

# Obtaining Faithful Interpretations from Compositional Neural Networks

Sanjay Subramanian<sup>\*1</sup> Ben Bogin<sup>\*2</sup> Nitish Gupta<sup>\*3</sup>  
 Tomer Wolfson<sup>1,2</sup> Sameer Singh<sup>4</sup> Jonathan Berant<sup>1,2</sup> Matt Gardner<sup>1</sup>  
<sup>1</sup>Allen Institute for AI <sup>2</sup>Tel-Aviv University  
<sup>3</sup>University of Pennsylvania <sup>4</sup>University of California, Irvine  
 {sanjays,mattg}@allenai.org, {ben.bogin,joberant}@cs.tau.ac.il,  
 nitishg@seas.upenn.edu, tomerwol@mail.tau.ac.il, sameer@uci.edu

## Abstract

Neural module networks (NMNs) are a popular approach for modeling compositionality: they achieve high accuracy when applied to problems in language and vision, while reflecting the compositional structure of the problem in the network architecture. However, prior work implicitly assumed that the structure of the network modules, describing the abstract reasoning process, provides a faithful explanation of the model’s reasoning; that is, that all modules perform their intended behaviour. In this work, we propose and conduct a systematic evaluation of the intermediate outputs of NMNs on NLVR2 and DROP, two datasets which require composing multiple reasoning steps. We find that the intermediate outputs differ from the expected output, illustrating that the network structure does not provide a faithful explanation of model behaviour. To remedy that, we train the model with auxiliary supervision and propose particular choices for module architecture that yield much better faithfulness, at a minimal cost to accuracy.

## 1 Introduction

Models that can read text and reason about it in a particular context (such as an image, a paragraph, or a table) have been recently gaining increased attention, leading to the creation of multiple datasets that require reasoning in both the visual and textual domain (Johnson et al., 2016; Suhr et al., 2017; Talmor and Berant, 2018; Yang et al., 2018a; Suhr et al., 2019; Hudson and Manning, 2019; Dua et al., 2019). Consider the example in Figure 1 from NLVR2: a model must understand the compositional sentence in order to then ground *dogs* in the input, count those that are *black* and verify that the count of all dogs in the image is equal to the number of black dogs.

<sup>\*</sup> Equal Contribution

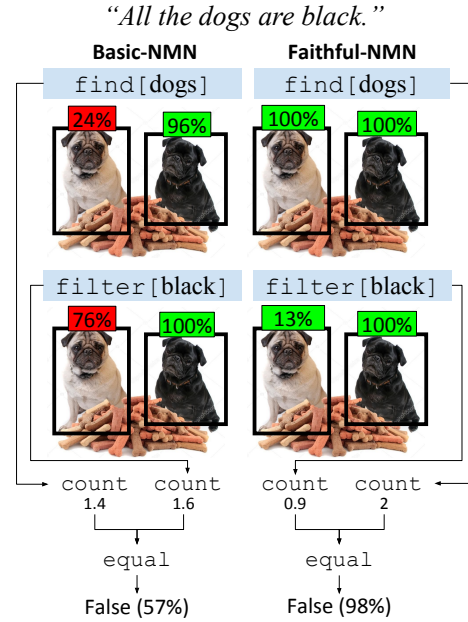


Figure 1: An example for a visual reasoning problem where both the Basic and Faithful NMNs produce the correct answer. The Basic NMN, however, fails to give meaningful intermediate outputs for the *find* and *filter* modules, whereas our improved Faithful-NMN assigns correct probabilities in all cases. Boxes are green if probabilities are as expected, red otherwise.

Both models that assume an intermediate structure (Andreas et al., 2016; Jiang and Bansal, 2019) and models without such structure (Tan and Bansal, 2019; Hu et al., 2019; Min et al., 2019) have been proposed for these reasoning problems. While good performance can be obtained without a structured representation, an advantage of structured approaches is that the reasoning process in such approaches is more *interpretable*. For example, a structured model can explicitly denote that there are two *dogs* in the image, but that one of them is *not black*. Such interpretability improves our scientific understanding, aids in model development, and improves overall trust in a model.

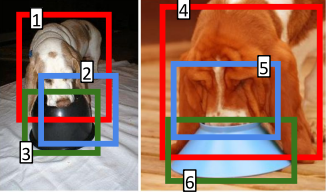
<p>two dogs are touching a food dish with their face</p> 		<b>Program</b> equal count with-relation [is touching] relocate [face] find [dog] find [food dish] number [two]	<b>Output</b> True 2 [2, 5] [2, 5] [1, 4] [3, 6] 2
<p>Who threw the longest touchdown pass in the second half?</p> <p>In the first quarter, the Texans trailed early after QB Kerry Collins threw a 19-yard TD pass [1] to WR Nate Washington. Second quarter started with kicker Neil Rackers made a 37-yard field goal, and the quarter closed with kicker Rob Bironas hitting a 30-yard field goal. The Texans tried to cut the lead with QB Matt Schaub getting a 8-yard TD pass [2] to WR Andre Johnson, but the Titans would pull away with RB Javon Ringer throwing a 7-yard TD pass [3]. The Texans tried to come back into the game in the fourth quarter, but only came away with Schaub [4] throwing a 12-yard TD pass [5] to WR Kevin Walter.</p>		<b>Program</b> relocate[who threw] find-max-num filter [the second half] find [touchdown pass]	<b>Output</b> Schaub [4] [5] [2, 3, 5] [1, 2, 3, 5]

Figure 2: An example for a mapping of an utterance to a gold program and a perfect execution in a reasoning problem from NLVR2 (top) and DROP (bottom).

Neural module networks (NMNs; Andreas et al., 2016) parse an input utterance into an executable program composed of learnable modules that are designed to perform atomic reasoning tasks and can be composed to perform complex reasoning against an unstructured context. NMNs are appealing since their output is interpretable; they provide a logical meaning representation of the utterance and also the outputs of the intermediate steps (modules) to reach the final answer. However, because module parameters are typically learned from end-task supervision only, it is possible that the program will not be a *faithful* explanation of the behaviour of the model (Ross et al., 2017; Wiegrefe and Pinter, 2019), i.e., the model will solve the task by executing modules according to the program structure, but the modules will not perform the reasoning steps *as intended*. For example, in Figure 1, a *basic NMN* predicts the correct answer False, but incorrectly predicts the output of the `find[dogs]` operation. It does not correctly locate one of the *dogs* in the image because two of the reasoning steps (`find` and `filter`) are collapsed into one module (`find`). This behavior of the `find` module is not faithful to its intended reasoning operation; a human reading the program would expect `find[dogs]` to *locate* all dogs. Such unfaithful module behaviour yields an unfaithful explanation of the model behaviour.

