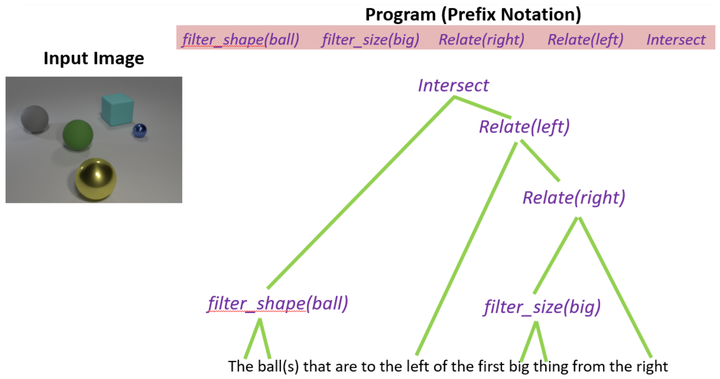
Robust Visual Reasoning via Language Guided Neural Module Networks

Arjun R Akula[1] Varun Jampani[2] Soravit(Beer) Changpinyo [2] Song-Chun Zhu[3]

[1]UCLA, [2]Google Research, [3]BIGAI; Peking University; Tsinghua University, [3]Mila; McGill University

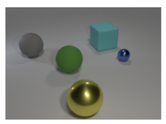
NEURAL INFORMATION PROCESSING SYSTEMS

Neural Module Network (NMN)



NMNs use an explicit modular reasoning process and provide interpretable output. NMNs first extract logical expression/program from input expression. Then composes a neural network to ground referring expressions.

NMN Limitations



... yellow sphere in the front

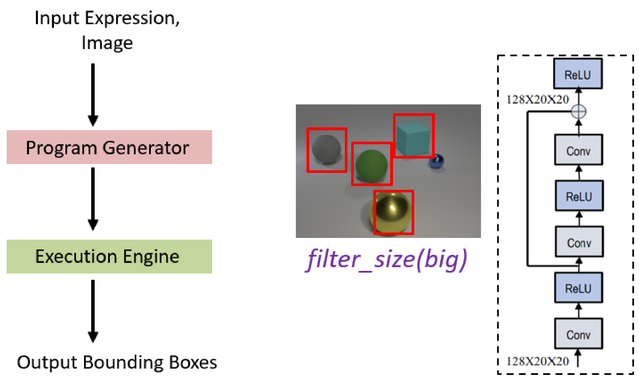
Seen Concepts in Training

... green sphere

... yellow cube

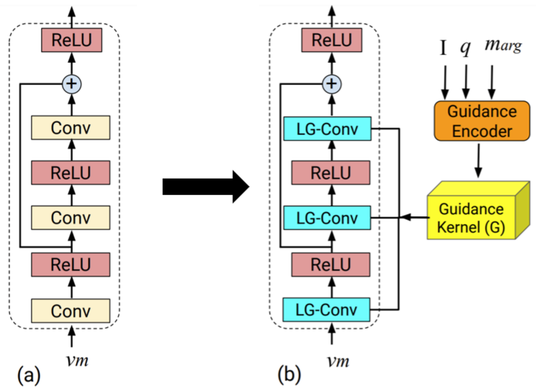
... cyan thing last from the front

Limitation 1: NMNs are less effective in generalizing to unseen but known language constructs



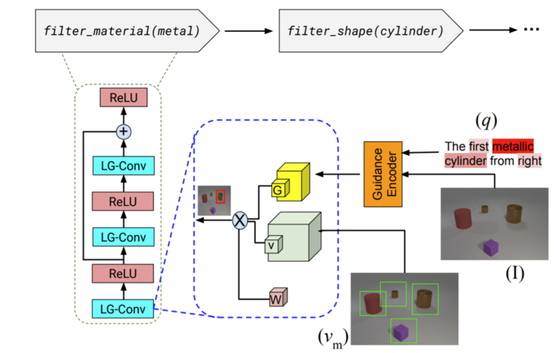
Limitation 2: NMNs use shallow, in-direct language guidance

Content Adaptive Convolutions

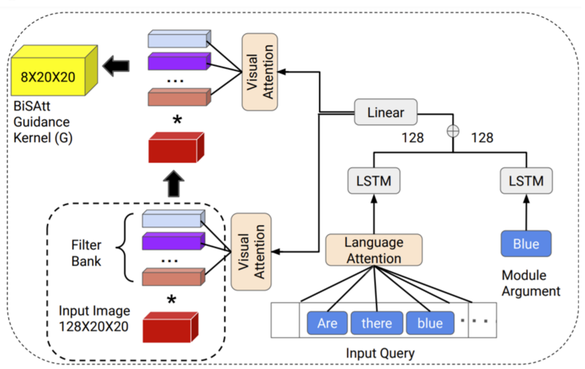


We replace the standard convolution layers with content adaptive convolution layers LG-Conv.

Bi-Salient Attention Guidance Encoder

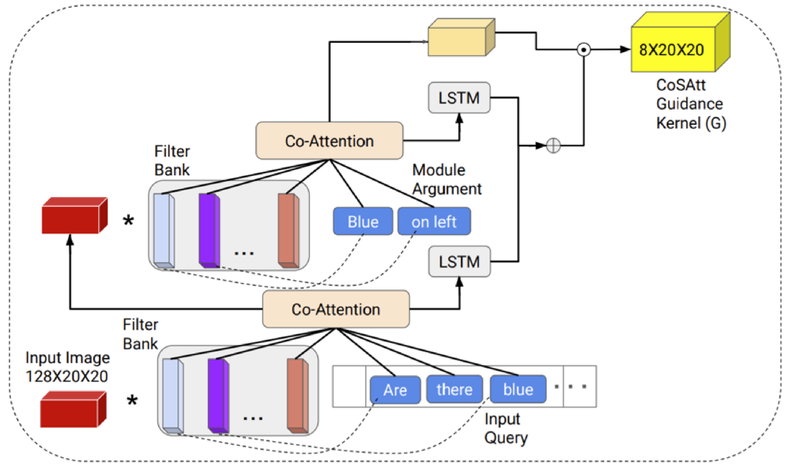


The spatial convolution W is adapted at each pixel in the visual input using the guidance kernel G.



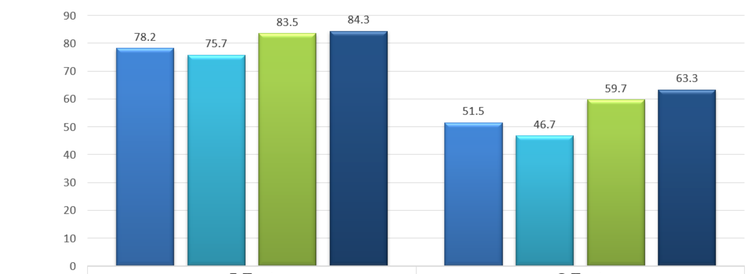
we first encode text inputs and then use it to learn a set of adaptive weights to linearly combine the basis filters.

Co-Salient Attention Guidance Encoder



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.

Experiment Results



	F-Test	S-Test
IEP-Ref	78.2	51.5
FiLM	75.7	46.7
BiSAtt	83.5	59.7
CoSAtt	84.3	63.3

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.

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.