Making the L in VQA Matter

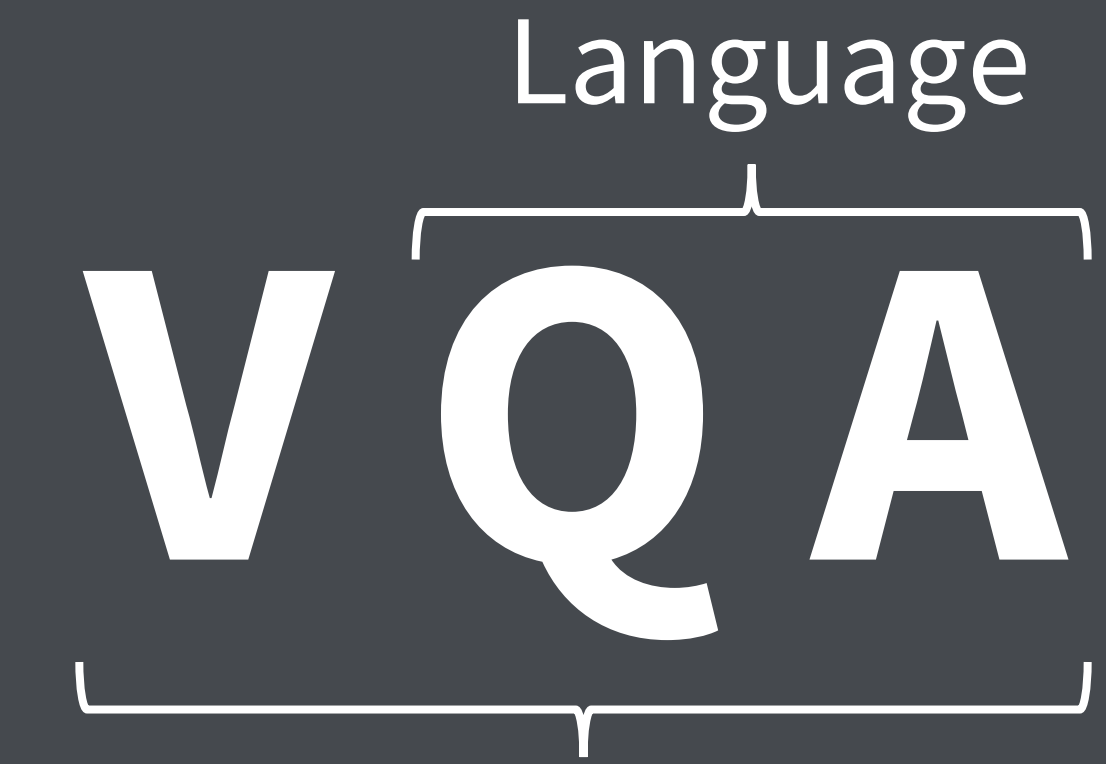
Elevating the Role of Language in Visual Question Answering

Christopher Manning

@chrmanning ❀ @stanfordnlp

VQA Workshop, June 2019

But there is no L in VQA


Language
VQA
Language of Thought

V Q A

Subtext: In a world where even I am able to fire up PyTorch and load a Faster R-CNN object detector, more non-vision people should do VQA !

Answers

- 38% yes/no
- 11% common numbers: 0–16, 18, 20, 24, 25, 30, 50, 100
- Objects, activities, spatial, animates, locations, materials, …

… just about all answers in a few classes

- Nearly all one word stuff

- Not exactly human language in all its grandeur

Answers: Not just one word

- Compound nouns (cf. German): **tennis racket**, **teddy bear**
- Adjective noun: **roman numerals**, **barbed wire**
- Color combinations: **black and white**, **red and white**
- Activities: **playing wii**, **laying down**, **taking selfie**
- Locatives: **on plate**, **on sidewalk**, **in bowl**, **across street**
- Object combinations: **fork and knife**, **ketchup and mustard**
- **I don’t know**

