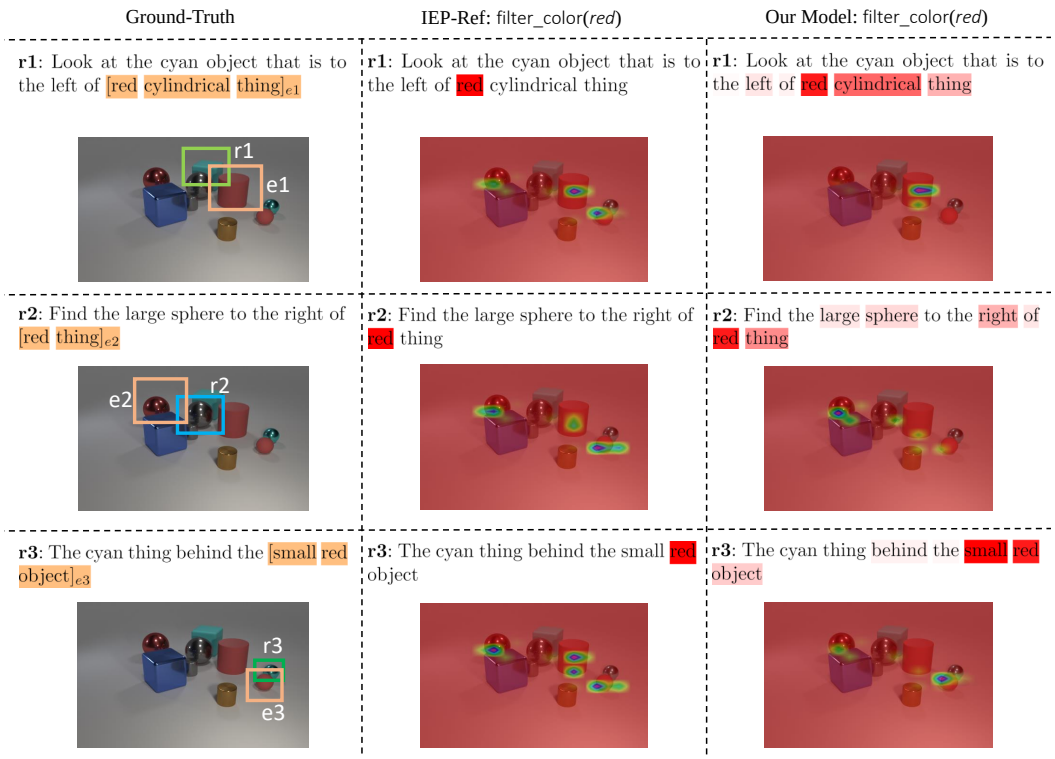


Figure 4: Qualitative examples showing attention heatmaps of `filter_color(red)` module outputs trained using IEP-Ref and our model on CLEVR-Ref+ dataset. $e1$, $e2$, $e3$ highlight the red objects that are referred in the input expressions $r1$, $r2$, and $r3$, respectively.

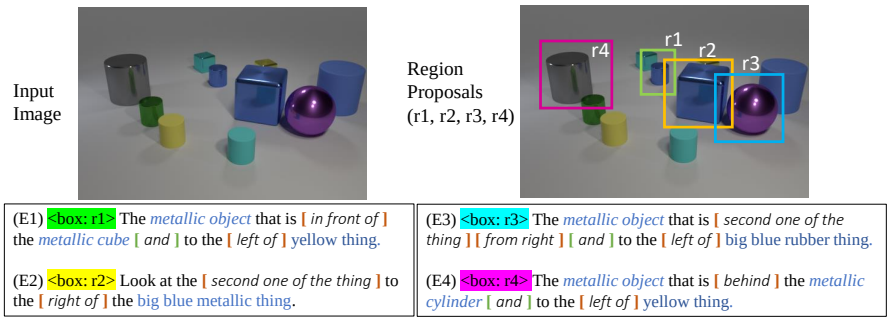
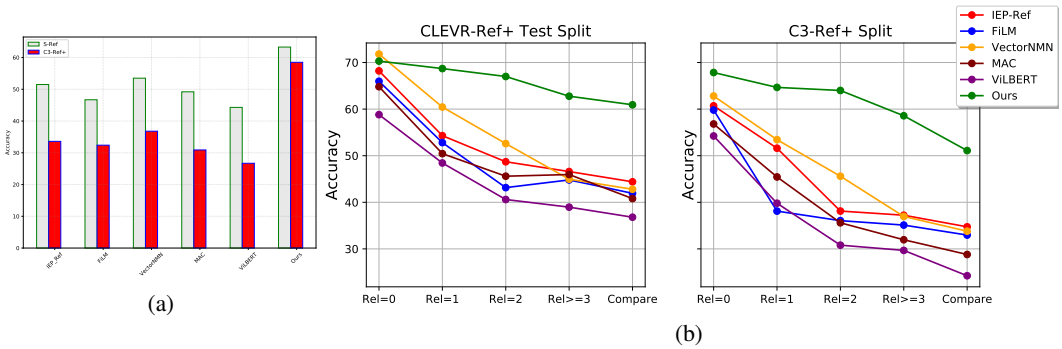


Figure 5: E1 and E2 are the original expressions in CLEVR-Ref+. E3 and E4 are examples from C3-Ref+. E3 is constructed using composition of E1 and E2. E4 is constructed by contrasting E1.

three human annotators are removed from the final split. In Figure 7, we summarize the size and complexity of our C3-Ref+ split and show the distribution of samples containing different types of spatial relations. More details are in Appendix.