Unfaithful behaviour of modules, such as multiple reasoning steps collapsing into one, are undesirable in terms of interpretability; when a model fails to answer some question correctly, it is hard to tell which modules are the sources of error. While recent work (Yang et al., 2018b; Jiang and Bansal,

2019) has shown that one can obtain good performance when using NMNs, the accuracy of individual module outputs was mostly evaluated through qualitative analysis, rather than systematically evaluating the intermediate outputs of each module.

We provide three primary contributions regarding faithfulness in NMNs. First, we propose the concept of module-wise faithfulness – a systematic evaluation of individual module performance in NMNs that judges whether they have learned their intended operations, and define metrics to quantify this for both visual and textual reasoning (§3). Empirically, we show on both NLVR2 (Suhr et al., 2019) and DROP (Dua et al., 2019) that training a NMN using end-task supervision, even using *gold programs*, does *not* yield module-wise faithfulness, i.e., the modules do not perform their intended reasoning task. Second, we provide strategies for improving module-wise faithfulness in NMNs (§4). Specifically, (a) we demonstrate how module architecture affects faithfulness (§4.1), (b) propose supervising module outputs with either a proxy task or heuristically generated data (§4.2), and (c) show that providing modules with uncontextualized token representations improves faithfulness (§4.3). Figure 1 shows an example where our approach (*Faithful-NMN*) results in expected module outputs as compared to the *Basic-NMN*. Last, we collect human-annotated intermediate outputs for 536 examples in NLVR2 and for 215 examples in DROP to measure the module-wise faithfulness of models, and publicly release them for future work. Our code and data are available at <https://github.com/allenai/faithful-nmn>.

## 2 Neural Module Networks

**Overview** Neural module networks (NMNs; Andreas et al., 2016) are a class of models that map a natural language utterance into an executable program, composed of learnable modules that can be executed against a given context (images, text, etc.), to produce the utterance’s denotation (truth value in NLVR2, or a text answer in DROP). Modules are designed to solve atomic reasoning tasks and can be composed to perform complex reasoning. For example, in Figure 1, the utterance “All the dogs are black” is mapped to the program `equal(count(find[dogs]), count(filter[black](find[dogs])))`. The `find` module is expected to find all *dogs* in the image and the `filter` module is expected to output only the *black* ones from its input. Figure 2 shows two other example programs with the expected output of each module in the program.

A NMN has two main components: (1) parser, which maps the utterance into an executable program; and (2) executor, which executes the program against the context to produce the denotation. In our setup, programs are always trees where each tree node is a module. In this work, we focus on the executor, and specifically the faithfulness of module execution. We examine NMNs for both text and images, and describe their modules next.

### 2.1 Modules for visual reasoning

In this task, given two images and a sentence that describes the images, the model should output True iff the sentence correctly describes the images. We base our model, the Visual-NMN, on LXMERT (Tan and Bansal, 2019), which takes as input the sentence  $x$  and raw pixels, uses Faster R-CNN (Ren et al., 2015) to propose a set of bounding boxes,  $\mathcal{B}$ , that cover the objects in the image, and passes the tokens of  $x$  and the bounding boxes through a Transformer (Vaswani et al., 2017), encoding the interaction between both modalities. This produces a contextualized representation  $t \in \mathbb{R}^{|x| \times h}$  for each one of the tokens, and a representation  $v \in \mathbb{R}^{|\mathcal{B}| \times h}$  for each one of the bounding boxes, for a given hidden dimension  $h$ .

We provide a full list of modules and their implementation in Appendix A. Broadly, modules take as input representations of utterance tokens through an *utterance attention* mechanism (Hu et al., 2017), i.e., whenever the parser outputs a module, it also predicts a distribution over the utterance to-

kens  $(p_1, \dots, p_{|x|})$ , and the module takes as input  $\sum_{i=1}^{|x|} p_i t_i$ , where  $t_i$  is the hidden representation of token  $i$ . In addition, modules produce as output (and take as input) vectors  $p \in [0, 1]^{|\mathcal{B}|}$ , indicating for each bounding box the probability that it should be output by the module (Mao et al., 2019). For example, in the program `filter[black](find[dog])`, the `find` module takes the word ‘dog’ (using *utterance attention*, which puts all probability mass on the word ‘dog’), and outputs a probability vector  $p \in [0, 1]^{|\mathcal{B}|}$ , where ideally all bounding boxes corresponding to dogs have high probability. Then, the `filter` module takes  $p$  as input as well as the word ‘black’, and is meant to output high probabilities for bounding boxes with ‘black dogs’.

For the Visual-NMN we do not use a parser, but rely on a collected set of *gold* programs (including *gold utterance attention*), as described in §5. We will see that despite this advantageous setup, a basic NMN does not produce interpretable outputs.

### 2.2 Modules for textual reasoning

Our Text-NMN is used to answer questions in the DROP dataset and uses the modules as designed for DROP in prior work (Gupta et al., 2020) along with three new modules we define in this work. The modules introduced in Gupta et al. (2020) and used as is in our Text-NMN are `find`, `filter`, `relocate`, `count`, `find-num`, `find-date`, `find-max-num`, `find-min-num`, `num-compare` and `date-compare`. All these modules are probabilistic and produce, as output, a distribution over the relevant support. For example, `find` outputs a distribution over the passage tokens and `find-num` outputs a distribution over the numbers in the passage. We extend their model and introduce additional modules; `addition` and `subtraction` to add or subtract passage numbers, and `extract-answer` which directly predicts an answer span from the representations of passage tokens without any explicit compositional reasoning. We use BERT-base (Devlin et al., 2019) to encode the input question and passage.

The Text-NMN does not have access to gold programs, and thus we implement a parser as an encoder-decoder model with attention similar to Krishnamurthy et al. (2017), which takes the utterance as input, and outputs a linearized abstract syntax tree of the predicted program.

### 3 Module-wise Faithfulness

Neural module networks (NMNs) facilitate interpretability of their predictions via the reasoning steps in the structured program and providing the outputs of those intermediate steps during execution. For example, in Figure 2, all reasoning steps taken by both the Visual-NMN and Text-NMN can be discerned from the program and the intermediate module outputs. However, because module parameters are learned from an end-task, there is no guarantee that the modules will learn to perform their intended reasoning operation. In such a scenario, when modules do not perform their intended reasoning, the program is no longer a faithful explanation of the model behavior since it is not possible to reliably predict the outputs of the intermediate reasoning steps given the program. Work on NMNs thus far (Yang et al., 2018b; Jiang and Bansal, 2019) has overlooked systematically evaluating faithfulness, performing only qualitative analysis of intermediate outputs.

We introduce the concept of *module-wise faithfulness* aimed at evaluating whether each module has correctly learned its intended operation by judging the correctness of its outputs in a trained NMN. For example, in Figure 2 (top), a model would be judged module-wise faithful if the outputs of all the modules, `find`, `relocate`, and `with_relation`, are correct – i.e. similar to the outputs that a human would expect. We provide gold programs when evaluating faithfulness, to not conflate faithfulness with parser accuracy.

