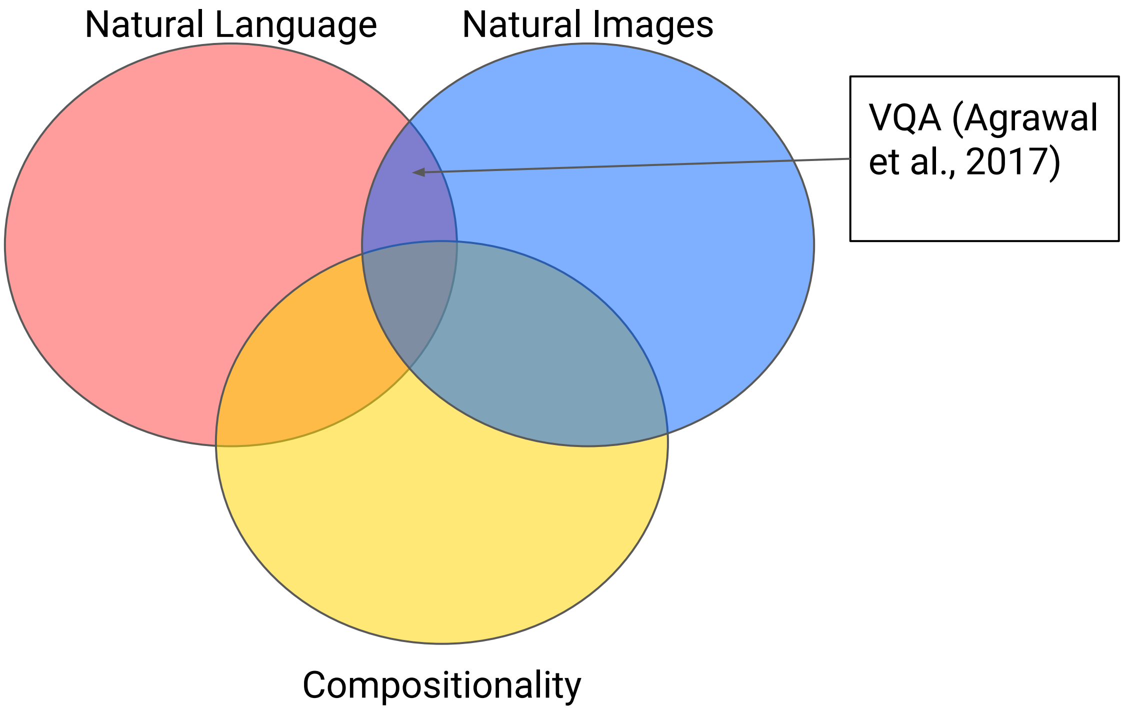
Compositional Visual Reasoning: Interpretability and Evaluation

Sanjay Subramanian, with many collaborators

AI2

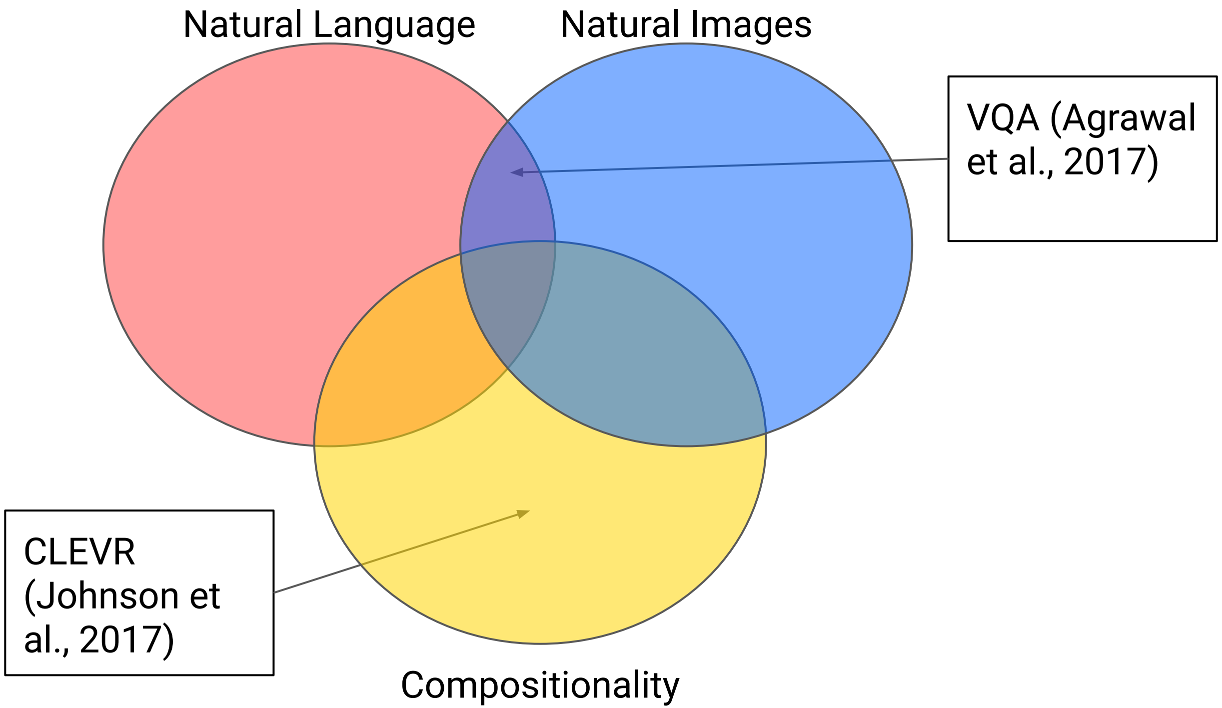
Recent developments in compositional visual reasoning

VQA datasets



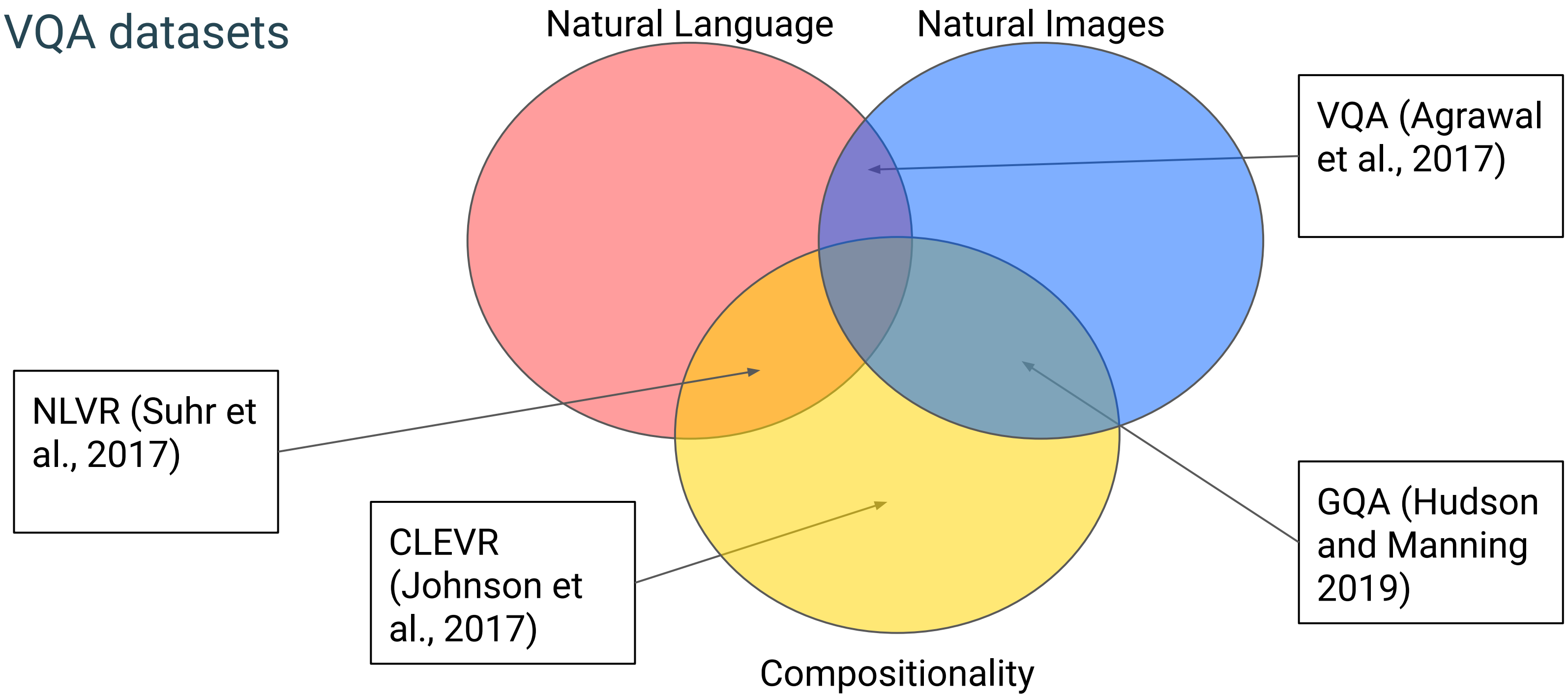
Recent developments in compositional visual reasoning

VQA datasets



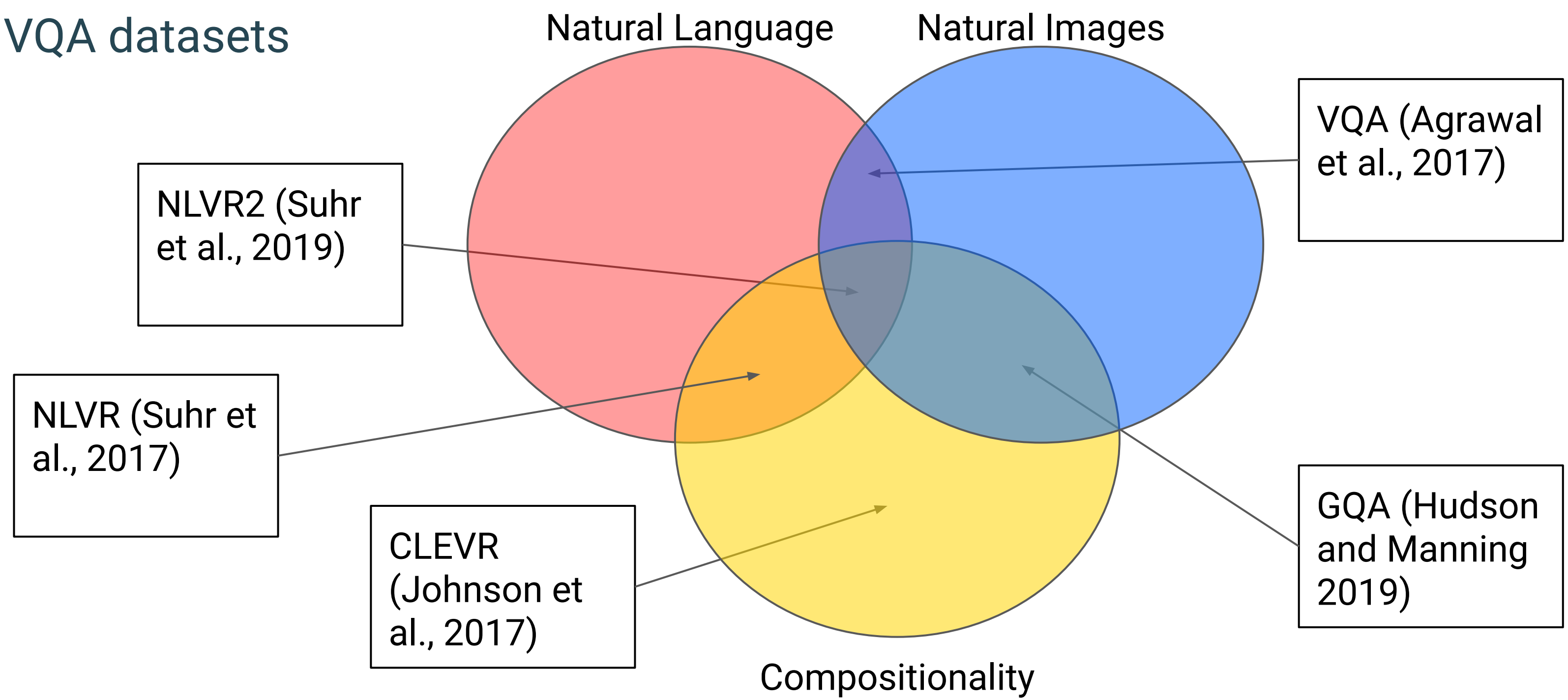
Recent developments in compositional visual reasoning

VQA datasets



Recent developments in compositional visual reasoning

VQA datasets



Recent developments in compositional visual reasoning

- Early VQA datasets were either simple and natural (e.g. VQA; Agrawal et al. 2017) or compositional and synthetic (e.g. CLEVR; Johnson et al. 2017)
- Recent compositional datasets:
  - NLVR2 (Suhr et al. 2019) -- two natural images paired with a sentence. True/false classification.
  - GQA (Hudson and Manning 2019) -- synthetic question with natural image. Classification and open-ended questions.

Recent developments in compositional visual reasoning

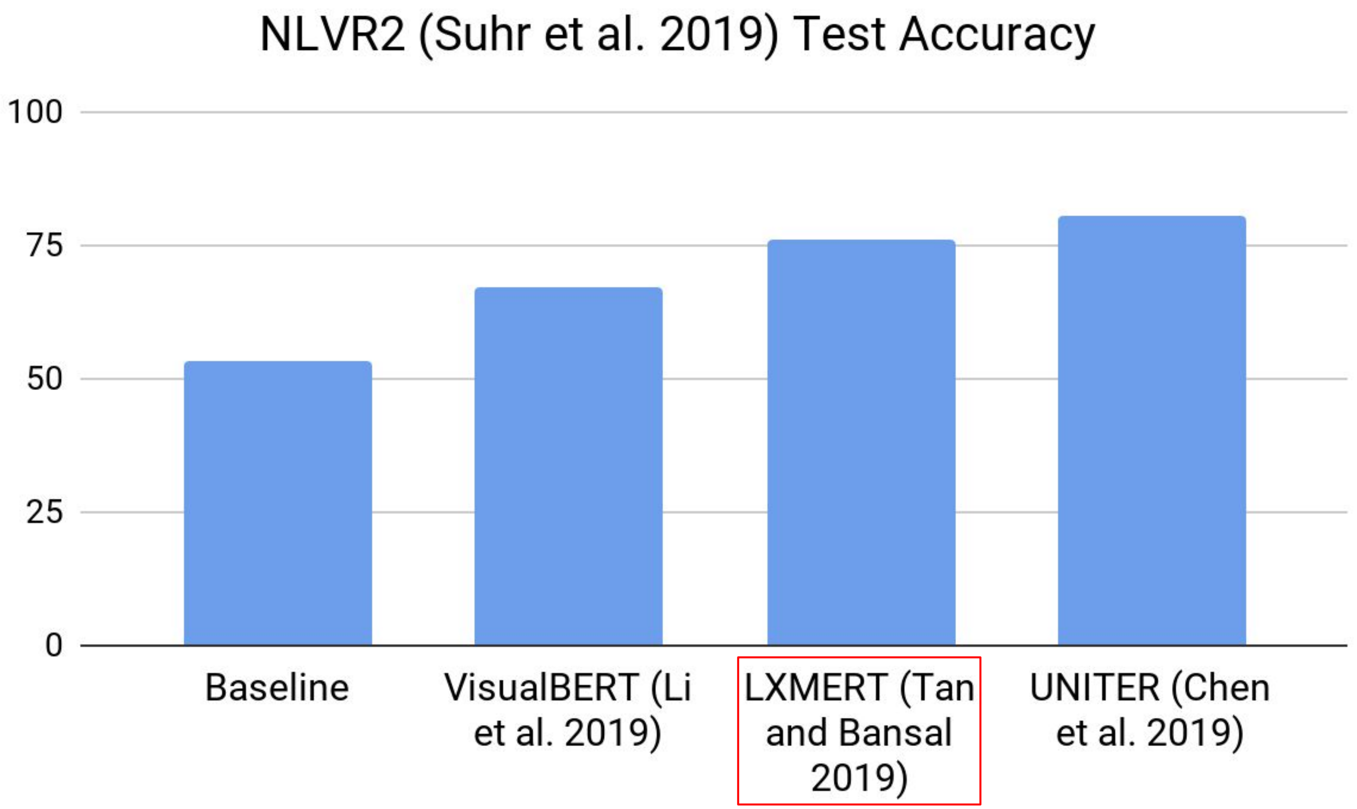
- Large-scale pre-trained transformers have been successful
- Example: LXMERT (Tan and Bansal, 2019)

LXMERT loss =

Masked language-modeling + Visual feature regression + Image-text matching + VQA

- Requires paired images and captions (COCO/Visual Genome) and VQA data
- SOTA on NLVR2, strong performance on GQA

Performance Gains from Pre-training



Issues raised by large pre-trained models

1. Interpretability: Can we make these models interpretable?
   - Unclear how to extract the steps of a vanilla Transformer
   - Particularly salient for compositional tasks
2. Evaluation: Are there shortcuts in these compositional datasets that enable models to perform well without going through the apparent reasoning steps?
   - Specifically: is object+attribute detection sufficient?