4.5 Evaluation on C3-Ref+

Overall, as shown in Figure 6a, performance of state-of-the-art models drop by up to 18% on C3-Ref+. Our proposed method shows least drop (<5%) in performance indicating its superiority in grounding new unseen compositions and adversarial perturbations. We also compare the performance in terms of number of spatial relations (e.g. *in the front*, *to the left*, *of same material as*) present in the expressions. Results in Figure 6b suggest that the existing models struggle to correctly ground the expressions in C3-Ref+ containing compare relations (e.g. *same_color*) or that containing more than 2 spatial



(a) (b)

Figure 6: (a) Accuracy of models on S-Ref and C3-Ref+; (b) Performance of Models on S-Ref and C3-Ref+ based on the number of spatial relations in the input query.

Referring Expressions	4300
Unique Images	2930
Vocabulary	104
Avg. Length of Expr	21.2

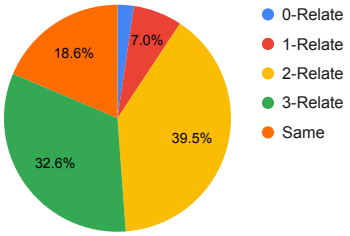


Figure 7: Statistics of C3-Ref+ dataset

relations (e.g. *front*, *left*) in the expression (see Appendix for more detailed analysis). Although our proposed method performs significantly well compared to baselines, there is still ample room for improvement. We hope that C3-Ref+ benchmark will foster more research in this area.

5 Conclusion

Neural module networks (NMNs) are widely used in language and vision tasks. We show that explicitly conditioning neural modules on the language guidance through adaptive convolutions improve their grounding and generalization abilities, achieving a new state-of-the-art results on the VQA and visual referring expressions tasks. Our analysis on CLOSURE, CLEVR-Ref+, Cops-Ref and a new compositional and contrastive split C3-Ref+ demonstrate that our proposed method enhances NMNs' ability in adaptively selecting and exploiting informative visiolinguistic relationships.

6 Broader Impact

This work contributes to improve the joint understanding of image and textual content, which in turn is a very important component in several vision and language grounding tasks. Our research can promote the development of a multi-modal interaction system and facilitate people's daily lives. This work does not present any foreseeable societal consequence.

References

[1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016.

[2] A. R. Akula, S. Gella, Y. Al-Onaizan, S.-C. Zhu, and S. Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Association for Computational Linguistics (ACL)*, 2020.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016.

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

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[7] V. Aurich and J. Weule. Non-linear gaussian filters performing edge preserving diffusion. In *Mustererkennung 1995*, pages 538–545. Springer, 1995.

[8] D. Bahdanau, H. de Vries, T. J. O'Donnell, S. Murty, P. Beaudoin, Y. Bengio, and A. Courville. Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783*, 2019.

[9] Z. Chen, P. Wang, L. Ma, K.-Y. K. Wong, and Q. Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095, 2020.

[10] V. Cirik, T. Berg-Kirkpatrick, and L.-P. Morency. Using syntax to ground referring expressions in natural images. In *AAAI Conference on Artificial Intelligence*, 2018.

[11] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[12] 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. 2016.

[13] M. Gardner, Y. Artzi, V. Basmova, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, et al. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

[14] K. He, J. Sun, and X. Tang. Guided image filtering. In *European conference on computer vision*, pages 1–14. Springer, 2010.

[15] 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*, pages 804–813, 2017.

[16] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124, 2017.

[17] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10294–10303, 2019.

[18] D. A. Hudson and C. D. Manning. Learning by abstraction: The neural state machine. In *Proceedings of NeurIPS*, 2019.

[19] V. Jampani, M. Kiefel, and P. V. Gehler. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4452–4461, 2016.

[20] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. In *Advances in neural information processing systems*, pages 667–675, 2016.

[21] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

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

[23] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

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

[25] B. M. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*, 2017.

[26] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[27] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

[28] R. Liu, C. Liu, Y. Bai, and A. L. Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 4185–4194, 2019.

[29] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Adv. Neural Inform. Process. Syst.*, 2016.

[30] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[31] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[32] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[33] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

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

[35] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen, et al. Dynamic sampling in convolutional neural networks for imbalanced data classification. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 112–117. IEEE, 2018.

[36] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. *arXiv preprint arXiv:1904.10151*, 2019.

[37] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Association for Computational Linguistics*, pages 2556–2565, 2018.

[38] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019.

[39] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[40] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[41] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016.

[42] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in neural information processing systems*, pages 1031–1042, 2018.

[43] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Checklist

- Did you include the license to the code and datasets? [N/A]
- Did you include the license to the code and datasets? [N/A]
- Did you include the license to the code and datasets? [N/A]

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]