#### 3.1 Measuring faithfulness in Visual-NMN

Modules in Visual-NMN provide for each bounding box a probability for whether it should be a module output. To evaluate intermediate outputs, we sampled examples from the development set, and annotated gold bounding boxes for each instance of `find`, `filter`, `with_relation` and `relocate`. The annotator draws the correct bounding-boxes for each module in the gold program, similar to the output in Figure 2 (top).

A module of a faithful model should assign high probability to bounding-boxes that are aligned with the annotated bounding boxes and low probabilities to other boxes. Since the annotated bounding boxes do not align perfectly with the model’s bounding boxes, our evaluation must first induce an alignment. We consider two bounding boxes as “aligned” if the intersection-over-union (IOU) between them

exceeds a pre-defined threshold  $T = 0.5$ . Note that it is possible for an annotated bounding box to be aligned with several proposed bounding boxes and vice versa. Next, we consider an *annotated* bounding box  $B_A$  as “matched” w.r.t a module output if  $B_A$  is aligned with a proposed bounding box  $B_P$ , and  $B_P$  is assigned by the module a probability  $> 0.5$ . Similarly, we consider a *proposed* bounding box  $B_P$  as “matched” if  $B_P$  is assigned by the module a probability  $> 0.5$  and is aligned with some annotated bounding box  $B_A$ .

We compute precision and recall for each module type (e.g. `find`) in a particular example by considering all instances of the module in that example. We define *precision* as the ratio between the number of matched proposed bounding boxes and the number of proposed bounding boxes assigned a probability of more than 0.5. We define *recall* as the ratio between the number of matched annotated bounding boxes and the total number of annotated bounding boxes.<sup>1</sup>  $F_1$  is the harmonic mean of precision and recall. Similarly, we compute an “overall” precision, recall, and  $F_1$  score for an example by considering all instances of all module types in that example. The final score is an average over all examples. Please see Appendix B.2 for further discussion on this averaging.

#### 3.2 Measuring faithfulness in Text-NMN

Each module in Text-NMN produces a distribution over passage tokens (§2.2) which is a soft distributed representation for the selected spans. To measure module-wise faithfulness in Text-NMN, we obtain annotations for the set of spans that should be output by each module in the gold program (as seen in Figure 2 (bottom)) Ideally, all modules (`find`, `filter`, etc.) should predict high probability for tokens that appear in the gold spans and *zero* probability for other tokens.

To measure a module output’s correctness, we use a metric akin to cross-entropy loss to measure the deviation of the predicted module output  $p_{\text{att}}$  from the gold spans  $S = [s_1, \dots, s_N]$ . Here each span  $s_i = (t_s^i, t_e^i)$  is annotated as the start and end tokens. Faithfulness of a module is measured by: 
$$I = - \sum_{i=1}^N \left( \log \sum_{j=t_s^i}^{t_e^i} p_{\text{att}}^j \right)$$
. Lower cross-entropy corresponds to better faithfulness of a module.

<sup>1</sup>The numerators of the precision and the recall are different. Please see Appendix B.1 for an explanation.



## 4 Improving Faithfulness in NMNs

Module-wise faithfulness is affected by various factors: the choice of modules and their implementation (§ 4.1), use of auxiliary supervision (§ 4.2), and the use of contextual utterance embeddings (§ 4.3). We discuss ways of improving faithfulness of NMNs across these dimensions.

### 4.1 Choice of modules

**Visual reasoning** The count module always appears in NLVR2 as one of the top-level modules (see Figures 1 and 2).<sup>2</sup> We now discuss how its architecture affects faithfulness. Consider the program, `count(filter[black](find[dogs]))`. Its gold denotation (correct count value) would provide minimal feedback using which the *descendant* modules in the program tree, such as `filter` and `find`, need to learn their intended behavior. However, if `count` is implemented as an expressive neural network, it might learn to perform tasks designated for `find` and `filter`, hurting faithfulness. Thus, an architecture that allows counting, but also encourages *descendant* modules to learn their intended behaviour through backpropagation, is desirable. We discuss three possible count architectures, which take as input the bounding box probability vector  $\mathbf{p} \in [0, 1]^{|B|}$  and the visual features  $\mathbf{v} \in \mathbb{R}^{|B| \times h}$ .

**Layer-count module** is motivated by the count architecture of Hu et al. (2017), which uses a linear projection from image attention, followed by a softmax. This architecture explicitly uses the visual features,  $\mathbf{v}$ , giving it greater expressivity compared to simpler methods. First we compute  $\mathbf{p} \cdot \mathbf{v}$ , the weighted sum of the visual representations, based on their probabilities, and then output a scalar count using:  $\text{FF}_1(\text{LayerNorm}(\text{FF}_2(\mathbf{p} \cdot \mathbf{v})))$ , where  $\text{FF}_1$  and  $\text{FF}_2$  are feed-forward networks, and the activation function of  $\text{FF}_1$  is ReLU in order to output positive numbers only.

As discussed, since this implementation has access to the visual features of the bounding boxes, it can learn to perform certain tasks itself, without providing proper feedback to descendant modules. We show in §5 this indeed hurts faithfulness.

**Sum-count module** on the other extreme, ignores  $\mathbf{v}$ , and simply computes the sum  $\sum_{i=1}^{|B|} p_i$ . Being

<sup>2</sup>Top-level modules are Boolean quantifiers, such as number comparisons like `equal` (which require count) or `exist`. We implement `exist` using a call to `count` and `greater-equal` (see Appendix A), so `count` always occurs in the program.

parameter-less, this architecture provides direct feedback to descendant modules on how to change their output to produce better probabilities. However, such a simple functional-form ignores the fact that bounding boxes are overlapping, which might lead to over-counting objects. In addition, we would want `count` to ignore boxes with low probability. For example, if `filter` predicts a 5% probability for 20 different bounding boxes, we would not want the output of `count` to be 1.0.

**Graph-count module** (Zhang et al., 2018) is a middle ground between both approaches - the naïve *Sum-Count* and the flexible *Layer-Count*. Like *Sum-Count*, it does not use visual features, but learns to ignore overlapping and low-confidence bounding boxes while introducing only a minimal number of parameters (less than 300). It does so by treating each bounding box as a node in a graph, and then learning to prune edges and cluster nodes based on the amount of overlap between their bounding boxes (see paper for further details). Because this is a light-weight implementation that does not access visual features, proper feedback from the module can propagate to its descendants, encouraging them to produce better predictions.

**Textual reasoning** In the context of Text-NMN (on DROP), we study the effect of several modules on interpretability.

First, we introduce an `extract-answer` module. This module bypasses all compositional reasoning and directly predicts an answer from the input contextualized representations. This has potential to improve performance, in cases where a question describes reasoning that cannot be captured by pre-defined modules, in which case the program can be the `extract-answer` module only. However, introducing `extract-answer` adversely affects interpretability and learning of other modules, specifically in the absence of gold programs. First, `extract-answer` does not provide any interpretability. Second, whenever the parser predicts the `extract-answer` module, the parameters of the more interpretable modules are not trained. Moreover, the parameters of the encoder are trained to perform reasoning *internally* in a non-interpretable manner. We study the interpretability vs. performance trade-off by training Text-NMN with and without `extract-answer`.