Obtaining Faithful Interpretations from Compositional Neural Networks

SS*, Ben Bogin*, Nitish Gupta*

Tomer Wolfson, Sameer Singh, Jonathan Berant, Matt Gardner

ACL 2020

AI2

Compositional reasoning

All dogs are black

Compositional reasoning

All dogs are black

LXMERT
(black-box neural network)

False

Tan and Bansal, EMNLP 2019

Compositional reasoning

All dogs are black

LXMERT
(black-box neural network)

False

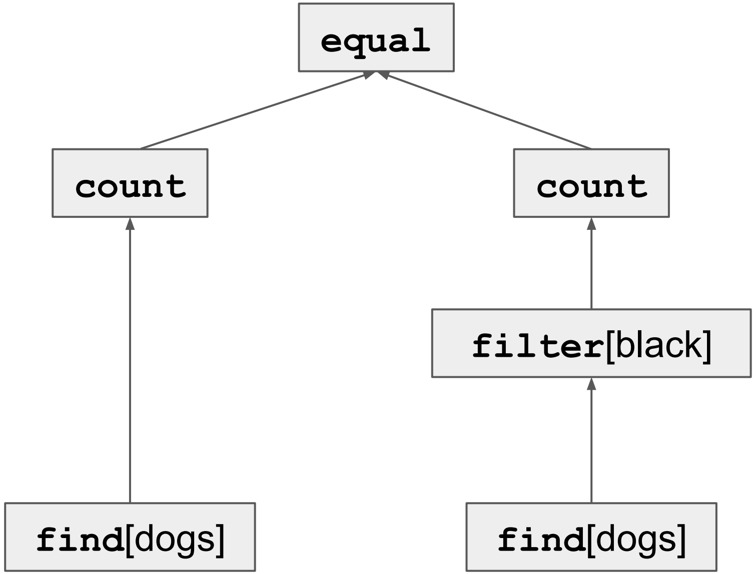
Tan and Bansal, EMNLP 2019

Not Interpretable

Neural Module Networks (NMN)

All dogs are black

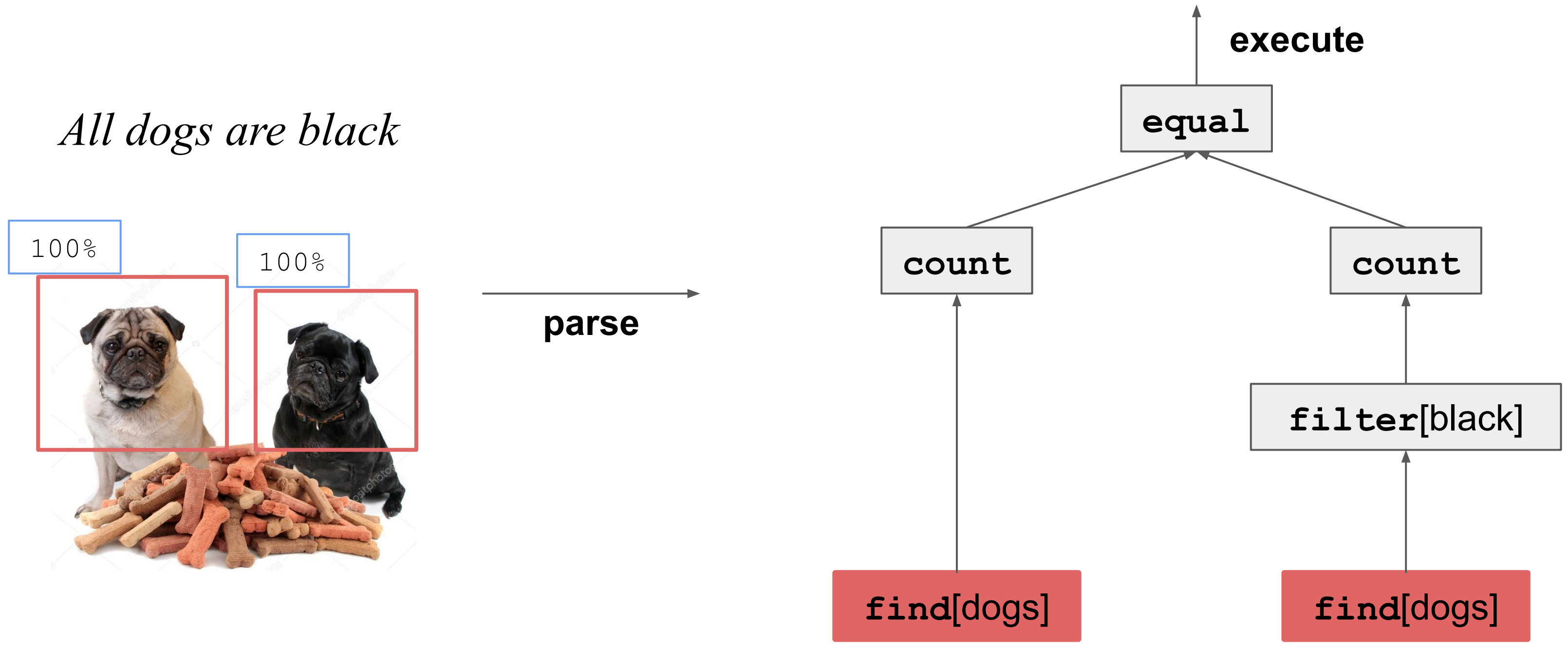
parse



Ai2

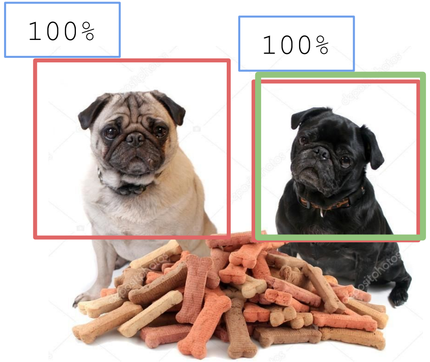
[Andreas et al., NAACL 2016]

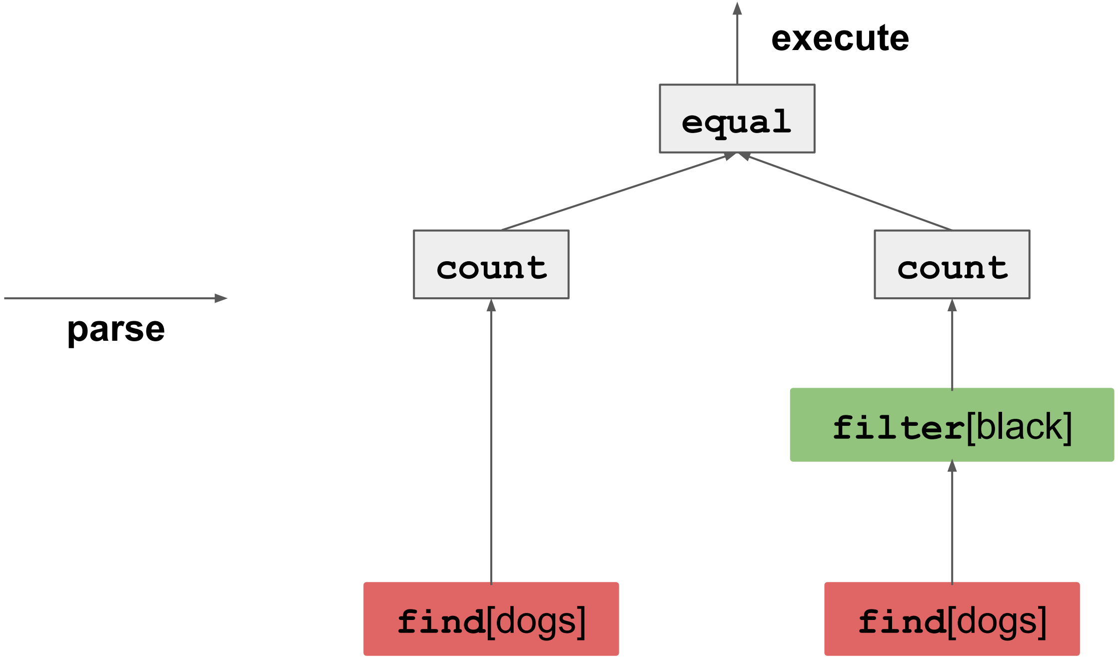
Neural Module Networks (NMN)



Neural Module Networks (NMN)

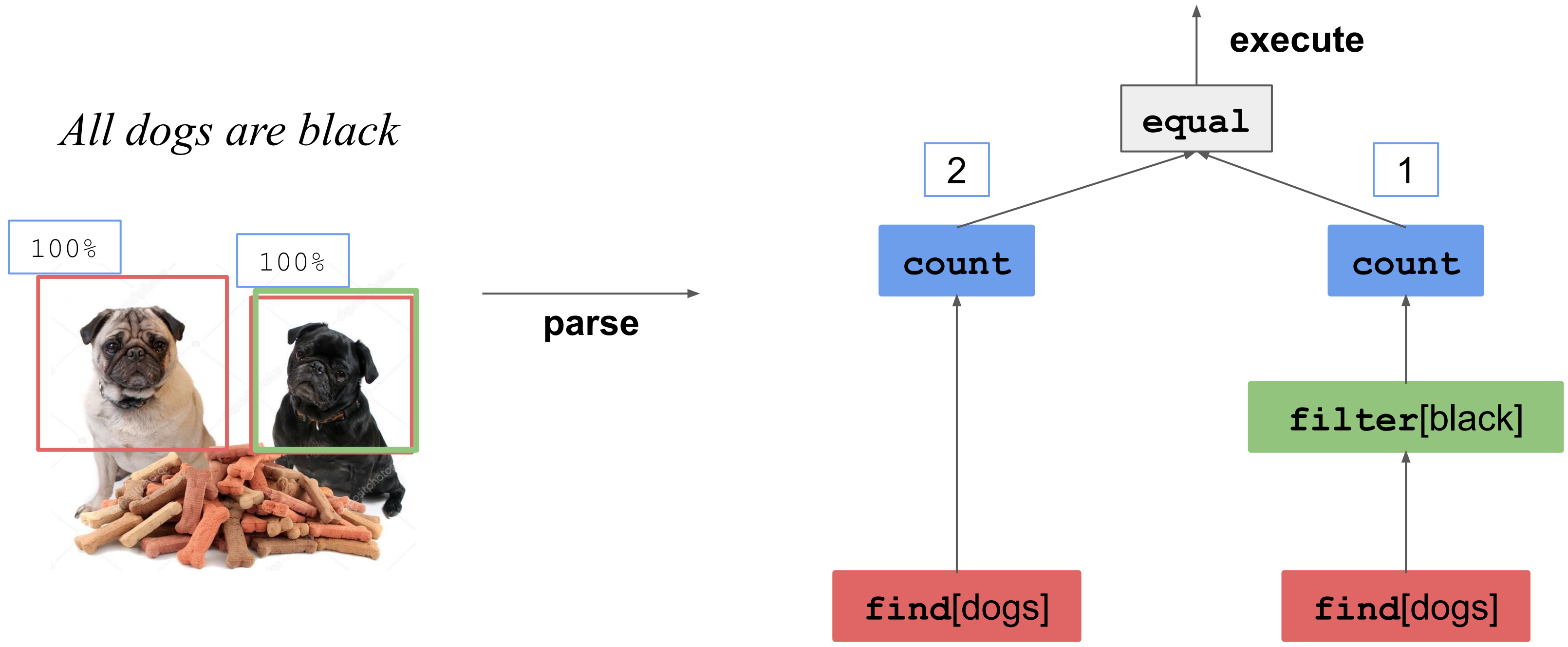
All dogs are black



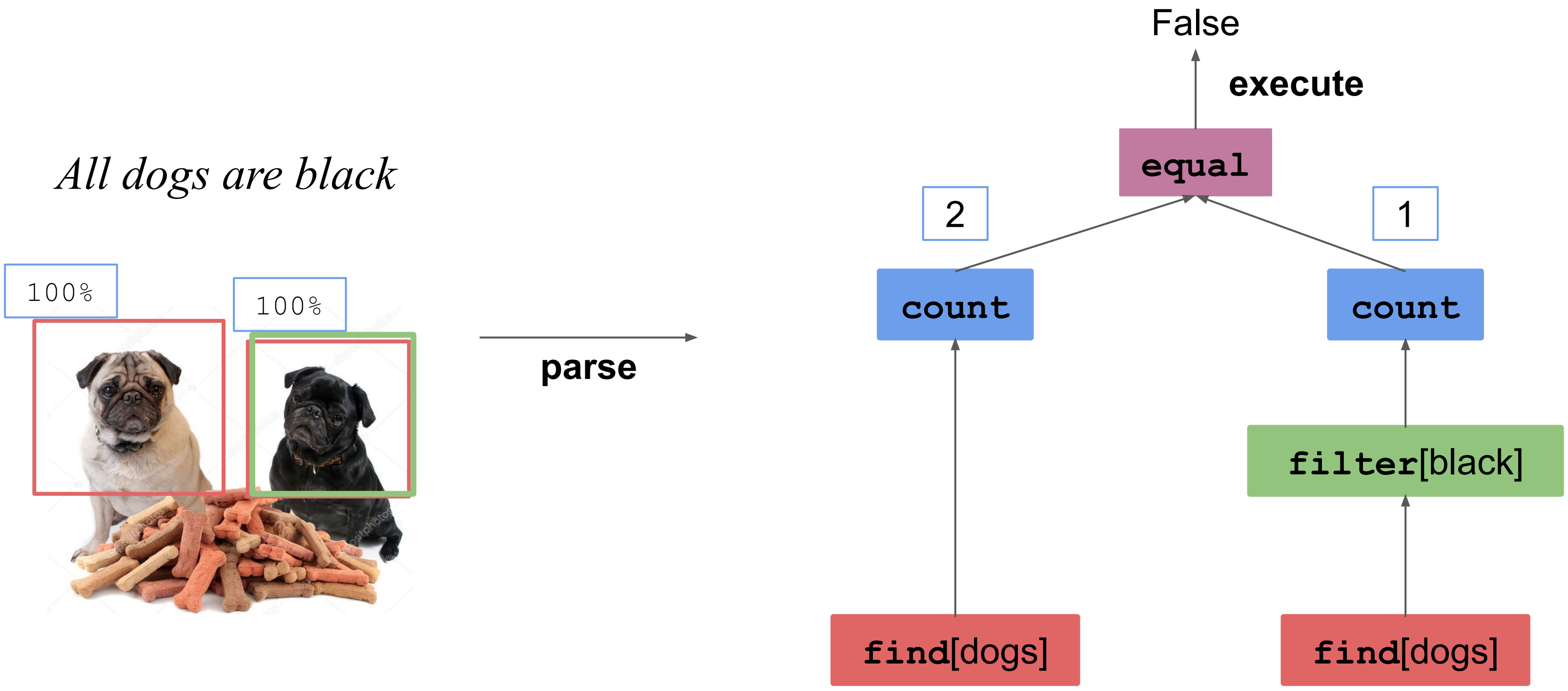


Ai2

Neural Module Networks (NMN)



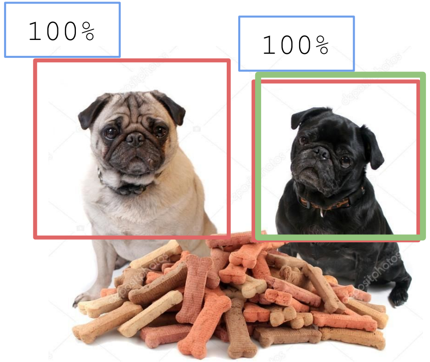
Neural Module Networks (NMN)

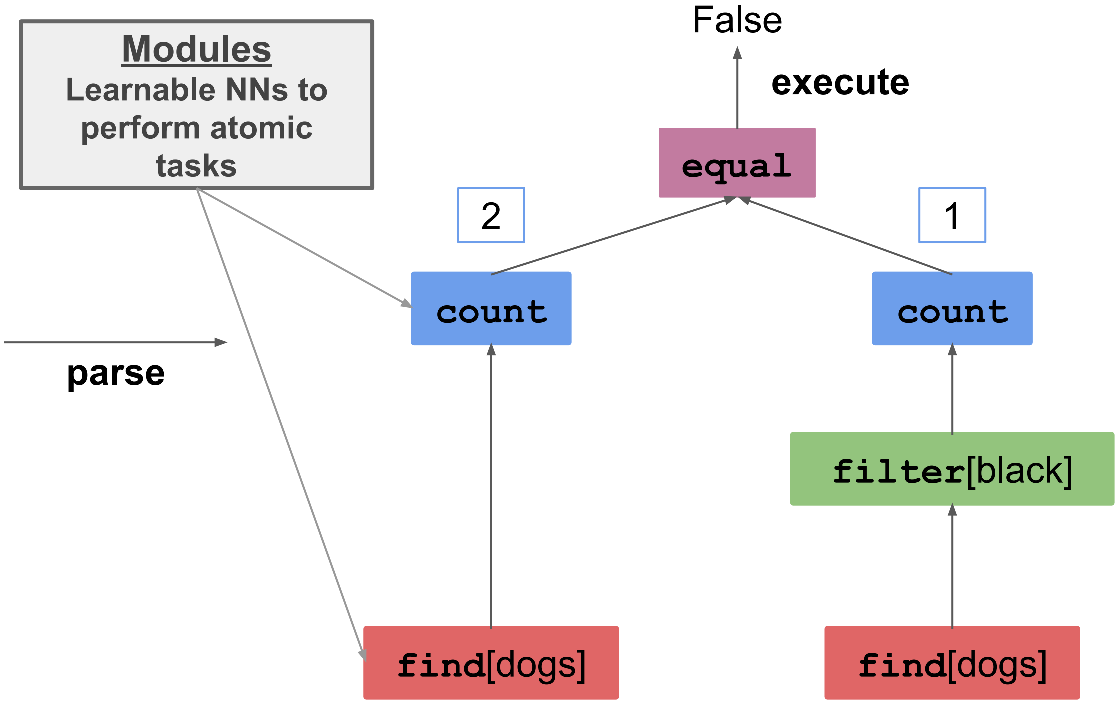


Ai2

Neural Module Networks (NMN)

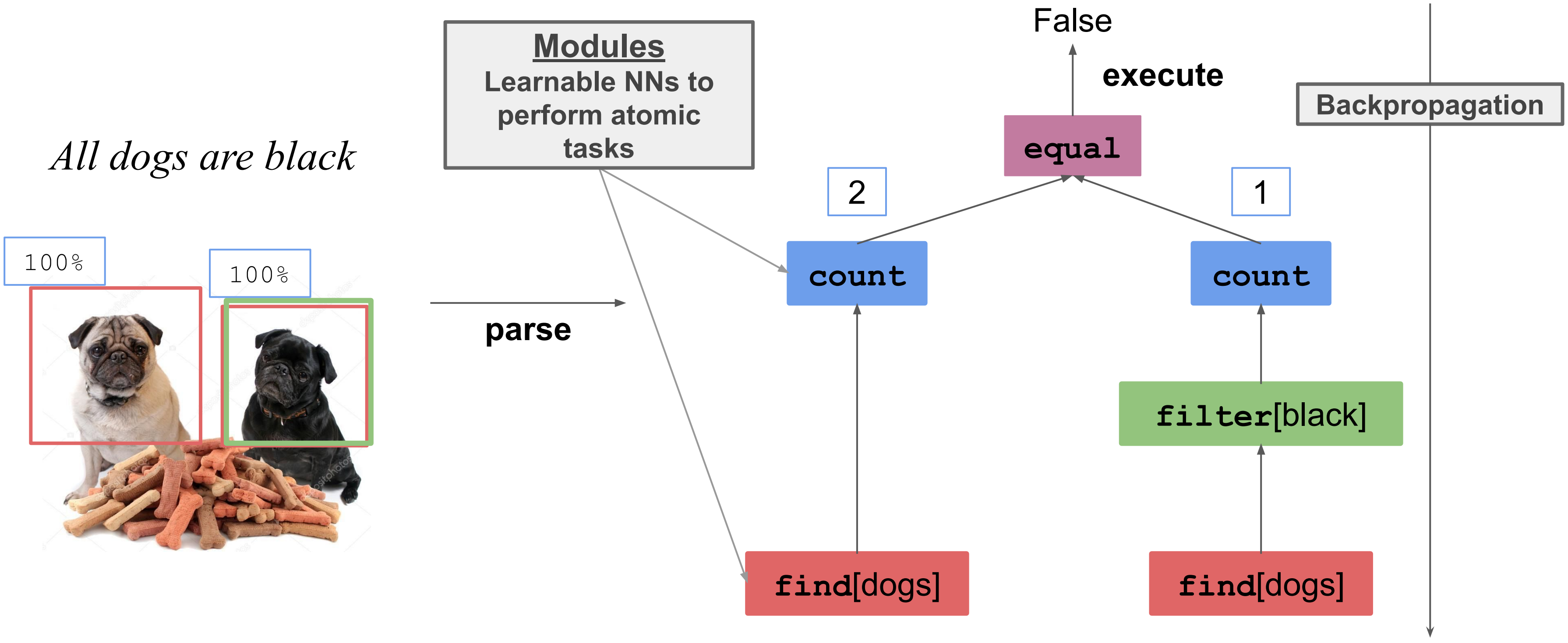
All dogs are black





Neural Module Networks (NMN)

Learn parameters for all modules based on the answer as weak signal

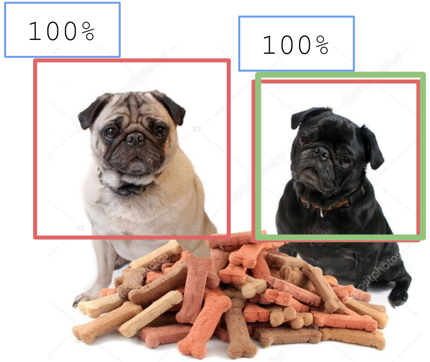


Neural Module Networks (NMN)

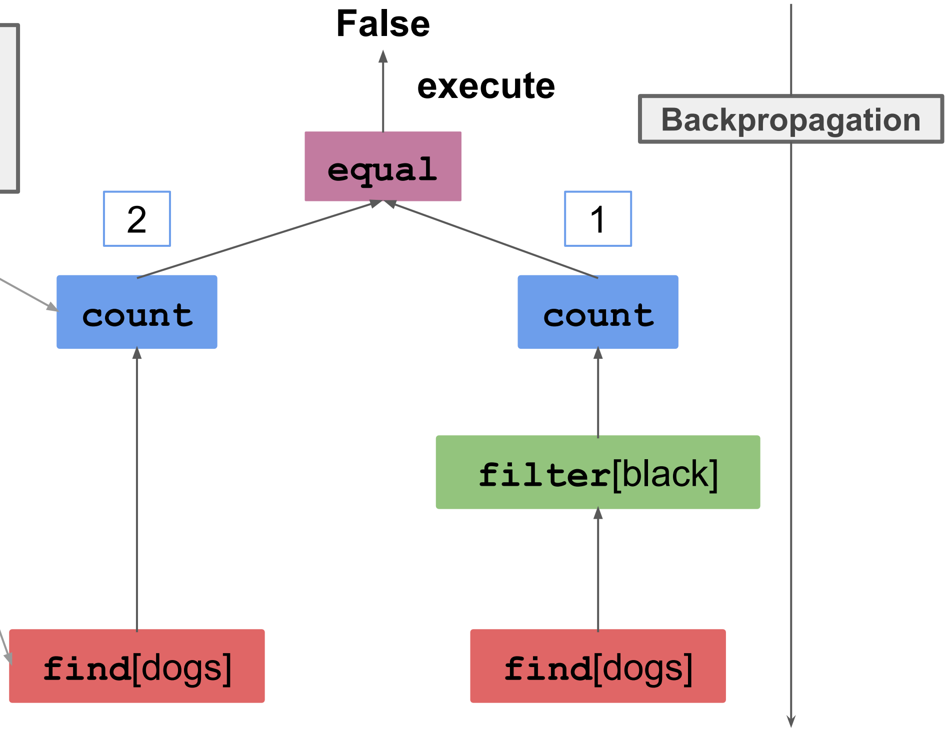
Learn parameters for all modules based on the answer as weak signal

Modules
Learnable NNs to perform atomic tasks

All dogs are black



parse

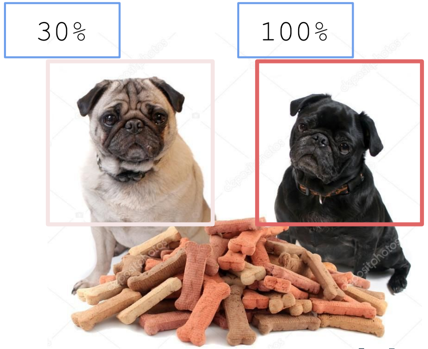


Interpretable!

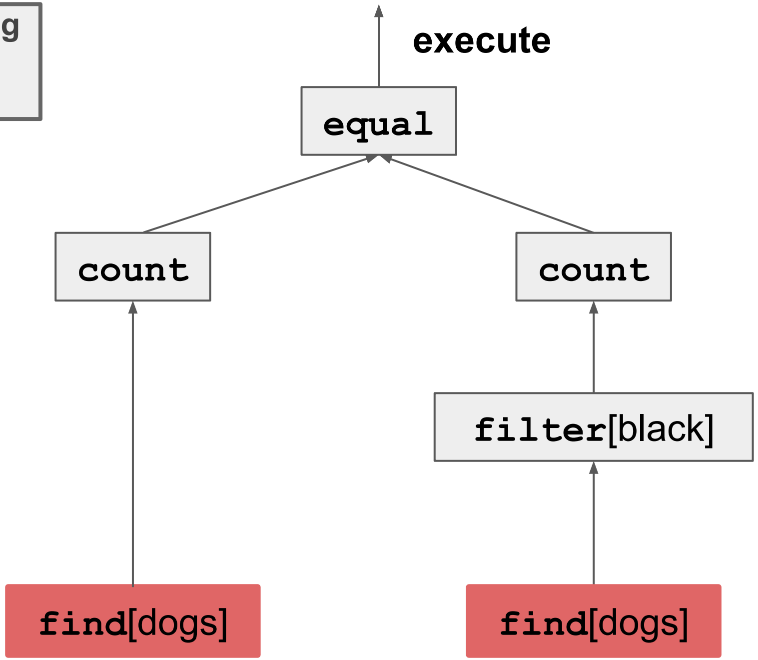
Module execution is not faithful!

After training using only end-task supervision

All dogs are black



parse

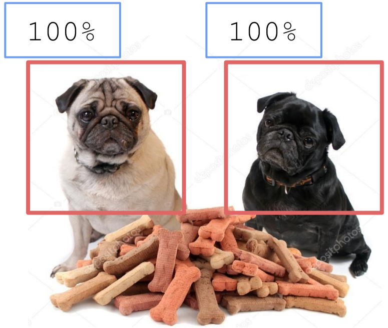


Module execution is not faithful to its intended reasoning

Faithful module execution

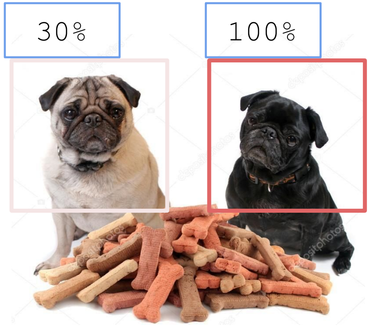
✓ Module performs its intended operation; hence faithful

find[dogs]



✗ Module does not perform its intended operation; hence not-faithful

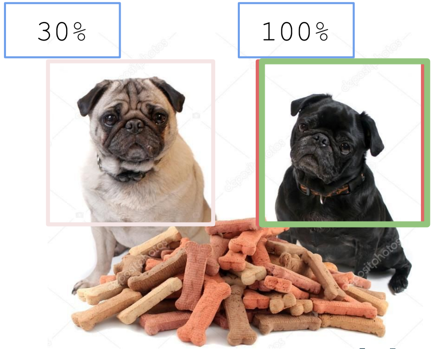
find[dogs]



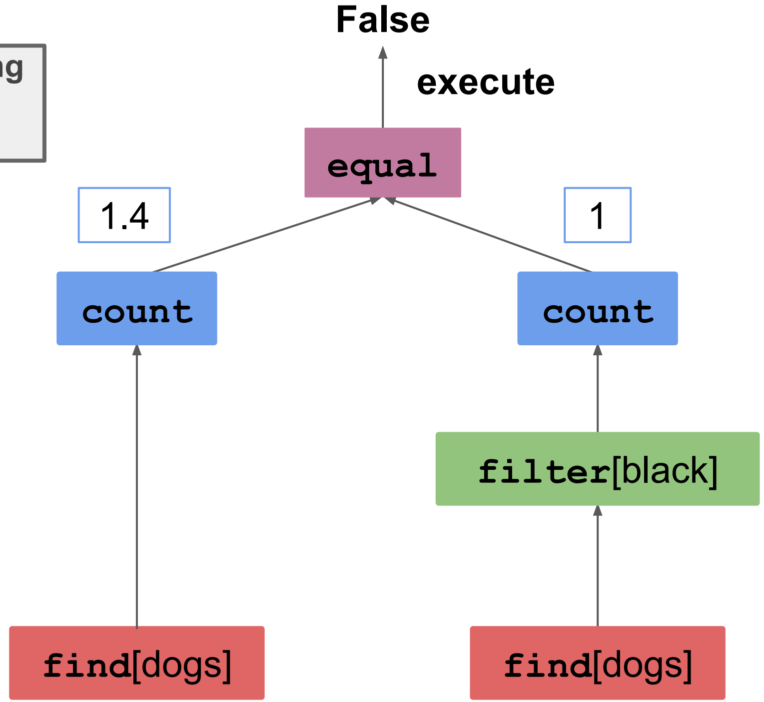
Module execution is not faithful!

After training using only end-task supervision

All dogs are black



parse



Module execution is not faithful to its intended reasoning

Program is *not a faithful explanation* of the model behavior

In this work …

We propose,

(1) Ways to improve module-wise faithfulness

(2) Systematic evaluation of intermediate module execution

In this work …

We propose,

(1) **Ways to improve module-wise faithfulness**

(2) Systematic evaluation of intermediate module execution

What’s causing the unfaithful interpretations?

- Model gets high accuracy but low faithfulness → multiple reasoning steps are being collapsed within one module (or in the contextualizing model)
- Possible causes for the collapsed reasoning:
    - Count architecture is too expressive
    - Contextualized representations already reflect the reasoning
- Supervising module outputs directly is another method

Dataset and Implementation

NLVR2 (Suhr et al., 2019)

two dogs are touching a food dish with their face

Train and evaluate on examples with QDMR program annotation
~32,000 examples

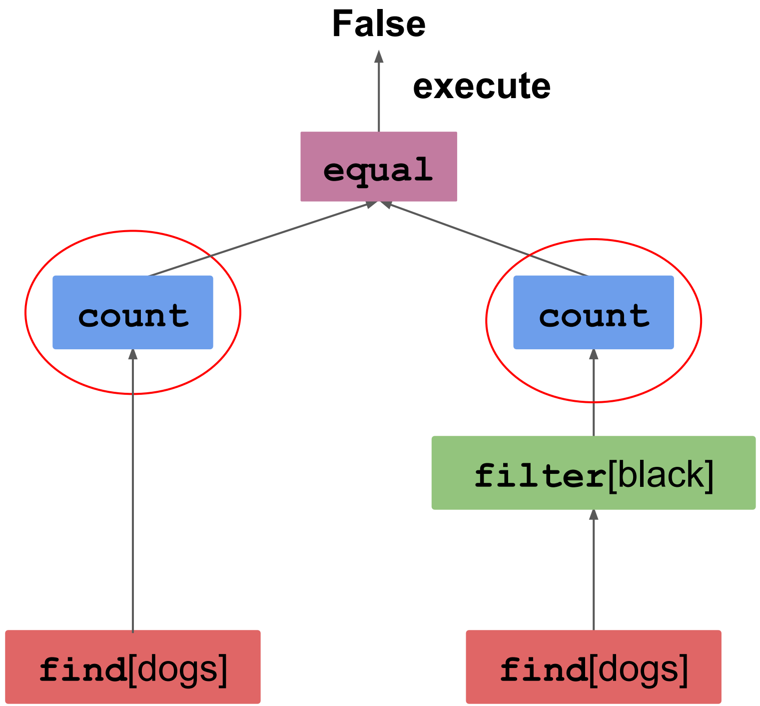
[BREAK; Wolfson et al. 2020]

Module List:

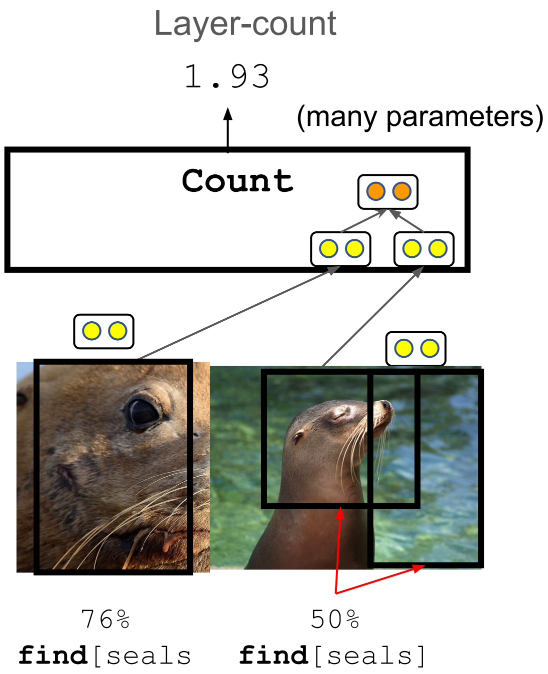
- Find() → ObjectSet
- Filter(ObjectSet) → ObjectSet
- Relation(ObjectSet, ObjectSet) → ObjectSet
- Project(ObjectSet) → ObjectSet
- Count(ObjectSet) → number
- Parameter-less: Equals, Greater-than, etc.
- Macros: In-each-image, In-at-least-one-image

1) Visual-NMN: Count module mediates backprop

all dogs are black



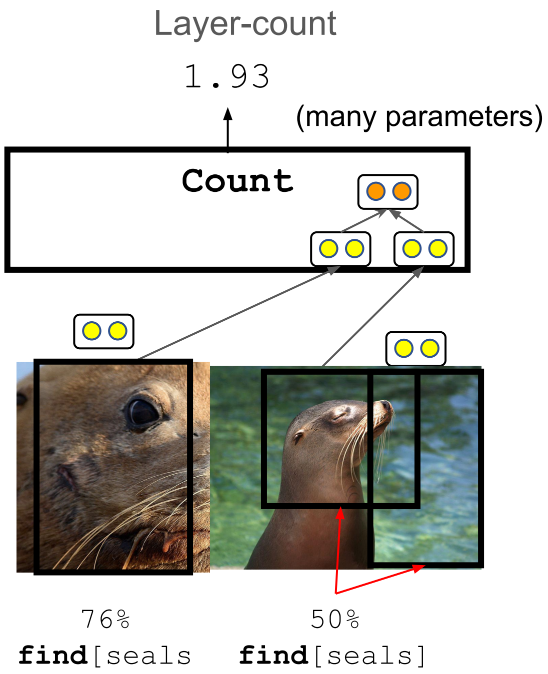
1) Visual-NMN: Lower-capacity Count Module improves faithfulness



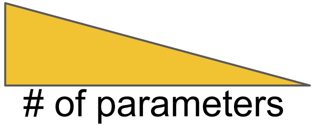
there are three seals in the image pair.

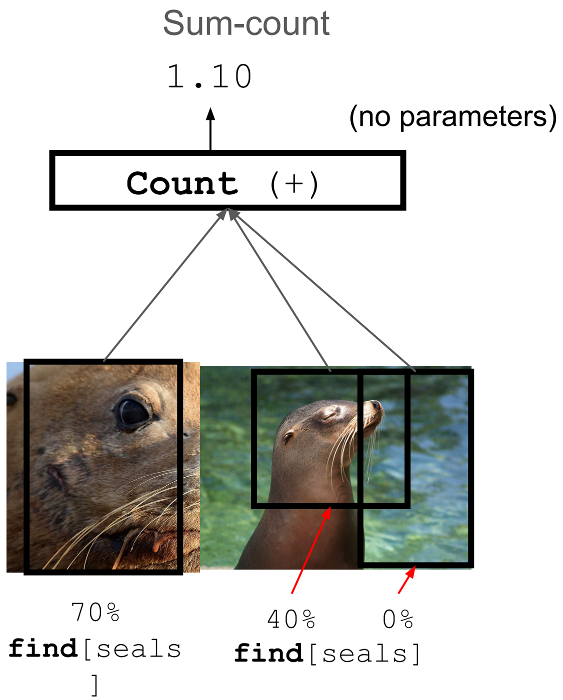
→ **Answer: False**

1) Visual-NMN: Lower-capacity Count Module improves faithfulness

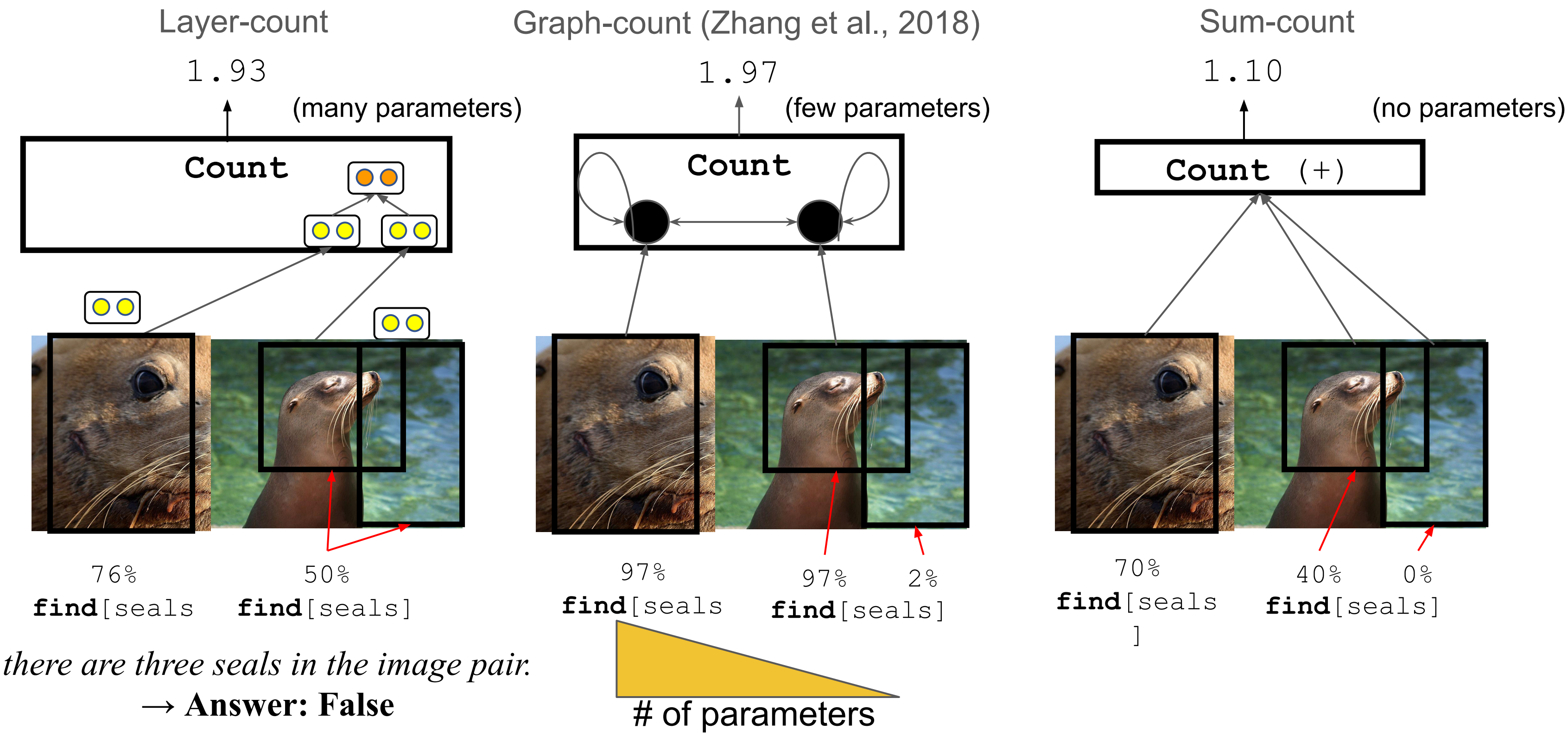


there are three seals in the image pair.
→ **Answer: False**





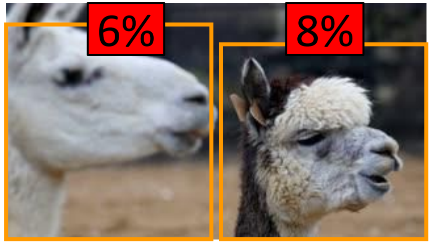
1) Visual-NMN: Lower-capacity Count Module improves faithfulness



there are three seals in the image pair.

→ **Answer: False**

2) Decontextualized Word Vectors Improve Faithfulness

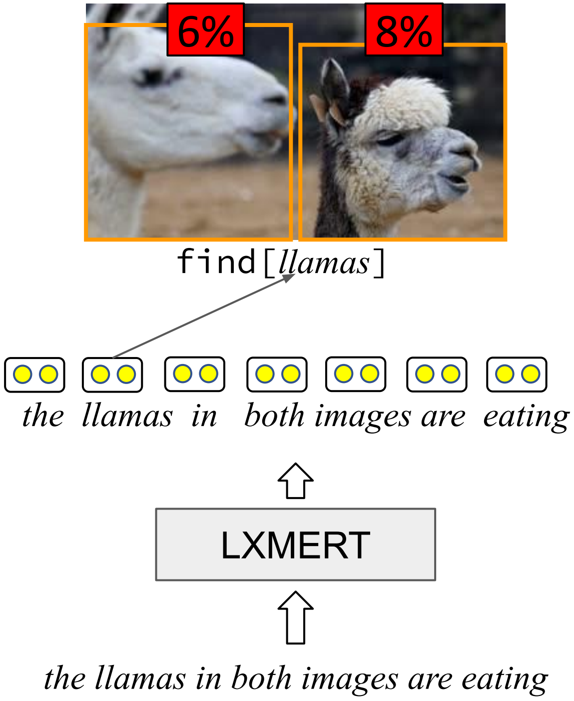


find[*llamas*]

the llamas in both images are eating

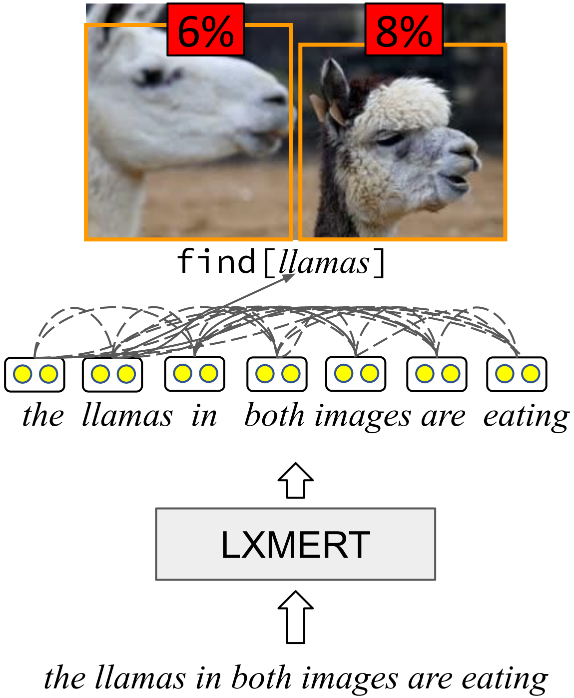
doesn't find llamas

2) Decontextualized Word Vectors Improve Faithfulness



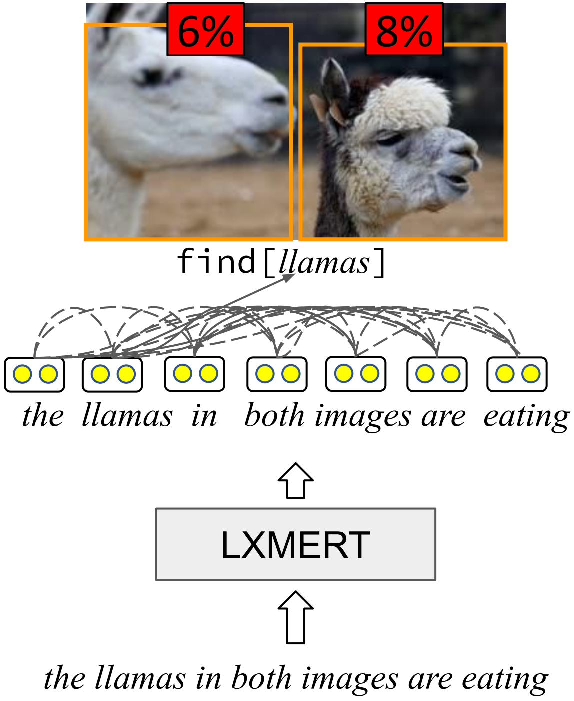
doesn't find llamas

2) Decontextualized Word Vectors Improve Faithfulness

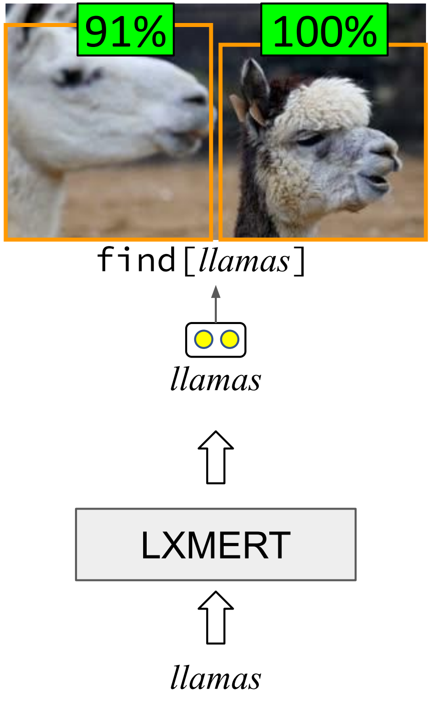


doesn’t find llamas, effectively searching for eating llamas

2) Decontextualized Word Vectors Improve Faithfulness



doesn't find llamas, effectively searching for eating llamas

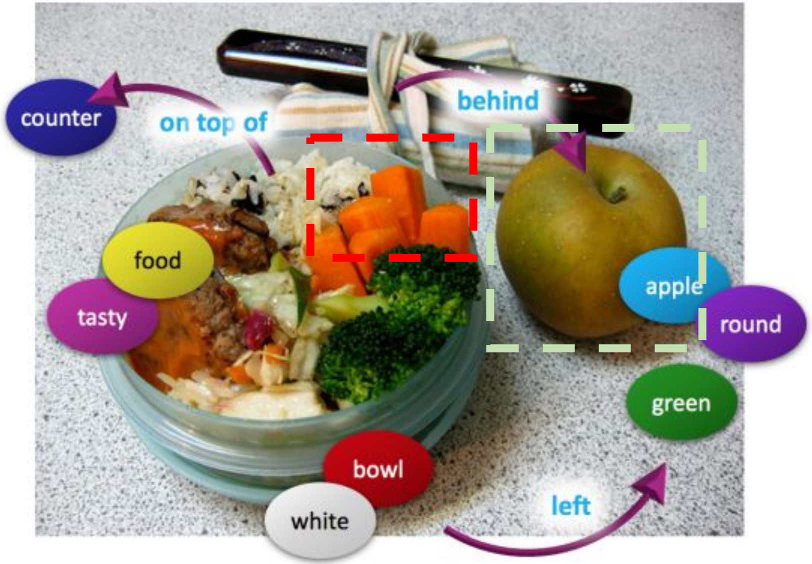


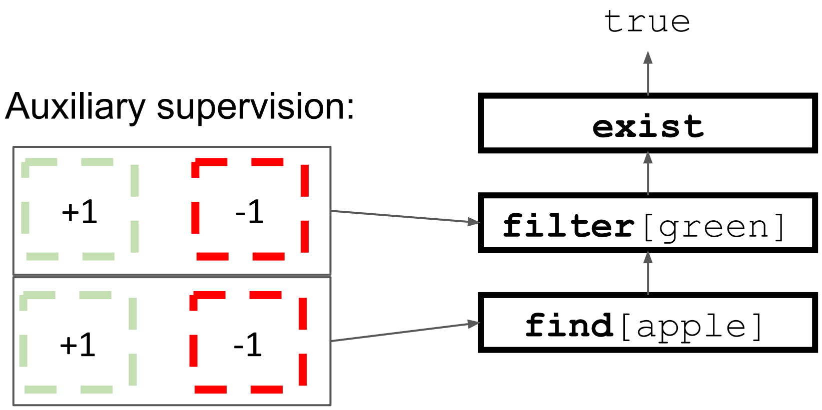
Correctly finds the llamas

3) Supervising module output improves faithfulness

Pre-train find and filter with auxiliary module supervision on different dataset (GQA)

Is there a green apple?