Second, consider the program `find-max-num(find[touchdown])` that aims to find the *longest touchdown*. `find-max-num`

should sort spans by their value and return the maximal one; if we remove `find-max-num`, the program would reduce to `find[touchdown]`, and the `find` module would have to select the longest touchdown rather than all touchdowns, following the true denotation. More generally, omitting atomic reasoning modules pushes other modules to compensate and perform complex tasks that were not intended for them, hurting faithfulness. To study this, we train Text-NMN by removing sorting and comparison modules (e.g., `find-max-num` and `num-compare`), and evaluate how this affects module-wise interpretability.

## 4.2 Supervising module output

As explained, given end-task supervision only, modules may not act as intended, since their parameters are only trained for minimizing the end-task loss. Thus, a straightforward way to improve interpretability is to train modules with additional atomic-task supervision.

**Visual reasoning** For Visual-NMN, we pre-train `find` and `filter` modules with explicit intermediate supervision, obtained from the GQA balanced dataset (Hudson and Manning, 2019). Note that this supervision is used only during pre-training – we do not assume we have full-supervision for the actual task at hand. GQA questions are annotated by gold programs; we focus on “exist” questions that use `find` and `filter` modules only, such as “Are there any red cars?”.

Given gold annotations from Visual Genome (Krishna et al., 2017), we can compute a label for each of the bounding boxes proposed by Faster-RCNN. We label a proposed bounding box as ‘positive’ if its IOU with a gold bounding box is  $> 0.75$ , and ‘negative’ if it is  $< 0.25$ . We then train on GQA examples, minimizing both the usual denotation loss, as well as an auxiliary loss for each instance of `find` and `filter`, which is binary cross entropy for the labeled boxes. This loss rewards high probabilities for ‘positive’ bounding boxes and low probabilities for ‘negative’ ones.

**Textual reasoning** Prior work (Gupta et al., 2020) proposed heuristic methods to extract supervision for the `find-num` and `find-date` modules in DROP. On top of the end-to-end objective, they use an auxiliary objective that encourages these modules to output the “gold” numbers and dates according to the heuristic supervision. They show that supervising intermediate module outputs helps

improve model performance. In this work, we evaluate the effect of such supervision on the faithfulness of both the supervised modules, as well as other modules that are trained jointly.

## 4.3 Decontextualized word representations

The goal of decomposing reasoning into multiple steps, each focusing on different parts of the utterance, is at odds with the widespread use of contextualized representations such as BERT or LXMERT. While the *utterance attention* is meant to capture information only from tokens relevant for the module’s reasoning, contextualized token representations carry global information. For example, consider the program `filter[red](find[car])` for the phrase *red car*. Even if `find` attends only to the token *car*, its representation might also express the attribute *red*, so `find` might learn to find just *red cars*, rather than all *cars*, rendering the `filter` module useless, and harming faithfulness. To avoid such contextualization in Visual-NMN, we zero out the representations of tokens that are unattended, thus the input to the module is computed (with LXMERT) from the remaining tokens only.

# 5 Experiments

We first introduce the datasets used and the experimental setup for measuring faithfulness (§ 5.1). We demonstrate that training NMNs using end-task supervision only does not yield module-wise faithfulness both for visual and textual reasoning. We then show that the methods from §4 are crucial for achieving faithfulness and how different design choices affect it (§ 5.2). Finally, we qualitatively show examples of improved faithfulness and analyze possible reasons for errors (§ 5.3).

## 5.1 Experimental setup

Please see Appendix C for further detail about the experimental setups.

**Visual reasoning** We automatically generate gold program annotations for 26,311 training set examples and for 5,772 development set examples from NLVR2. The input to this generation process is the set of crowdsourced question decompositions from the BREAK dataset (Wolfson et al., 2020). See Appendix C.1 for details. For module-wise faithfulness evaluation, 536 examples from the development set were annotated with the gold output for each module by experts.

Model	Performance (Accuracy)	Overall Faithful. ( $\uparrow$ )			Module-wise Faithfulness $F_1(\uparrow)$			
		Prec.	Rec.	$F_1$	find	filter	with-relation	relocate
LXMERT	<b>71.7</b>							
Upper Bound		1	0.84	0.89	0.89	0.92	0.95	0.75
NMN w/ Layer-count	71.2	0.39	0.39	0.11	0.12	0.20	0.37	<b>0.27</b>
NMN w/ Sum-count	68.4	<b>0.49</b>	0.31	0.28	0.31	0.32	0.44	0.26
NMN w/ Graph-count	69.6	0.37	0.39	0.28	0.31	0.29	0.37	0.19
NMN w/ Graph-count + decont.	67.3	0.29	0.51	0.33	0.38	0.30	0.36	0.13
NMN w/ Graph-count + pretraining	69.6	0.44	0.49	0.36	0.39	0.34	0.42	0.21
NMN w/ Graph-count + decont. + pretraining	68.7	0.42	<b>0.66</b>	<b>0.47</b>	<b>0.52</b>	<b>0.41</b>	<b>0.47</b>	0.21

Table 1: Faithfulness and accuracy on NLVR2. “decont.” refers to decontextualized word representations. Precision, recall, and  $F_1$  are averages across examples, and thus  $F_1$  is **not** the harmonic mean of the corresponding precision and recall.

Model	Performance ( $F_1$ Score)	Overall Faithful. (cross-entropy* $\downarrow$ )	Module-wise Faithfulness* ( $\downarrow$ )				
			find	filter	relocate	min-max <sup>†</sup>	find-arg <sup>†</sup>
Text-NMN w/o prog-sup							
w/ extract-answer	63.5	9.5	13.3	9.5	3.5	2.6	9.9
w/o extract-answer	60.8	<b>6.9</b>	8.1	7.3	1.3	1.7	8.5
Text-NMN w/ prog-sup							
no auxiliary sup	65.3	11.2	13.7	16.9	1.5	2.2	13.0
w/o sorting & comparison	63.8	8.4	9.6	11.1	1.6	1.3	10.6
w/ module-output-sup	<b>65.7</b>	<b>6.5</b>	<b>7.6</b>	<b>10.7</b>	<b>1.3</b>	<b>1.2</b>	<b>7.6</b>

Table 2: Faithfulness and performance scores for various NMNs on DROP. \*lower is better. <sup>†</sup>min-max is average faithfulness of find-min-num and find-max-num; find-arg of find-num and find-date.