[Hudson and Manning, 2019]

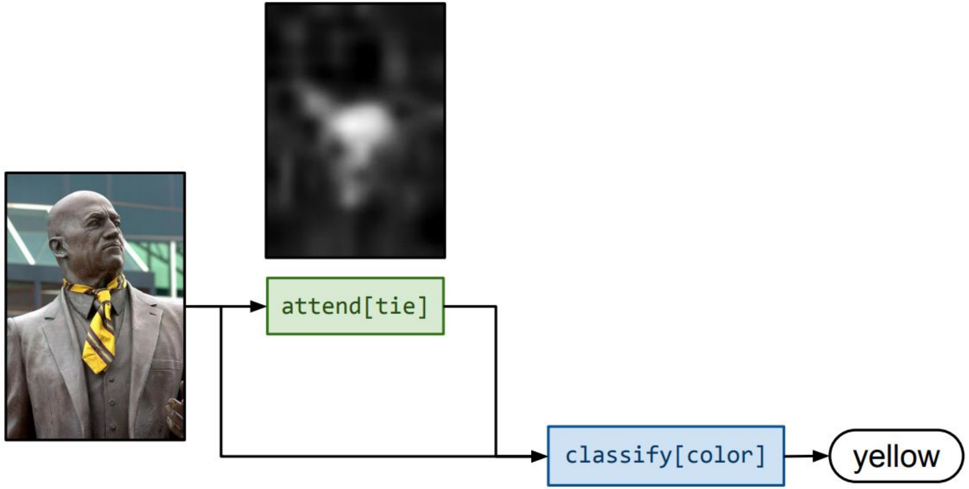
In this work …

We propose,

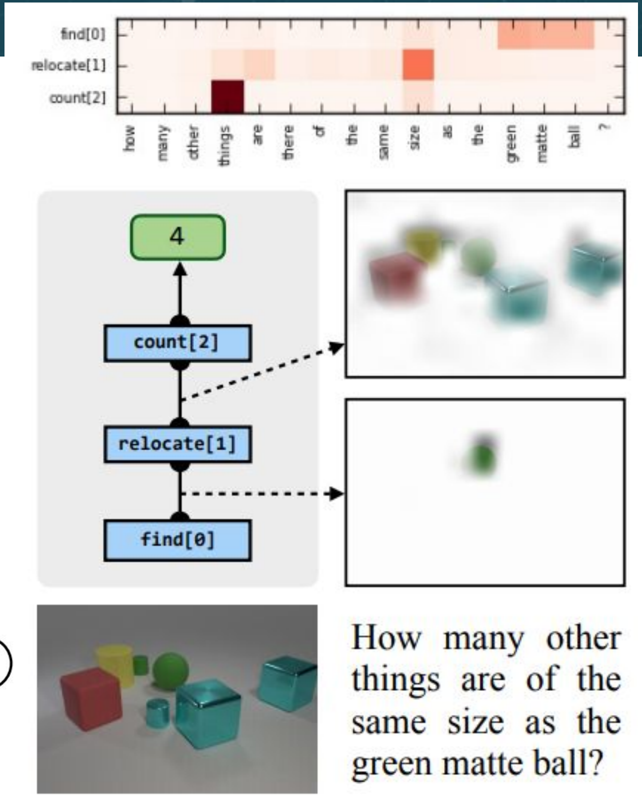
(1) Ways to improve module-wise faithfulness

(2) Systematic evaluation of intermediate module execution

Previous work



[Andreas et al. 2016]



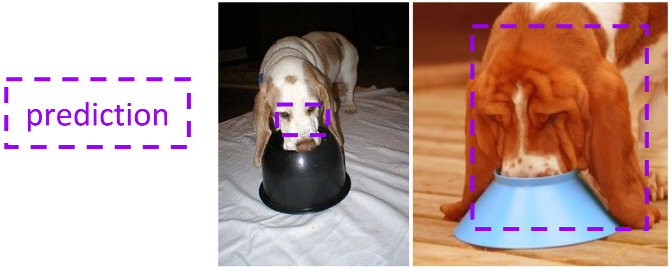
[Hu et al. 2017]

Previous work: Human evaluation of module outputs

- One exception in previous work: Hu et al. 2018 asks humans to evaluate module outputs in two ways:
  - Subjective understanding: Rate (on a 4-point scale) how well you can understand the model's reasoning via the module outputs
  - Forward prediction: Predict the model's output and failure based on the module outputs
- Our approach allows evaluation of multiple models without any additional annotations.

How do we evaluate faithfulness?

two dogs are touching a food dish with their face

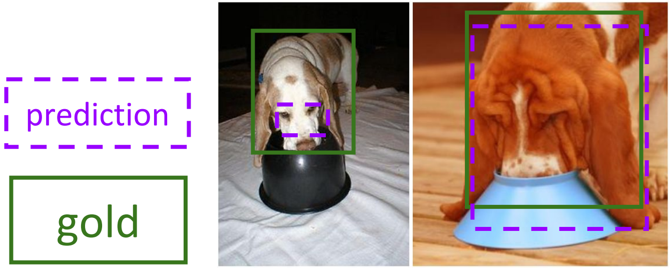


Gold Program

```
equal
  count
    with-relation [is touching]
      relocate [face]
          find [dog]
      find [food dish]
  number [two]
```

How do we evaluate faithfulness?

two dogs are touching a food dish with their face



Gold Program

```
equal
  count
    with-relation [is touching]
      relocate [face]
         find [dog]
      find [food dish]
  number [two]
```

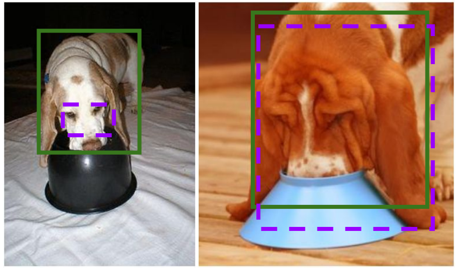
We collect intermediate outputs for 536 programs

How do we evaluate faithfulness?

two dogs are touching a food dish with their face

prediction

gold



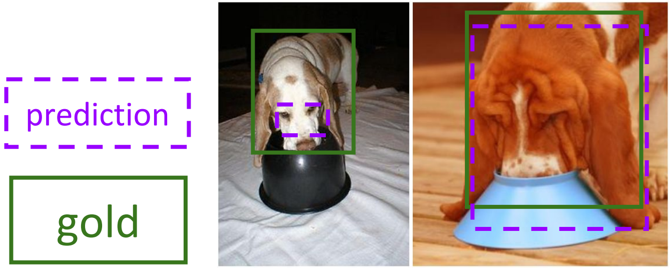
Gold Program

```
equal
  count
    with-relation [is touching]
      relocate [face]
          find [dog]
      find [food dish]
  number [two]
```

We collect intermediate outputs for 536 programs → Compute precision, recall, F_1

How do we evaluate faithfulness?

two dogs are touching a food dish with their face



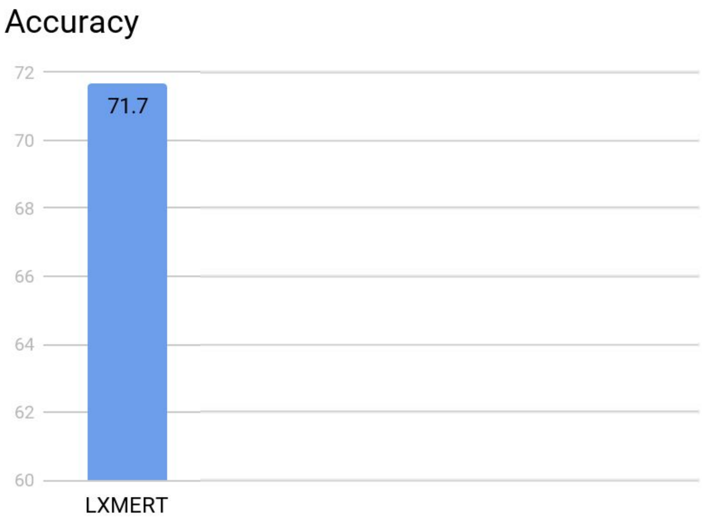
F_1: 0.5

Gold Program

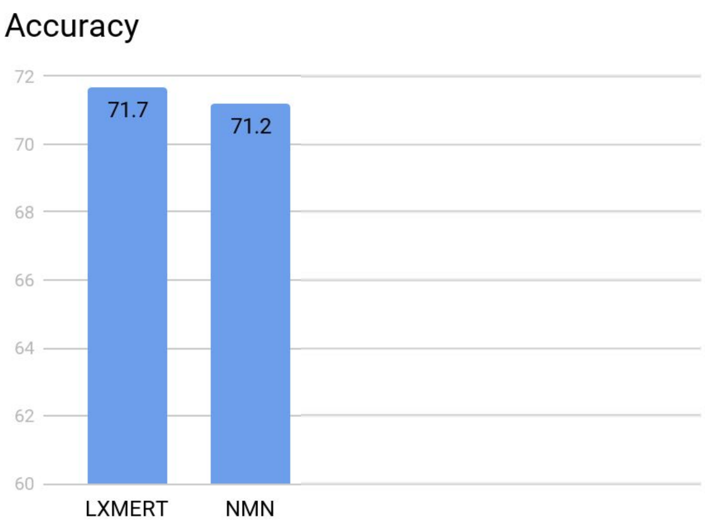
```
equal
  count
    with-relation [is touching]
      relocate [face]
          find [dog]
      find [food dish]
  number [two]
```

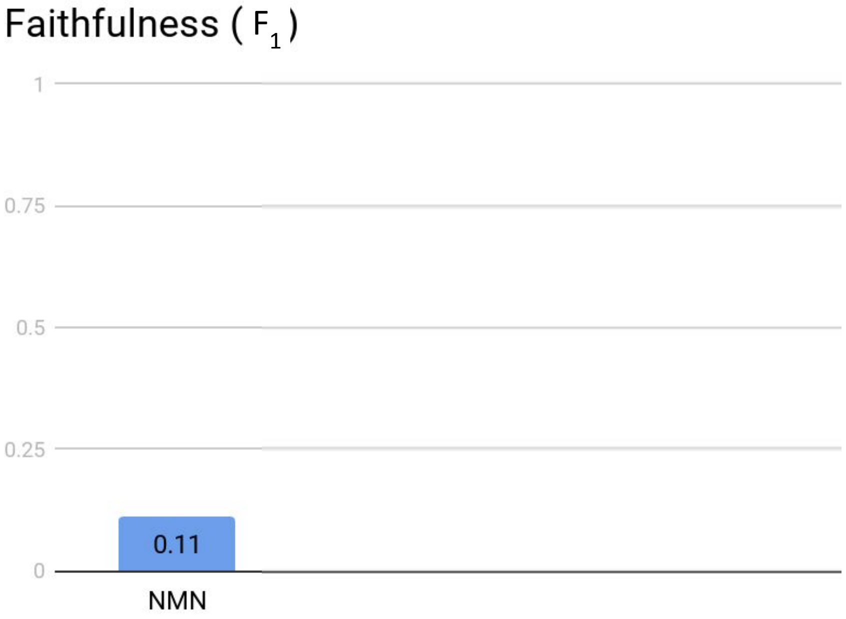
We collect intermediate outputs for 536 programs → Compute precision, recall, F_1

Results - NLVR2



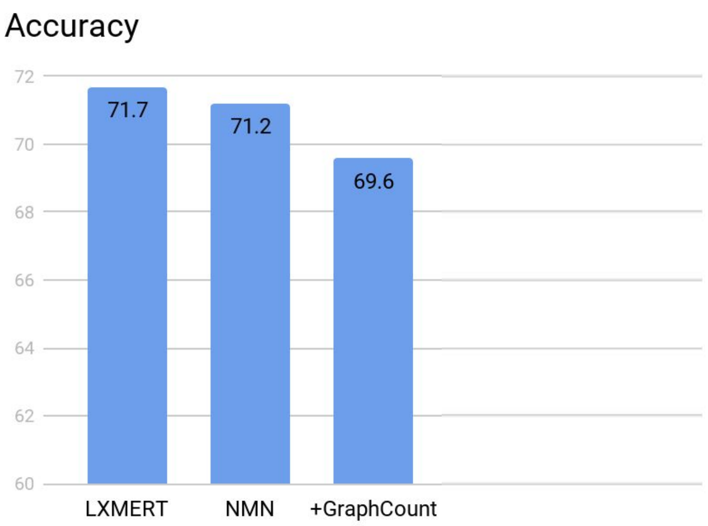
Results - NLVR2

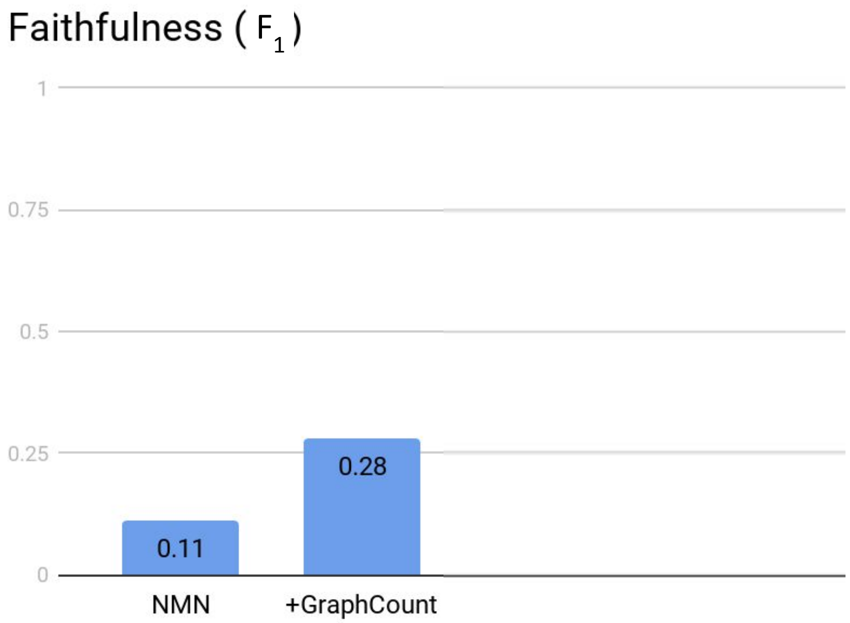




(average across modules)

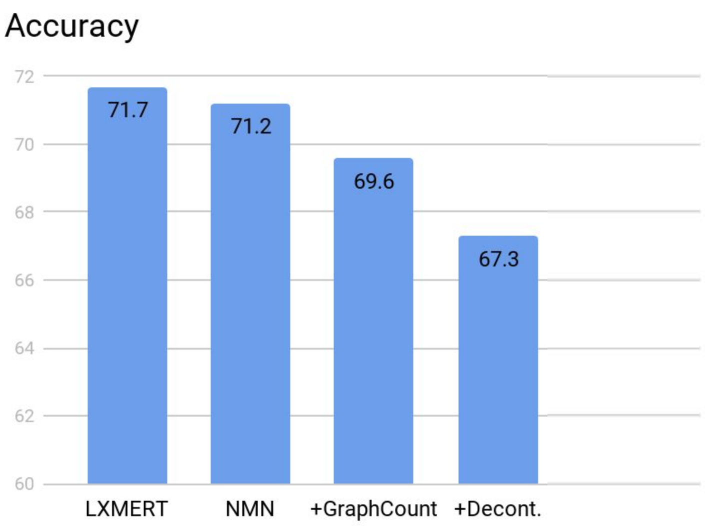
Results - NLVR2

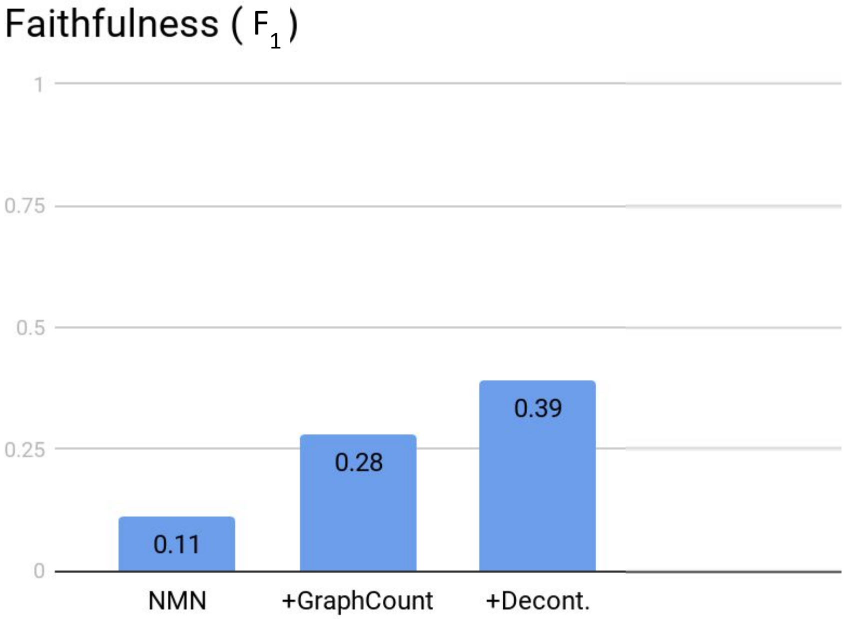




(average across modules)

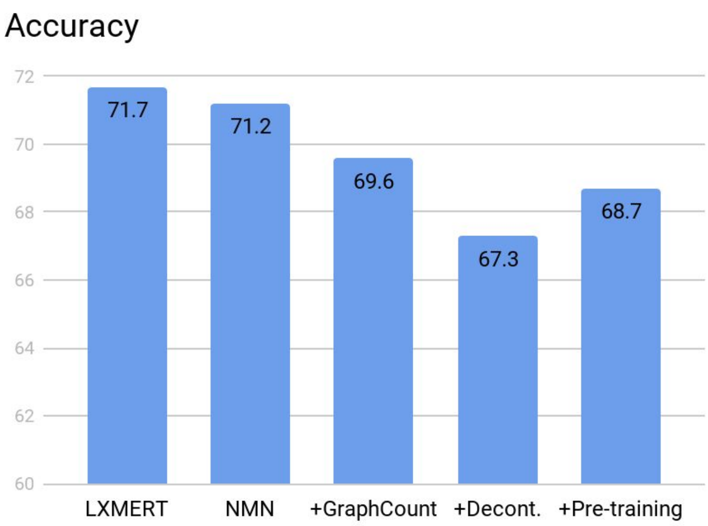
Results - NLVR2

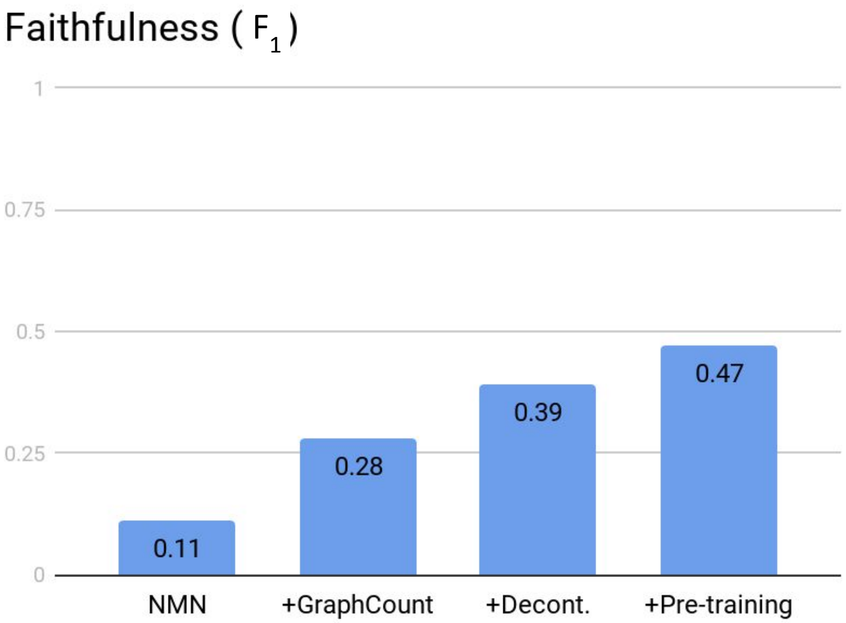




(average across modules)

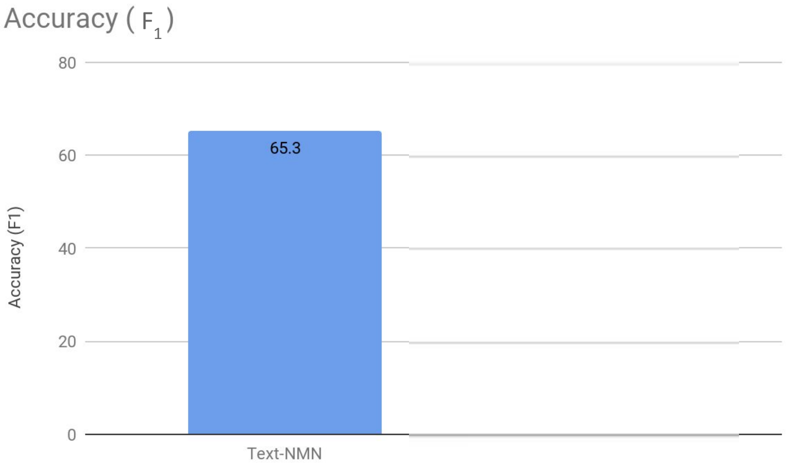
Results - NLVR2

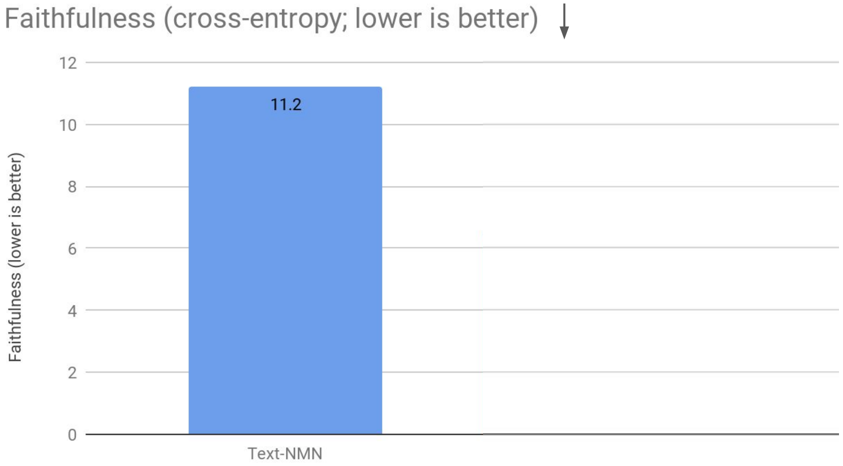




(average across modules)

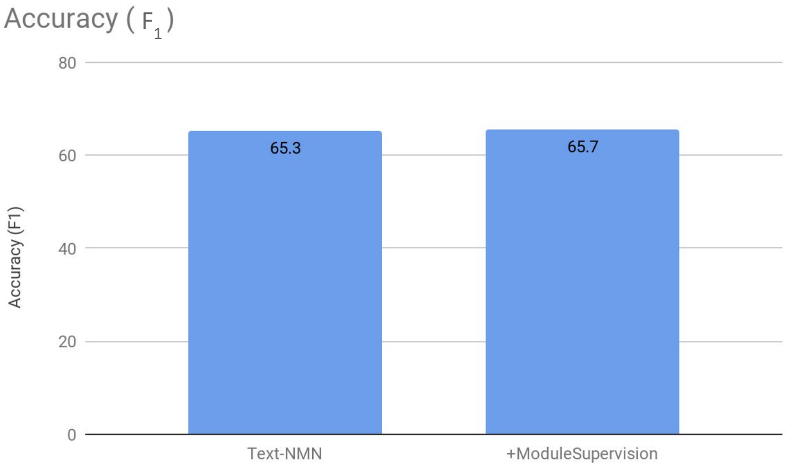
Results - DROP

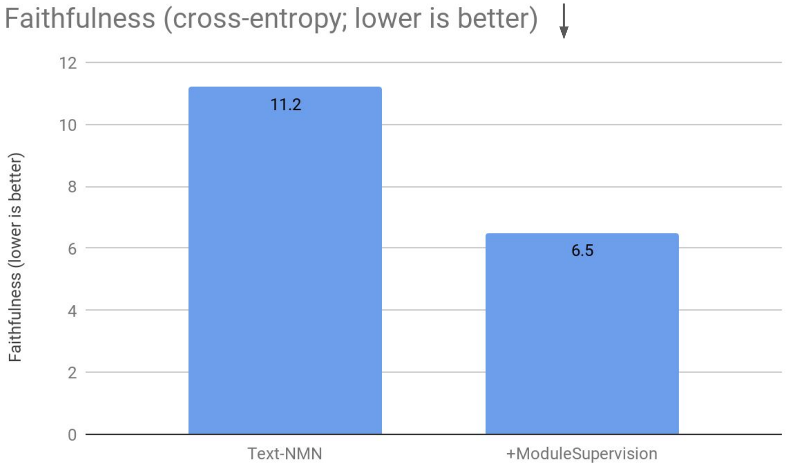




(average across modules)

Results - DROP





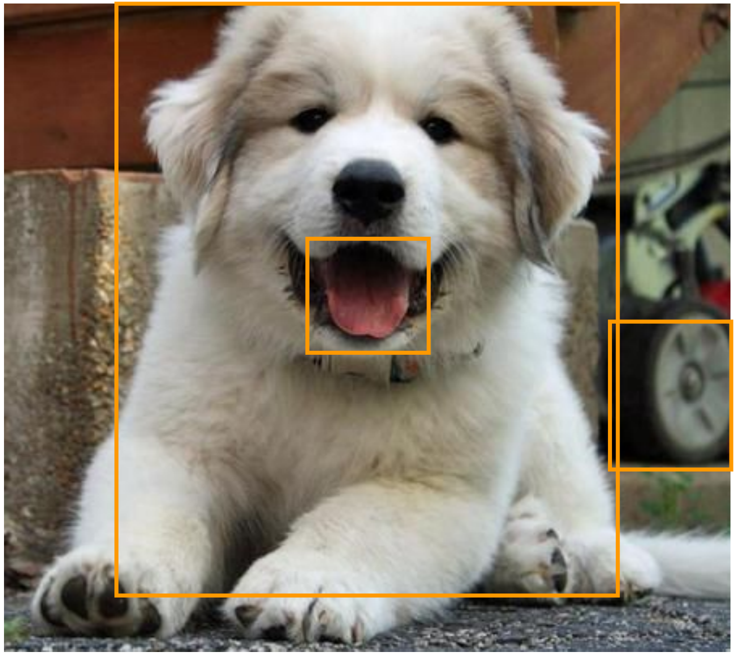
(average across modules)

NLVR2 Example

"a small white puppy is laying down with its tongue out."

NLVR2 Example

"a small white puppy is laying down with its tongue out."

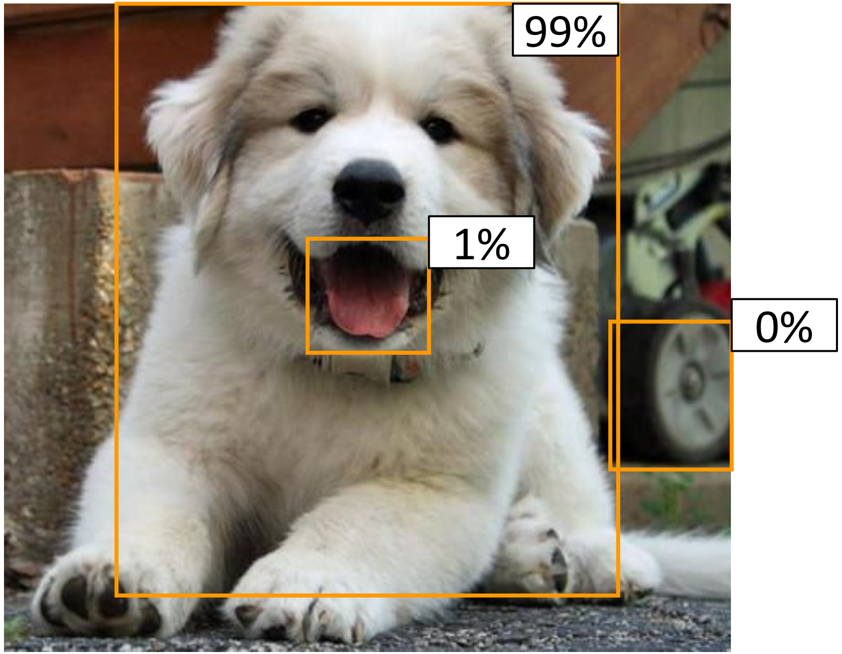


NLVR2 Example

"a small white puppy is laying down with its tongue out."

Gold Program

`find [puppy]` ←

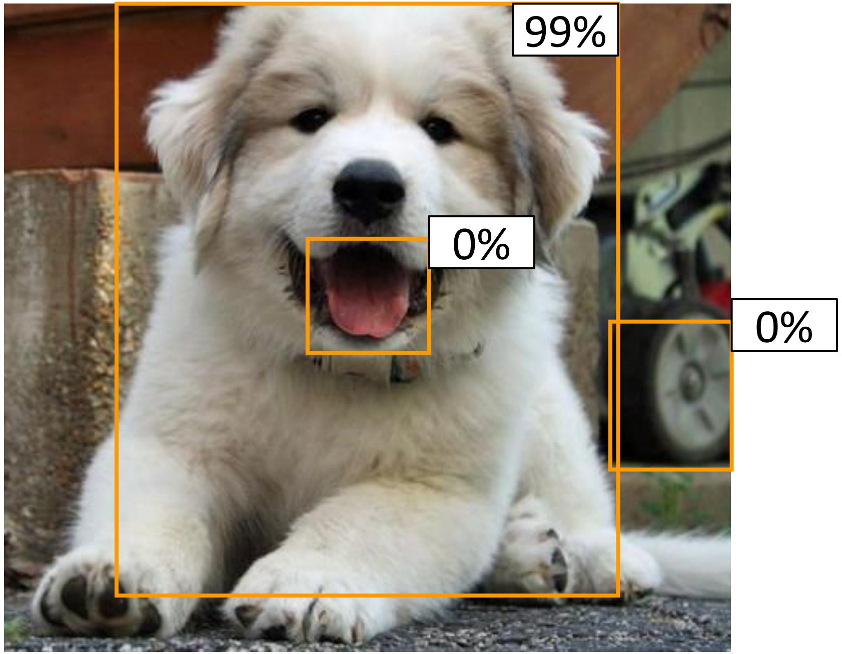


NLVR2 Example

"a small white puppy is laying down with its tongue out."

Gold Program

```
filter [laying down]
    filter [white]
        filter[small]
            find [puppy]
```

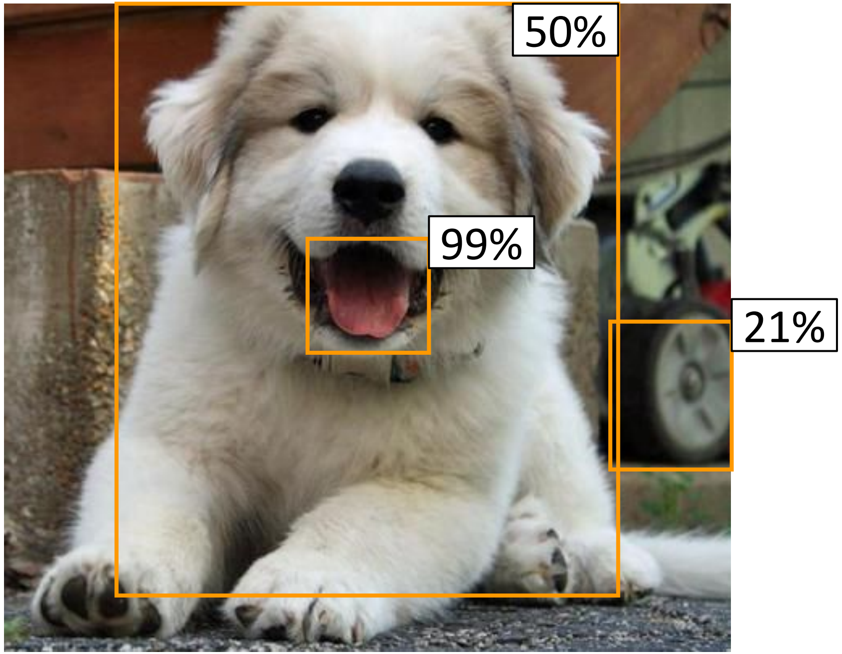


NLVR2 Example

"a small white puppy is laying down with its tongue out."

Gold Program

```
project [tongue]
  filter [laying down]
    filter [white]
      filter[small]
        find [puppy]
```

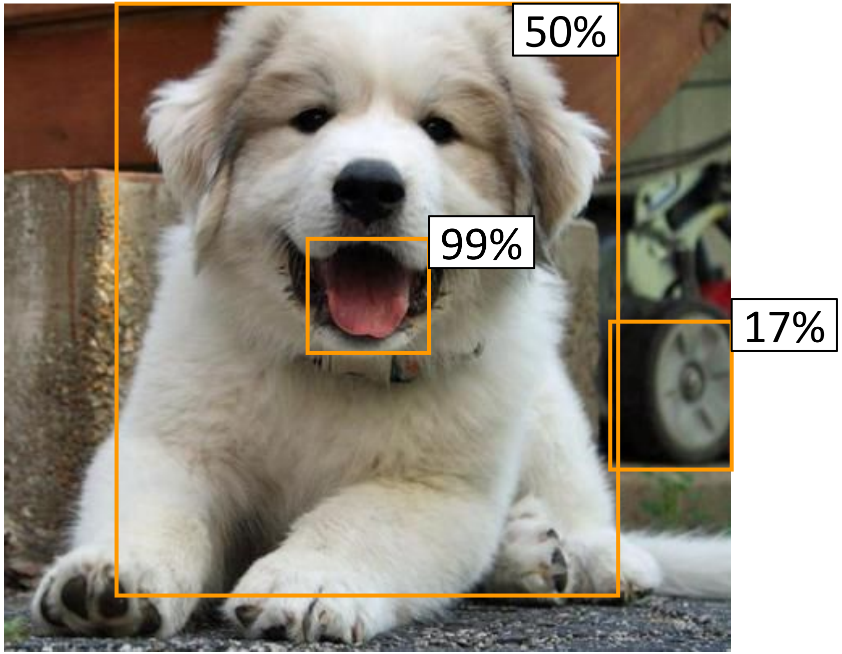


NLVR2 Example

"a small white puppy is laying down with its tongue out."