**Textual reasoning** We train Text-NMN on DROP, which is augmented with program supervision for 4,000 training questions collected heuristically as described in Gupta et al. (2020). The model is evaluated on the complete development set of DROP which does not contain any program supervision. Module-wise faithfulness is measured on 215 manually-labeled questions from the development set, which are annotated with gold programs and module outputs (passage spans).

## 5.2 Faithfulness evaluation

**Visual reasoning** Results are seen in Table 1. Accuracy for LXMERT, when trained and evaluated on the same subset of data, is 71.7%; slightly higher than NMNs, but without providing evidence for the compositional structure of the problem.

For faithfulness, we measure an upper-bound on the faithfulness score. Recall that this score measures the similarity between module outputs and annotated outputs. Since module outputs are constrained by the bounding boxes proposed by Faster-RCNN (§2.1), while annotated boxes are not, perfect faithfulness could only be achieved by a model if there are suitable bounding boxes. *Upper Bound* shows the maximal faithfulness score

conditioned on the proposed bounding boxes.

We now compare the performance and faithfulness scores of the different components. When training our NMN with the most flexible count module, (*NMN w/ Layer-count*), an accuracy of 71.2% is achieved, a slight drop compared to LXMERT but with low faithfulness scores. Using *Sum-count* drops about 3% of performance, but increases faithfulness. Using *Graph-count* increases accuracy while faithfulness remains similar.

Next, we analyze the effect of decontextualized word representations (abbreviated “decont.”) and pre-training. First, we observe that *NMN w/ Graph-count + decont.* increases faithfulness score to 0.33  $F_1$  at the expense of accuracy, which drops to 67.3%. Pre-training (*NMN w/ Graph-count + pre-training*) achieves higher faithfulness scores with a higher accuracy of 69.6%. Combining the two achieves the best faithfulness (0.47  $F_1$ ) with a minimal accuracy drop. We perform a paired permutation test to compare *NMN w/ Graph-count + decont. + pretraining* with *NMN w/ Layer-count* and find that the difference in  $F_1$  is statistically significant ( $p < 0.001$ ). Please see Appendix D.1 for further details.

**Textual reasoning** As seen in Table 2, when trained on DROP using question-program supervision, the model achieves 65.3  $F_1$  performance and a faithfulness score of 11.2. When adding supervision for intermediate modules (§4.2), we find that the module-wise faithfulness score improves to 6.5. Similar to Visual-NMN, this shows that supervising intermediate modules in a program leads to better faithfulness.

To analyze how choice of modules affects faithfulness, we train without sorting and comparison modules (`find-max-num`, `num-compare`, etc.). We find that while performance drops slightly, faithfulness deteriorates significantly to 8.4, showing that modules that perform atomic reasoning are crucial for faithfulness. When trained without program supervision, removing `extract-answer` improves faithfulness (9.5  $\rightarrow$  6.9) but at the cost of performance (63.5  $\rightarrow$  60.8  $F_1$ ). This shows that such a black-box module encourages reasoning in an opaque manner, but can improve performance by overcoming the limitations of pre-defined modules. All improvements in faithfulness are significant as measured using paired permutation tests ( $p < 0.001$ ).

**Generalization** A natural question is whether models that are more faithful also generalize better. We conducted a few experiments to see whether this is true for our models. For NLVR2, we performed (1) an experiment in which programs in training have length at most 7, and programs at test time have length greater than 7, (2) an experiment in which programs in training have at most 1 `filter` module and programs at test time have at least 2 `filter` modules, and (3) an experiment in which programs in training do not have both `filter` and `with-relation` modules in the same program, while each program in test has both modules. We compared three of our models – *NMN w/ Layer-count*, *NMN w/ Sum-count*, and *NMN w/ Graph-count + decont. + pretraining*. We did not observe that faithful models generalize better (in fact, the most unfaithful model tended to achieve the best generalization).

To measure if faithful model behavior leads to better generalization in Text-NMN we conducted the following experiment. We selected the subset of data for which we have gold programs and split the data such that questions that require maximum and greater-than operations are present in the training data while questions that require com-

puting minimum and less-than are in the test data. We train and test our model by providing gold-programs under two conditions, in the presence and absence of additional module supervision. We find that providing auxiliary module supervision (that leads to better module faithfulness; see above) also greatly helps in model generalization (performance increases from 32.3  $F_1 \rightarrow$  78.3  $F_1$ ).

### 5.3 Qualitative analysis

**Model comparisons** We analyze outputs of different modules in Figure 3. Figures 3a, 3b show the output of `find[llamas]` when trained with contextualized and decontextualized word representations. With contextualized representations (3a), the `find` fails to select any of the *llamas*, presumably because it can observe the word *eating*, thus effectively searching for *eating llamas*, which are not in the image. Conversely, the decontextualized model correctly selects the boxes. Figure 3c shows that `find` outputs meaningless probabilities for most of the bounding boxes when trained with *Layer-count*, yet the count module produces the correct value (three). Figure 3d shows that `find` fails to predict all relevant spans when trained without sorting modules in Text-NMN.

**Error analysis** We analyze cases where outputs were unfaithful. First, for visual reasoning, we notice that faithfulness scores are lower for long-tail objects. For example, for *dogs*, a frequent noun in NLVR2, the execution of `find[dogs]` yields an average faithfulness score of 0.71, while items such as *roll of toilet paper*, *barbell* and *safety pin* receive lower scores (0.22, 0.29 and 0.05 respectively; example for a failure case for *safety pin* in Fig. 3e). In addition, some objects are harder to annotate with a box (*water*, *grass*, *ground*) and therefore receive low scores. The issue of small objects can also explain the low scores of `relocate`. In the gold box annotations used for evaluation, the average areas for `find`, `filter`, `with-relation`, and `relocate` (as a fraction of the total image area) are 0.19, 0.19, 0.15, and 0.07, respectively. Evidently, `relocate` is executed with small objects that are harder to annotate (*tongue*, *spots*, *top of*), and indeed the upper-bound and model scores for `relocate` are lowest among the module types.