Gold Program

```
filter [is out]
  project [tongue]
    filter [laying down]
        filter [white]
            filter[small]
                find [puppy]
```

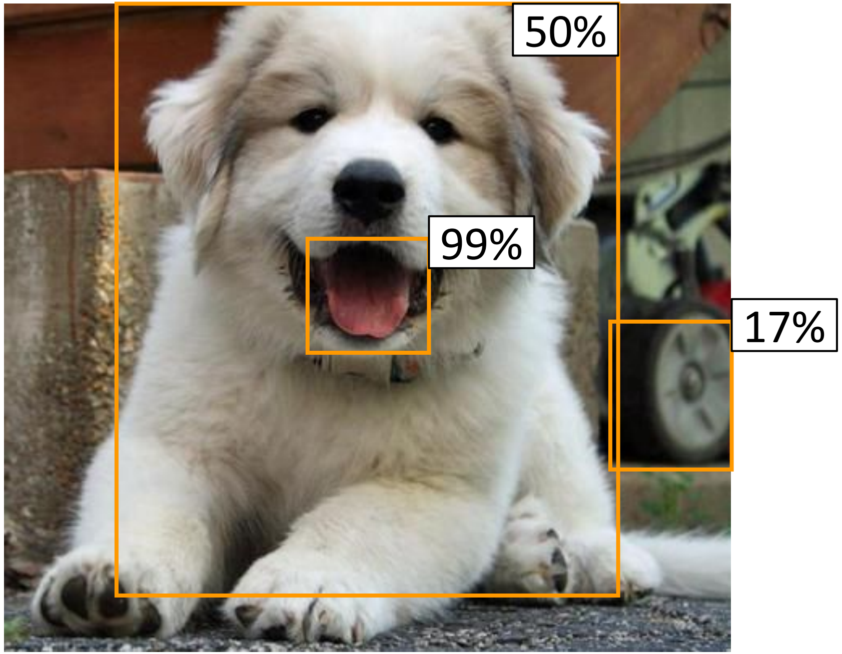


NLVR2 Example

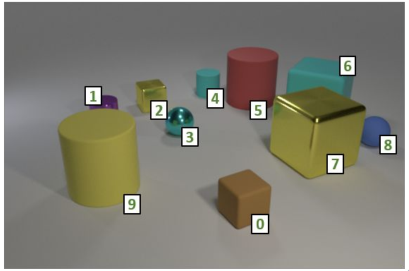
"a small white puppy is laying down with its tongue out."

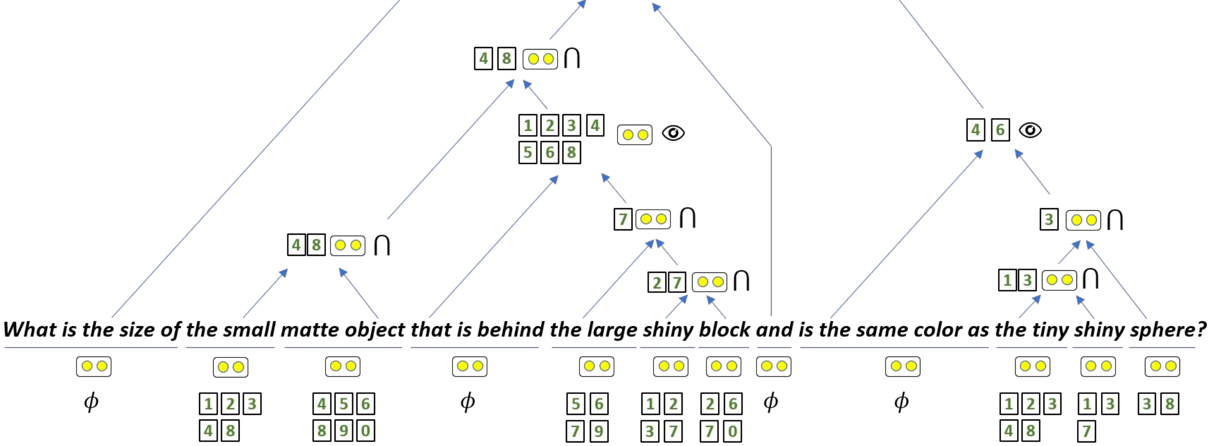
Gold Program

```
Exist True
  filter [is out]
    project [tongue]
      filter [laying down]
          filter [white]
              filter[small]
                  find [puppy]
```



Another example of interpretable compositional reasoning: Grounded Chart Parser





Ben Bogin,

SS, Matt Gardner, Jonathan Berant, Accepted to TACL

Ai2

Grounded Chart Parser Results

Accuracy

	CLEVR	CLOSURE
MAC	98.5	72.4
FiLM	97.0	60.1
GLT (our model)	99.1	**96.1** $\pm$ 2.5
NS-VQA † ∓	**100**	77.2
PG+EE (18K prog.) †	95.4	-
PG-Vector-NMN †	98.0	71.3
GT-Vector-NMN † ‡	98.0	94.4

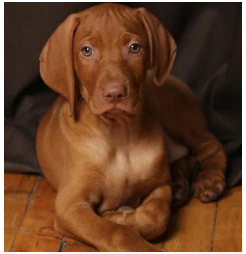
Interpretability

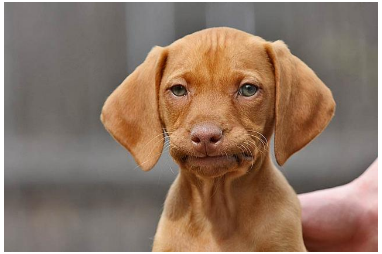
	CLEVR	CLOSURE
Constituents Recall (%)	83.1	81.6
Denotation (F1)	95.9%	94.7

Evaluation: Are pre-trained systems doing compositional reasoning?

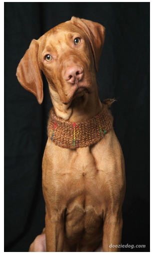
NLVR2 Example: “The dog in the image on the right is wearing a collar.”

Label: False





Label: True

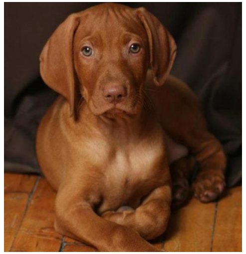


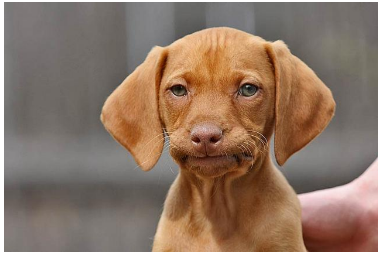
SS, Sameer Singh, Matt Gardner; ViGIL @ NeurIPS 2019

Are pre-trained systems doing compositional reasoning?

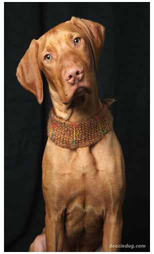
NLVR2 Example: “The dog in the image on the right is wearing a collar.”

Label: False





Label: True



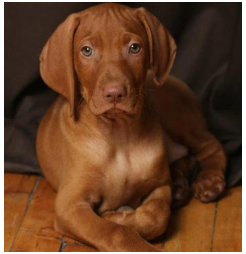
Harder image
(Not Taken from
NLVR2)

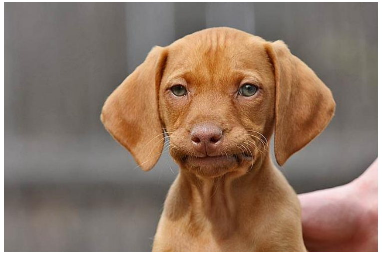
Are pre-trained systems doing compositional reasoning?

NLVR2 Example: “The dog in the image on the right is wearing a collar.”

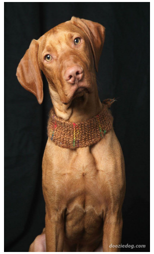
<u>Relation “wearing” is not necessary to answer these correctly</u>

Label: False





Label: True

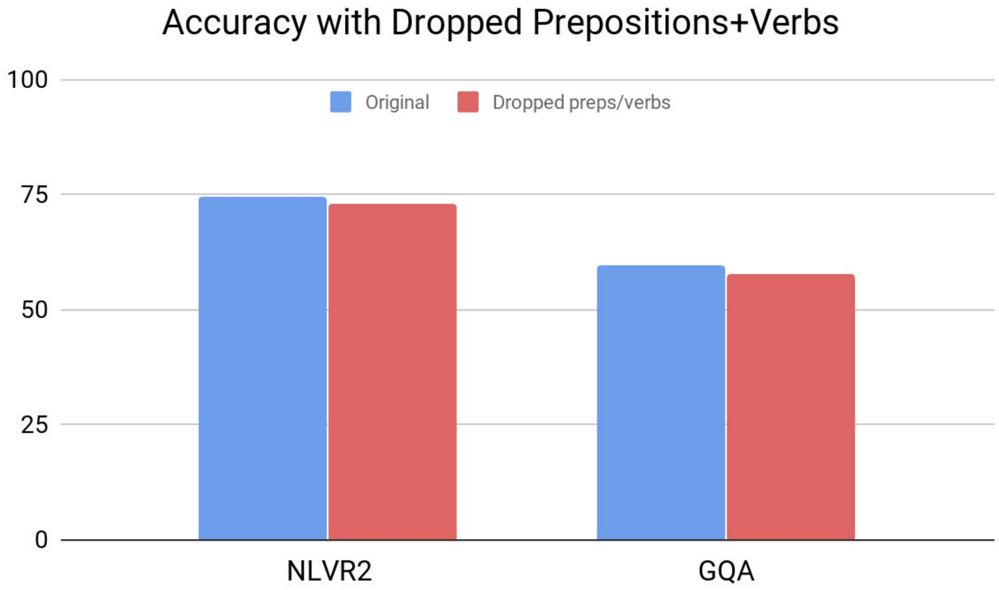


Harder image (Not Taken from NLVR2)

Experiment: Remove relational cues

- Mask/drop prepositions and verbs across all sentences
- LXMERT's Performance is nearly the same!
- Similar result on GQA

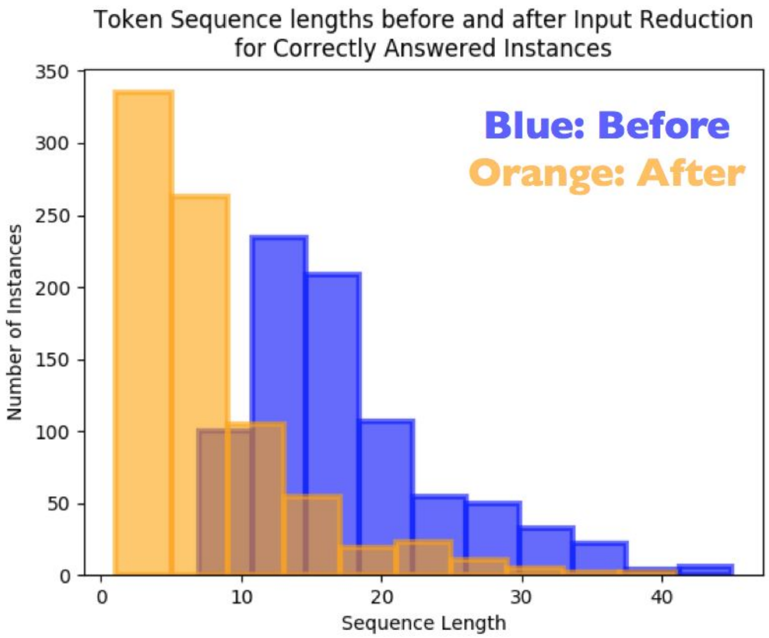
Example
[CLS] the dog ~~in~~ the image ~~on~~ the right ~~is wearing~~ a collar. [SEP]



Experiment: Input Reduction

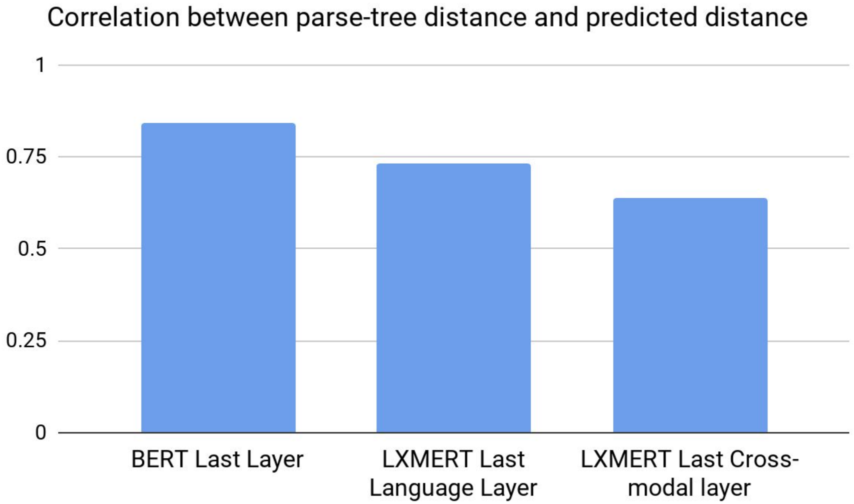
Remove token from NLVR2 sentence with least gradient iteratively **without changing prediction** on any image pair (Feng et al. 2018)

Examples
[CLS] ~~a~~ silver spoon ~~has~~ cookie ~~dough in it .~~ ~~[SEP]~~
~~[CLS] at least one human is wearing~~ eye ~~glasses . [SEP]~~
[CLS] ~~the left~~ and right ~~image contains~~ no more ~~than three~~ bottles ~~of~~ lot ~~##ion . [SEP]~~



Experiment: Syntax Probe

- Compositionality presumably requires some knowledge of syntax
- How well does LXMERT encode syntax trees?
- Structural probe (Hewitt and Manning 2019) learns to map from encoder representations to pairwise parse-tree distance



Evaluation: Contrast sets for NLVR2

- What happens when we modify slightly the input language or images for NLVR2?
- Contrast sets: non-i.i.d. test data for many NLP tasks to evaluate how well models do around local decision boundaries

Matt Gardner and many others, EMNLP-Findings 2020

Evaluation: Contrast sets for NLVR2

- What happens when we modify slightly the input language or images for NLVR2?
- Contrast sets: non-i.i.d. test data for many NLP tasks to evaluate how well models do around local decision boundaries
- NLVR2 Results (for LXMERT):

# of Examples	994
# of Sets	479
Original Test Accuracy	76.4
Contrast Test Accuracy	61.1 (-15.3)
Consistency	30.1

Matt Gardner and many others, EMNLP-Findings 2020

Evaluation: Contrast sets for NLVR2

Example:

Two similarly-colored and similarly-posed chow dogs are face to face in one image.

Two differently-colored but similarly-posed chow dogs are face to face in one image.

Evaluation: Contrast sets for NLVR2

Example:

> Two similarly-colored and similarly-posed chow dogs are face to face in one image.

Conclusion

- Interpretability: Interpretability is still feasible using previous methods (e.g. NMNs) on top of recent pre-trained models
    - Our work relies heavily on gold programs; how well can we do without them?
- Evaluation: Pre-training seems to be very good for grounding nouns and adjectives (and perhaps for counting), but relations seem to need more work

Vision+Language in Scientific Documents

MedICaT: A Dataset of Medical Images, Captions, and Textual References (EMNLP-Findings 2020)

Unique features:

- Subfigure-subcaption alignment annotations for > 2000 figures
- Figure references in main body text for > 70% of figures

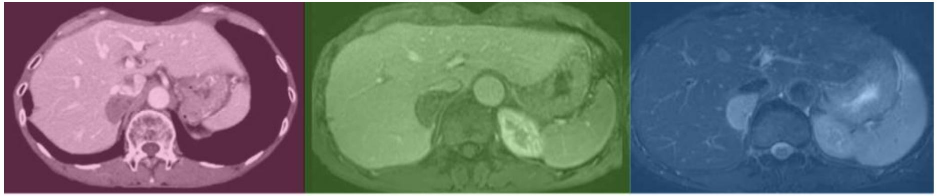
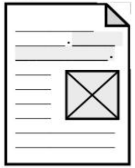


FIGURE 1. The tumor (approximately 40mm in diameter) was hypovascular on enhanced computed tomography scan (right), indicated low intensity on T1-weighted MRI (center), and high intensity on T2-weighted or diffusion MRI (left). Dynamic study revealed peripheral enhancement on a late phase. The tumor located close to the inferior vena cava. MRI = magnetic resonance imaging.