## 6 Related Work

NMNs were originally introduced for visual question answering and applied to datasets with syn-



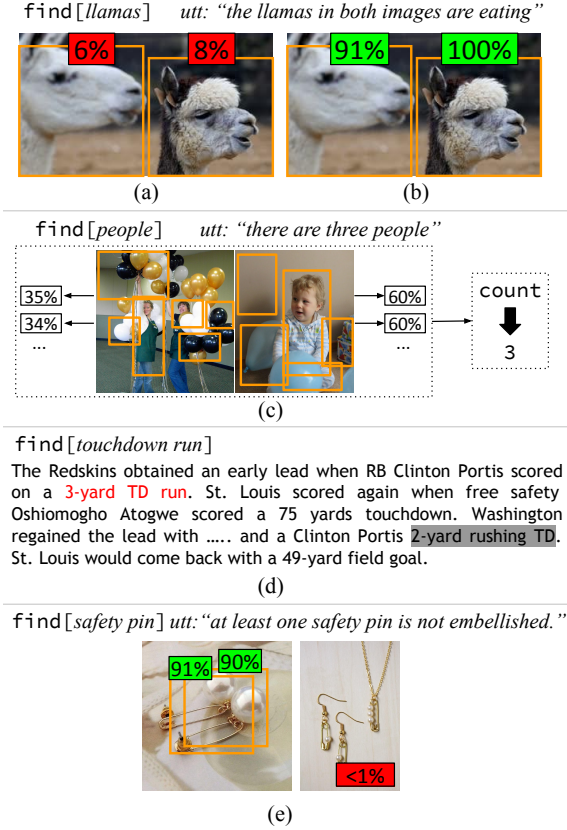


Figure 3: Comparison of module outputs between NMN versions: (a) Visual-NMN with contextualized representations, (b) Visual-NMN with decontextualized representations, (c) model using a parameter-rich count layer (Layer-Count), (d) Text-NMN trained without sorting module produces an incorrect `find` output (misses *2-yard rushing TD*), and (e) Visual-NMN failure case with a rare object (of w/ *Graph-count + decont. + pretraining*)

thetic language and images as well as VQA (Antol et al., 2015), whose questions require few reasoning steps (Andreas et al., 2016; Hu et al., 2017; Yang et al., 2018b). In such prior work, module-wise faithfulness was mostly assessed via qualitative analysis of a few examples (Jiang and Bansal, 2019; Gupta et al., 2020). Yang et al. (2018b) did an evaluation where humans rated the clarity of the reasoning process and also tested whether humans could detect model failures based on module outputs. In contrast, we quantitatively measure each module’s predicted output against the annotated gold outputs.

A related systematic evaluation of interpretability in VQA was conducted by Trott et al. (2018). They evaluated the interpretability of their VQA counting model, where the interpretability score is given by the semantic similarity between the gold

label for a bounding box and the relevant word(s) in the question. However, they studied only counting questions, which were also far less compositional than those in NLVR2 and DROP.

Similar to the gold module output annotations that we provide and evaluate against, HotPotQA (Yang et al., 2018a) and CoQA (Reddy et al., 2019) datasets include supporting facts or rationales for the answers to their questions, which can be used for both supervision and evaluation.

In concurrent work, Jacovi and Goldberg (2020) recommend studying faithfulness on a scale rather than as a binary concept. Our evaluation method can be viewed as one example of this approach.

## 7 Conclusion

We introduce the concept of *module-wise faithfulness*, a systematic evaluation of faithfulness in neural module networks (NMNs) for visual and textual reasoning. We show that naïve training of NMNs does not produce faithful modules and propose several techniques to improve module-wise faithfulness in NMNs. We show how our approach leads to much higher module-wise faithfulness at a low cost to performance. We encourage future work to judge model interpretability using the proposed evaluation and publicly published annotations, and explore techniques for improving faithfulness and interpretability in compositional models.

## Acknowledgements

We thank members of UCI NLP, TAU NLP, and the AllenNLP teams as well as Daniel Khashabi for comments on earlier drafts of this paper. We also thank the anonymous reviewers for their comments. This research was partially supported by The Yandex Initiative for Machine Learning, the European Research Council (ERC) under the European Union Horizons 2020 research and innovation programme (grant ERC DELPHI 802800), funding by the ONR under Contract No. N00014-19-1-2620, and by sponsorship from the LwLL DARPA program under Contract No. FA8750-19-2-0201. This work was completed in partial fulfillment for the Ph.D degree of Ben Bogin.

## References

- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT*, pages 1545–1554.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.
- Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural Module Networks for Reasoning over Text. In *International Conference on Learning Representations (ICLR)*.
- Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.
- Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. 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 1596–1606, Hong Kong, China. Association for Computational Linguistics.
- Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. 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.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 2020 Conference of the Association for Computational Linguistics*.
- Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop reasoning. 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 4473–4483, Hong Kong, China. Association for Computational Linguistics.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.
- 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.
- Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526.
- Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*.
- Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.
- E.W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley.
- Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS’15*, pages 91–99, Cambridge, MA, USA. MIT Press.

- Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *IJCAI*.
- Howard Seltman. 2018. [Approximations for mean and variance of a ratio](#).
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223.
- Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.
- Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of NAACL-HLT*, pages 641–651.
- 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 5099–5110, Hong Kong, China. Association for Computational Linguistics.
- Alexander Trott, Caiming Xiong, and Richard Socher. 2018. [Interpretable counting for visual question answering](#). In *International Conference on Learning Representations*.
- 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Dan Ventura. 2007. [CS478 Paired Permutation Test Overview](#). Accessed April 29, 2020.
- Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). 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 11–20, Hong Kong, China. Association for Computational Linguistics.
- Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. [Break it down: A question understanding benchmark](#). *Transactions of the Association for Computational Linguistics*, 8:183–198.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018a. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics.
- Yan Zhang, Jonathon Hare, and Adam Prgel-Bennett. 2018. [Learning to count objects in natural images for visual question answering](#). In *International Conference on Learning Representations*.

## A Modules

We list all modules for Visual-NMN in Table 3.

For Text-NMN, as mentioned, we use all modules are described in Gupta et al. (2020). In this work, we introduce the (a) `addition` and `subtraction` modules that take as input two distributions over numbers mentioned in the passage and produce a distribution over all possible addition and subtraction values possible. The output distribution here is the expected distribution for the random variable  $Z = X + Y$  (for `addition`), and (b) `extract-answer` that produces two distributions over the passage tokens denoting the probabilities for the start and end of the answer span. This distribution is computed by mapping the passage token representations using a simple MLP and softmax operation.

## B Measuring Faithfulness in Visual-NMN

### B.1 Numerators of Precision and Recall

As stated in Section 3.1, for a given module type and a given example, precision is defined as the number of matched proposed bounding boxes divided by the number of proposed bounding boxes to which the module assigns a probability more than 0.5. Recall is defined as the number of matched annotated bounding boxes divided by the number of annotated bounding boxes. Therefore, the numerators of the precision and the recall need not be equal. In short, the reason for the discrepancy is that there is no one-to-one alignment between annotated and proposed bounding boxes. To further illustrate why we chose not to have a common numerator, we will consider two sensible choices for this shared numerator and explain the issues with them.

One choice for the common numerator is the number of matched proposed bounding boxes. If we were to keep the denominator of the recall the same, then the recall would be defined as the number of matched proposed bounding boxes divided by the number of annotated bounding boxes. Consider an example in which there is a single annotated bounding box that is aligned with five proposed bounding boxes. When this definition of recall is applied to this example, the numerator would exceed the denominator. Another choice would be to set the denominator to be the number of proposed bounding boxes that are aligned with

some annotated bounding box. In the example, this approach would penalize a module that gives high probability to only one of the five aligned proposed bounding boxes. However, it is not clear that a module giving high probability to all five proposed boxes is more faithful than a module giving high probability to only one bounding box (e.g. perhaps one proposed box has a much higher IOU with the annotated box than the other proposed boxes). Hence, this choice for the numerator does not make sense.

Another choice for the common numerator is the number of matched annotated bounding boxes. If we were to keep the denominator of the precision the same, then the precision would be defined as the number of matched annotated bounding boxes divided by the number of proposed bounding boxes to which the module assigns probability more than 0.5. Note that since a single proposed bounding box can align with multiple annotated bounding boxes, it is possible for the numerator to exceed the denominator.

Thus, these two choices for a common numerator have issues, and we avoid these issues by defining the numerators of precision and recall separately.

### B.2 Averaging Faithfulness Scores

The method described in Section 3.1 computes a precision, recall, and  $F_1$  score for each example for every module type occurring in that example. The faithfulness scores reported in Table 1 are averages across examples. We also considered two other ways of aggregating scores across examples:

1. Cumulative P/R/ $F_1$ : For each module type, we compute a single cumulative precision and recall across all examples. We then compute the dataset-wide  $F_1$  score as the harmonic mean of the precision and the recall. The results using this method are in Table 4. There are some differences between these results and those in Table 1, e.g. in these results, *NMN w/ Graph-count + decont. + pretraining* has the highest faithfulness score for every module type, including *relocate*.
2. Average over module occurrences: For each module type, for each occurrence of the module we compute a precision and recall and compute  $F_1$  as the harmonic mean of precision and recall. Then for each module type, we compute the overall precision as the average precision across module occurrences and



similarly compute the overall recall and  $F_1$ . Note that a module can occur multiple times in a single program and that each image is considered a separate occurrence. The results using this method are in Table 5. Again, there are some differences between these results and those in Table 1, e.g. *NMN w/ Sum-count* has a slightly higher score for *with-relation* than *NMN w/ Graph-count + decont. + pre-training*.

With both of these alternative score aggregation methods, we still obtained  $p < 0.001$  in our significance tests.

We also noticed qualitatively that the metric can penalize modules that assign high probability to proposed bounding boxes that have a relatively high IOU that does not quite pass the IOU threshold of 0.5. In such cases, while it may not make sense to give the model credit in its recall score, it also may not make sense to penalize the model in its precision score. Consequently, we also performed an evaluation in which for the precision calculation we set a separate “negative” IOU threshold of  $10^{-8}$  (effectively 0) and only penalized modules for high probabilities assigned to proposed boxes whose IOU is below this threshold. The results computed with example-wise averaging are provided in Table 6.

## C Details about Experiments

**Visual Reasoning** We use the published pre-trained weights and the same training configuration of LXMERT (Tan and Bansal, 2019), with 36 bounding boxes proposed per image. Due to memory constraints, we restrict training data to examples having a gold program with at most 13 modules.

### C.1 Program Annotations

We generated program annotations for NLVR2 by automatically canonicalizing its question decompositions in the BREAK dataset (Wolfson et al., 2020). Decompositions were originally annotated by Amazon Mechanical Turk workers. For each utterance, the workers were asked to produce the correct decomposition and an *utterance attention* for each operator (module), whenever relevant.

**Limitations of Program Annotations** Though our annotations for gold programs in NLVR2 are largely correct, we find that there are some examples for which the programs are unnecessarily

Program
exist
in_right_image
with-relation [with]
filter [brown]
find [dog]
filter[exposed]
relocate[tongue]
filter[brown]
find [dog]

Figure 4: An example of a gold program for NLVR2 that is unnecessarily complicated.

complicated. For instance, for the sentence “the right image contains a brown dog with its tongue extended.” the gold program is shown in Figure 4. This program could be simplified by replacing the *with-relation* with the second argument of *with-relation*. Programs like this make learning more difficult for the NMNs since they use modules (in this case, *with-relation*) in degenerate ways. There are also several sentences that are beyond the scope of our language, e.g. comparisons such as “the right image shows exactly two virtually identical trifle desserts.”

## D Significance tests

### D.1 Visual Reasoning

We perform a paired permutation test to test the hypothesis  $H_0$ : *NMN w/ Graph-count + decont. + pretraining* has the same inherent faithfulness as *NMN w/ Layer-count*. We follow the procedure described by Ventura (2007), which is similar to tests described by Yeh (2000) and Noreen (1989). Specifically, we perform  $N_{total} = 100,000$  trials in which we do the following. For every example, with probability 1/2 we swap the  $F_1$  scores obtained by the two models for that example. Then we check whether the difference in the aggregated  $F_1$  scores for the two models is at least as extreme as the original difference in the aggregated  $F_1$  scores of the two models. The p-value is given by  $N_{exceed}/N_{total}$ , where  $N_{exceed}$  is the number of trials in which the new difference is at least as extreme as the original difference.

Module	Output	Implementation
<code>find</code> [ $q_{att}$ ]	$p$	$W_1^T([x; v]) + b_1$
<code>filter</code> [ $q_{att}$ ]( $p$ )	$p$	$p \odot (W_1^T([x; v]) + b_1)$
<code>with-relation</code> [ $q_{att}$ ]( $p_1, p_2$ )	$p$	$\max(p_2)p_1 \odot \text{MLP}([x; v_1; v_2])$
<code>project</code> [ $q_{att}$ ]( $p$ )	$p$	$\max(p)\text{find}(q_{att}) \odot \text{MLP}([W_2; v_1; v_2])$
<code>count</code> ( $p$ )	N	$\text{number}(\sum(p), \sigma^2)$
<code>exist</code> ( $p$ )	B	<code>greater-equal</code> ( $p, 1$ )
<code>greater-equal</code> ( $a : N, b : N$ )	B	<code>greater</code> ( $a, b$ ) + <code>equal</code> ( $a, b$ )
<code>less-equal</code> ( $a : N, b : N$ )	B	<code>less</code> ( $a, b$ ) + <code>equal</code> ( $a, b$ )
<code>equal</code> ( $a : N, b : N$ )	B	$\sum_{k=0}^K \Pr[a = k] \Pr[b = k]$
<code>less</code> ( $a : N, b : N$ )	B	$\sum_{k=0}^K \Pr[a = k] \Pr[b > k]$
<code>greater</code> ( $a : N, b : N$ )	B	$\sum_{k=0}^K \Pr[a = k] \Pr[b < k]$
<code>and</code> ( $a : B, b : B$ )	B	$a * b$
<code>or</code> ( $a : B, b : B$ )	B	$a + b - a * b$
<code>number</code> ( $m : F, v : F$ )	N	<code>Normal</code> ( $\text{mean} = m, \text{var} = v$ )
<code>sum</code> ( $a : N, b : N$ )	N	$\text{number}(a_{\text{mean}} + b_{\text{mean}}, a_{\text{var}} + b_{\text{var}})$
<code>difference</code> ( $a : N, b : N$ )	N	$\text{number}(a_{\text{mean}} - b_{\text{mean}}, a_{\text{var}} + b_{\text{var}})$
<code>division</code> ( $a : N, b : N$ )	N	$\text{number}\left(\frac{a_{\text{mean}}}{b_{\text{mean}}} + \frac{b_{\text{var}}a_{\text{mean}}}{b_{\text{mean}}^3}, \frac{a_{\text{mean}}^2}{b_{\text{mean}}^2}\left(\frac{a_{\text{var}}}{a_{\text{mean}}^2} + \frac{b_{\text{var}}}{b_{\text{mean}}^2}\right)\right)$
<code>intersect</code> ( $p_1, p_2$ )	$p$	$p_1 \cdot p_2$
<code>discard</code> ( $p_1, p_2$ )	$p$	$\max(p_1 - p_2, 0)$
<code>in-left-image</code> ( $p$ )	$p$	$p$ s.t. probabilities for right image are 0
<code>in-right-image</code> ( $p$ )	$p$	$p$ s.t. probabilities for left image are 0
<code>in-at-least-one-image</code>	B	macro (see caption)
<code>in-each-image</code>	B	macro (see caption)
<code>in-one-other-image</code>	B	macro (see caption)