Corresponding inline reference: The tumor was hypovascular on enhanced CT scan (**Figure 1**) and indicated low intensity on T1-weighted magnetic resonance imaging (MRI) and high intensity on T2-weighted or diffusion MRI (**Figure 1**).

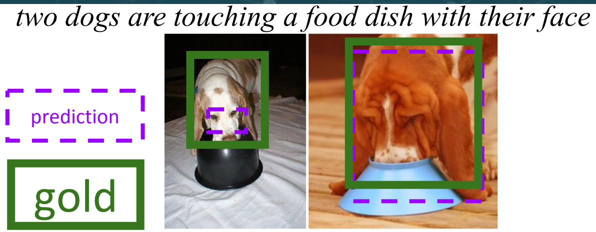
Collaborators



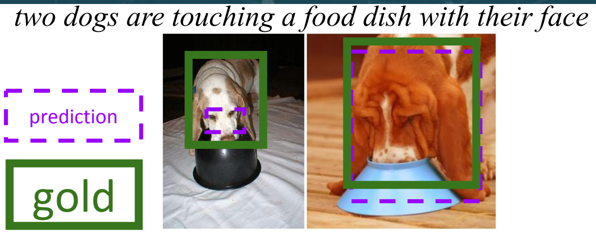



Conclusion

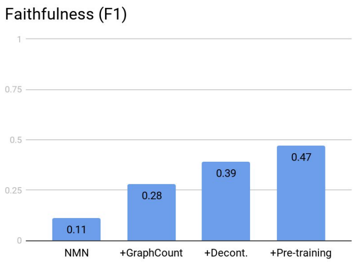
1. **We propose the concept of module-wise faithfulness and ways to systematically evaluate faithfulness in Visual and Text NMN**



Conclusion



1. We propose the concept of module-wise faithfulness and ways to systematically evaluate faithfulness in Visual and Text NMN



2. We propose various ways to improve module-wise faithfulness in NMNs.

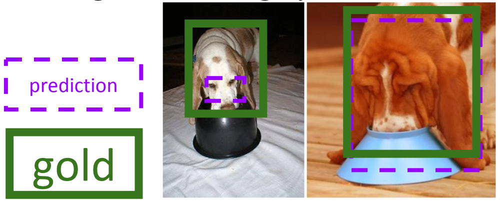
Conclusion

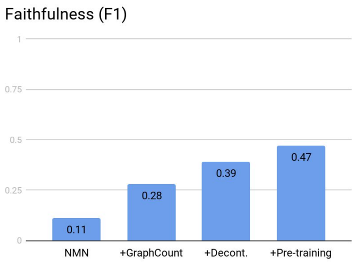
1. **We propose the concept of module-wise faithfulness and ways to systematically evaluate faithfulness in Visual and Text NMN**

2. **We propose various ways to improve module-wise faithfulness in NMNs.**

3. **We release over 700 human-annotated programs with intermediate outputs for NLVR2 and DROP to measure module-wise faithfulness**

two dogs are touching a food dish with their face





Gold Program

```
equal
  count
    with-relation [is touching]
      relocate [face]
          find [dog]
      find [food dish]
  number [two]
```

Conclusion

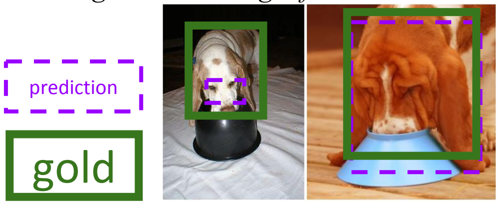
1. We propose the concept of module-wise faithfulness and ways to systematically evaluate faithfulness in Visual and Text NMN

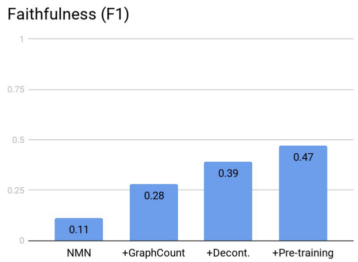
2. We propose various ways to improve module-wise faithfulness in NMNs.

3. We release over 700 human-annotated programs with intermediate outputs for NLVR2 and DROP to measure module-wise faithfulness

Code and annotations: github.com/allenai/faithful-nmn

two dogs are touching a food dish with their face





Gold Program

```
equal
  count
    with-relation [is touching]
      relocate [face]
          find [dog]
      find [food dish]
  number [two]
```

Neural Module Networks for Text Reasoning

Neural networks with learnable parameters to solve an atomic task

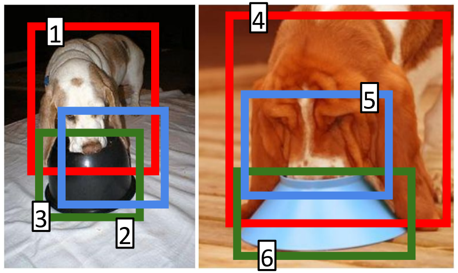
Modules for Visual Reasoning

`find`[arg]	Find bounding boxes corresponding to “arg”
`filter`[condition]	Filter input bounding boxes based on the “condition”
`count`	Count the input number of boxes

....

Two dogs example with what we're evaluating

two dogs are touching a food dish with their face

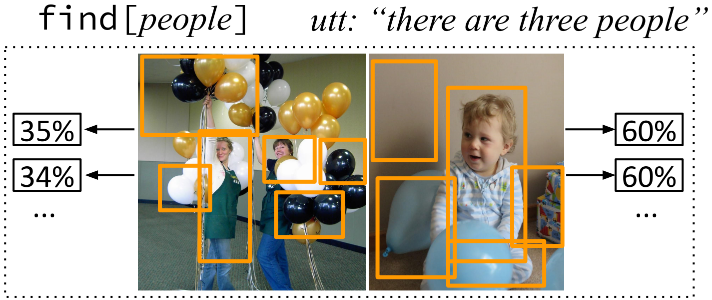


Program	Output
equal	True
count	2
with-relation [is touching]	[2, 5]
relocate [face]	[2, 5]
find [dog]	[1, 4]
find [food dish]	[3, 6]
number [two]	2

Person 2

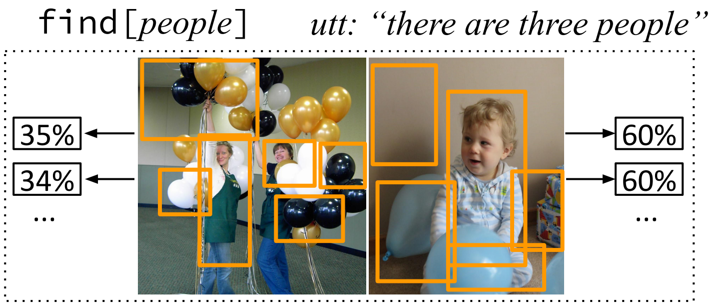
Improvement 1: Architectural choice

- Visual-NMN: Count module occurs in every program
  - Layer-count (most flexible): count = FFNN(box probabilities, box representations)



Improvement 1: Architectural choice

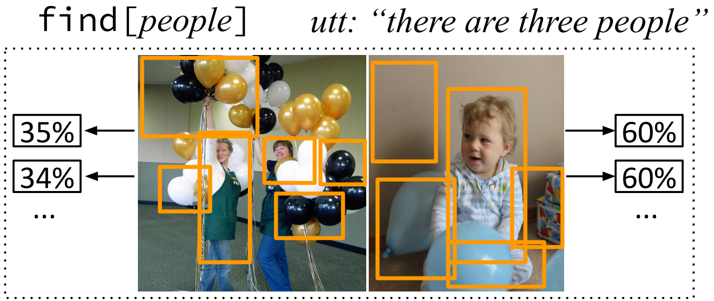
- Visual-NMN: Count module occurs in every program
  - Layer-count (most flexible): count = FFNN(box probabilities, box representations)



  - Sum-count (least flexible): count = Sum(box probabilities)

Improvement 1: Architectural choice

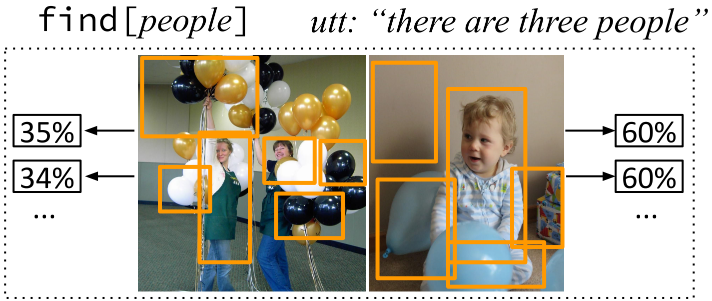
- Visual-NMN: Count module occurs in every program
  - Layer-count (most flexible): count = FFNN(box probabilities, box representations)



  - Sum-count (least flexible): count = Sum(box probabilities)
  - Graph-count: Like Sum-count but accounts for box overlap (Zhang et al., 2018)

Improvement 1: Architectural choice

- Visual-NMN: Count module occurs in every program
    - Layer-count (most flexible): count = FFNN(box probabilities, box representations)



    - Sum-count (least flexible): count = Sum(box probabilities)
    - Graph-count: Like Sum-count but accounts for box overlap (Zhang et al., 2018)
- Text-NMN: "extract-answer" module produces a direct answer without compositional reasoning
    - Can improve accuracy by handling reasoning out of scope of modules
    - Decreases faithfulness by collapsing several reasoning steps

Improvement 2: Supervising module output

- Include loss term for individual module outputs
- Visual-NMN: Supervise object box probabilities
  - Module-wise annotations are not available for NLVR2
  - We pre-train on GQA (Hudson et al., 2019) for which we can obtain annotations
- Text-NMN: Supervise token probabilities
  - We use heuristics (proposed by Gupta et al., 2020) to obtain gold spans for `find-num` and `find-date`

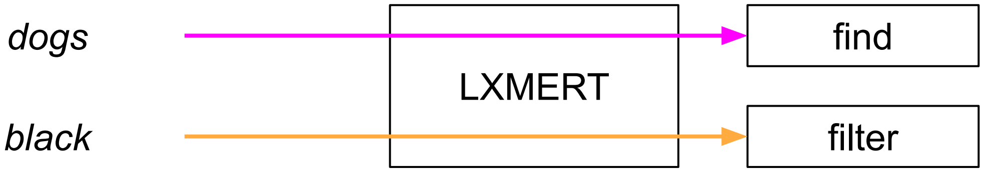
Improvement 3: Decontextualized Word Vectors

- Visual NMN: each module uses an attention over tokens to obtain a weighted average of LXMERT (Tan and Bansal, 2019) token representations
- However, LXMERT's outputs are already contextualized, so tokens outside the attention can still contribute to the attended representation

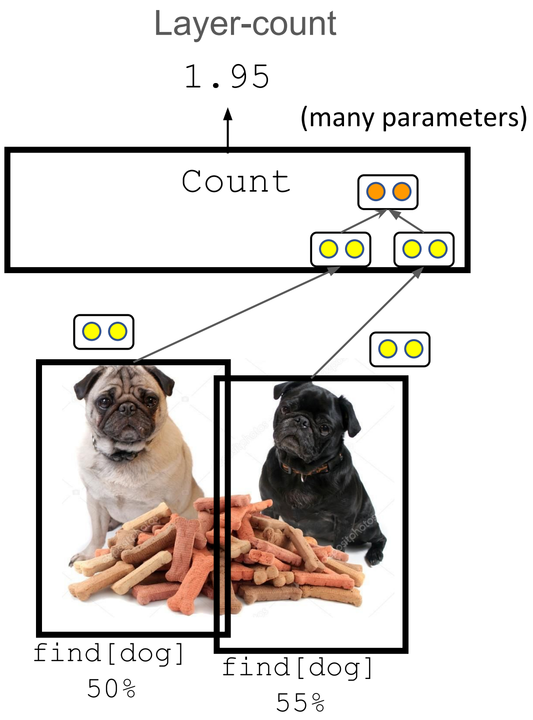
Improvement 3: Decontextualized Word Vectors

- Visual NMN: each module uses an attention over tokens to obtain a weighted average of LXMERT (Tan and Bansal, 2019) token representations
- However, LXMERT's outputs are already contextualized, so tokens outside the attention can still contribute to the attended representation
- Our proposal: Run LXMERT separately for each module, masking out all tokens outside the module's utterance attention

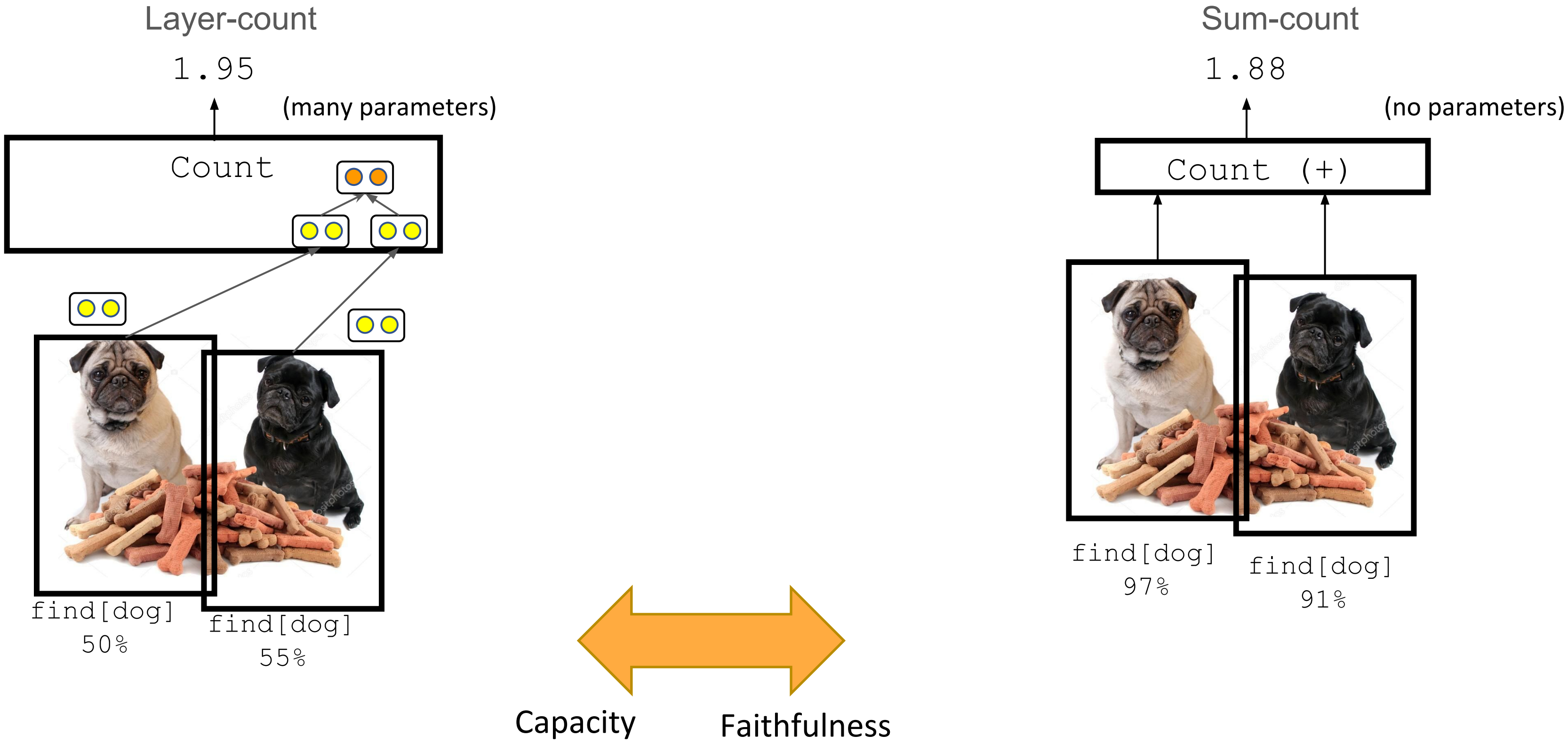
Example: All the dogs are black.



Visual-NMN: Lower-capacity Count Module improves faithfulness



Visual-NMN: Lower-capacity Count Module improves faithfulness



Visual-NMN: Lower-capacity Count Module improves faithfulness