Table 3: Implementations of modules for NLVR2 NMN. First five contain parameters, the rest are deterministic. The implementation of `count` shown here is the Sum-count version; please see Section 4 for a description of other count module varieties and a discussion of their differences. ‘B’ denotes the Boolean type, which is a probability value ( $[0..1]$ ). ‘N’ denotes the Number type which is a probability distribution.  $K = 72$  is the maximum count value supported by our model. To obtain probabilities, we first convert each Normal random variable  $X$  to a categorical distribution over  $\{0, 1, \dots, K\}$  by setting  $\Pr[X = k] = \Phi(k + 0.5) - \Phi(k - 0.5)$  if  $k \in \{1, 2, \dots, K - 1\}$ . We set  $\Pr[X = 0] = \Phi(0.5)$  and  $\Pr[X = K] = 1 - \Phi(K - 0.5)$ . Here  $\Phi(\cdot)$  denotes the cumulative distribution function of the Normal distribution.  $W_1, W_2$  are weight vectors with shapes  $2h \times 1$  and  $h \times 1$ , respectively. Here  $h = 768$  is the size of LXMERT’s representations.  $b_1$  is a scalar weight. MLP denotes a two-layer neural network with a GeLU activation (Hendrycks and Gimpel, 2016) between layers.  $x$  denotes a question representation, and  $v_i$  denotes encodings of objects in the image.  $x$  and  $v_i$  have shape  $h \times |\mathcal{B}|$ , where  $|\mathcal{B}|$  is the number of proposals.  $p$  denotes a vector of probabilities for each proposal and has shape  $1 \times |\mathcal{B}|$ .  $\odot$  and  $;$  represent elementwise multiplication and matrix concatenation, respectively. The expressions for the mean and variance in the division module are based on the approximations in Seltman (2018). The macros execute a given program on the two input images. `in-at-least-one-image` macro returns true iff the program returns true when executed on at least one of the images. `in-each-image` returns true iff the program returns true when executed on both of the images. `in-one-other-image` takes two programs and returns true iff one program return true on left image and second program returns true on right image, or vice-versa.

Model	Performance (Accuracy)	Overall Faithful.(↑)			Module-wise Faithfulness(↑)			
		Prec.	Rec.	F1	find	filter	with-relation	relocate
LXMERT	<b>71.7</b>							
Upper Bound		1	0.63	0.77	0.78	0.79	0.73	0.71
NMN w/ Layer-count	71.2	0.069	0.29	0.11	0.13	0.09	0.07	0.05
NMN w/ Sum-count	68.4	0.25	0.18	0.21	0.23	0.20	0.16	0.05
NMN w/ Graph-count	69.6	0.20	0.22	0.21	0.24	0.19	0.17	0.04
NMN w/ Graph-count + decont.	67.3	0.21	0.29	0.24	0.28	0.22	0.19	0.04
NMN w/ Graph-count + pretraining	69.6	0.28	0.31	0.30	0.34	0.27	0.25	0.09
NMN w/ Graph-count + decont. + pretraining	68.7	<b>0.34</b>	<b>0.43</b>	<b>0.38</b>	<b>0.43</b>	<b>0.34</b>	<b>0.29</b>	<b>0.11</b>

Table 4: Faithfulness scores on NLVR2 using the cumulative precision/recall/F<sub>1</sub> evaluation.

Model	Performance (Accuracy)	Overall Faithful.(↑)			Module-wise Faithfulness(↑)			
		Prec.	Rec.	F1	find	filter	with-relation	relocate
LXMERT	<b>71.7</b>							
Upper Bound		1	0.91	0.92	0.90	0.95	0.96	0.82
NMN w/ Layer-count	71.2	0.67	0.64	0.39	0.21	0.50	0.61	<b>0.50</b>
NMN w/ Sum-count	68.4	<b>0.70</b>	0.59	0.48	0.38	0.53	<b>0.63</b>	0.49
NMN w/ Graph-count	69.6	0.55	0.64	0.43	0.36	0.47	0.54	0.41
NMN w/ Graph-count + decont.	67.3	0.47	0.70	0.45	0.42	0.47	0.55	0.33
NMN w/ Graph-count + pretraining	69.6	0.58	0.70	0.47	0.42	0.49	0.58	0.41
NMN w/ Graph-count + decont. + pretraining	68.7	0.58	<b>0.79</b>	<b>0.55</b>	<b>0.54</b>	<b>0.55</b>	0.62	0.43

Table 5: Faithfulness scores on NLVR2 using the average over module occurrences evaluation.

Model	Performance (Accuracy)	Overall Faithful.(↑)			Module-wise Faithfulness(↑)			
		Prec.	Rec.	F1	find	filter	with-relation	relocate
LXMERT	<b>71.7</b>							
Upper Bound		1	0.8377	0.89	0.89	0.92	0.95	0.75
NMN w/ Layer-count	71.2	0.59	0.39	0.25	0.31	0.28	0.45	0.30
NMN w/ Sum-count	68.4	<b>0.79</b>	0.31	0.34	0.38	0.36	0.48	0.28
NMN w/ Graph-count	69.6	0.68	0.39	0.38	0.43	0.36	0.44	0.22
NMN w/ Graph-count + decont.	67.3	0.62	0.51	0.47	0.53	0.39	0.43	0.16
NMN w/ Graph-count + pretraining	69.6	0.70	0.49	0.47	0.52	0.41	0.51	0.27
NMN w/ Graph-count + decont. + pretraining	68.7	0.71	<b>0.66</b>	<b>0.62</b>	<b>0.68</b>	<b>0.50</b>	<b>0.55</b>	<b>0.31</b>

Table 6: Faithfulness scores on NLVR2 using a negative IOU threshold of  $10^{-8}$  and example-wise averaging.