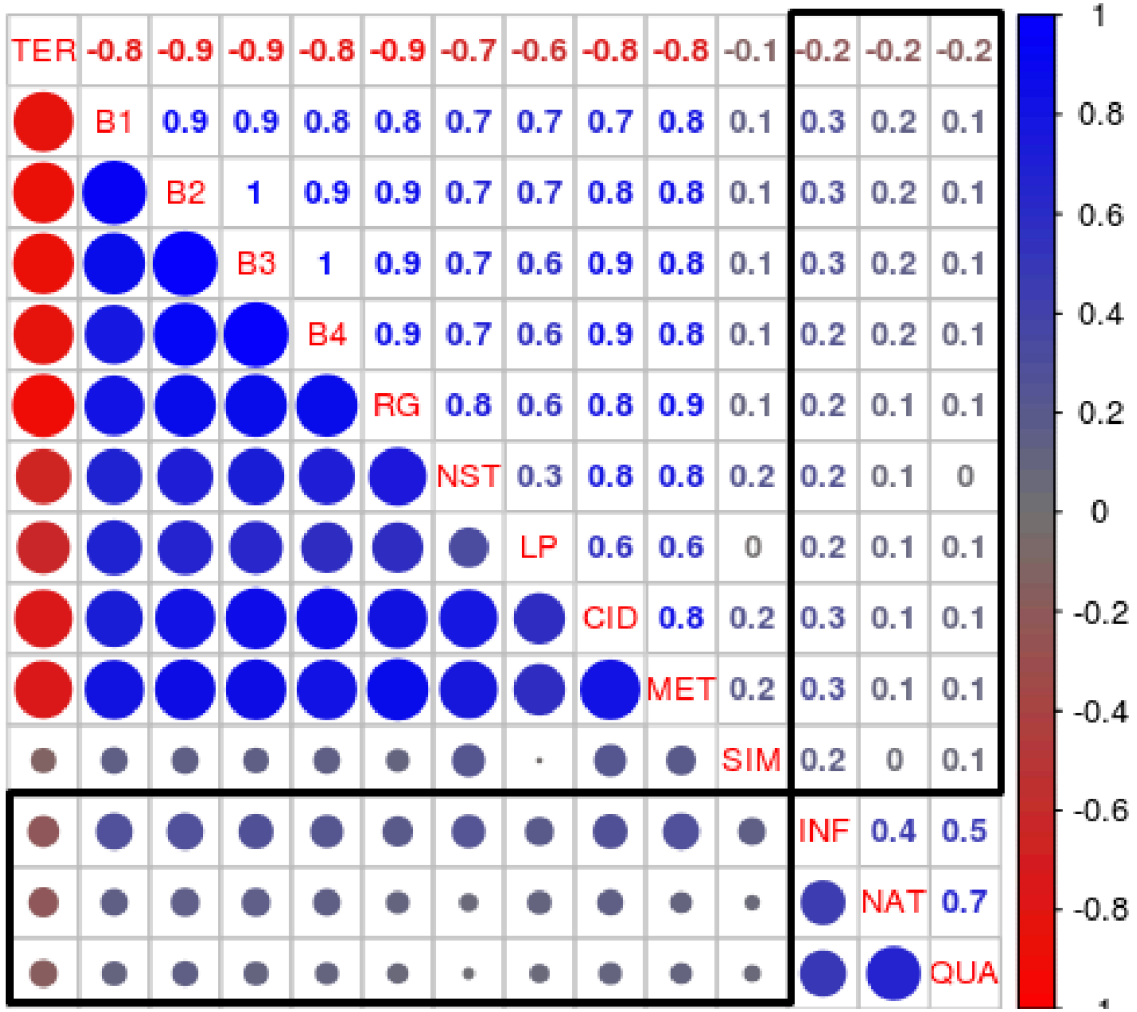
Answers: interesting

- Located objects: **man on right** (22), **one on left** (21)
- Purpose: **to catch frisbee** (8), **to catch ball** (8)
- Quotative (signs, etc.): **100 Year Party Ct** (9), **Via Ferlinghetti** (8)

  — 8 occurrences cut off —
- Q: Why are there no utensils?
  You can eat pizza with your hands,
  Do not need them, **finger food**, **it is pizza**, **unnecessary**, …

Answers: Anything to Change?

- Should we do real Natural La
- **Visual Dialog Challenge** [Da
  - “One fundamental challenge ir
  - NLG metrics are known to corr
    - [Novikova et al. 2017]
  - Solution: Have 100 candidates
    - Use Mean Reciprocal Rank (M
- Maybe choosing from keywo

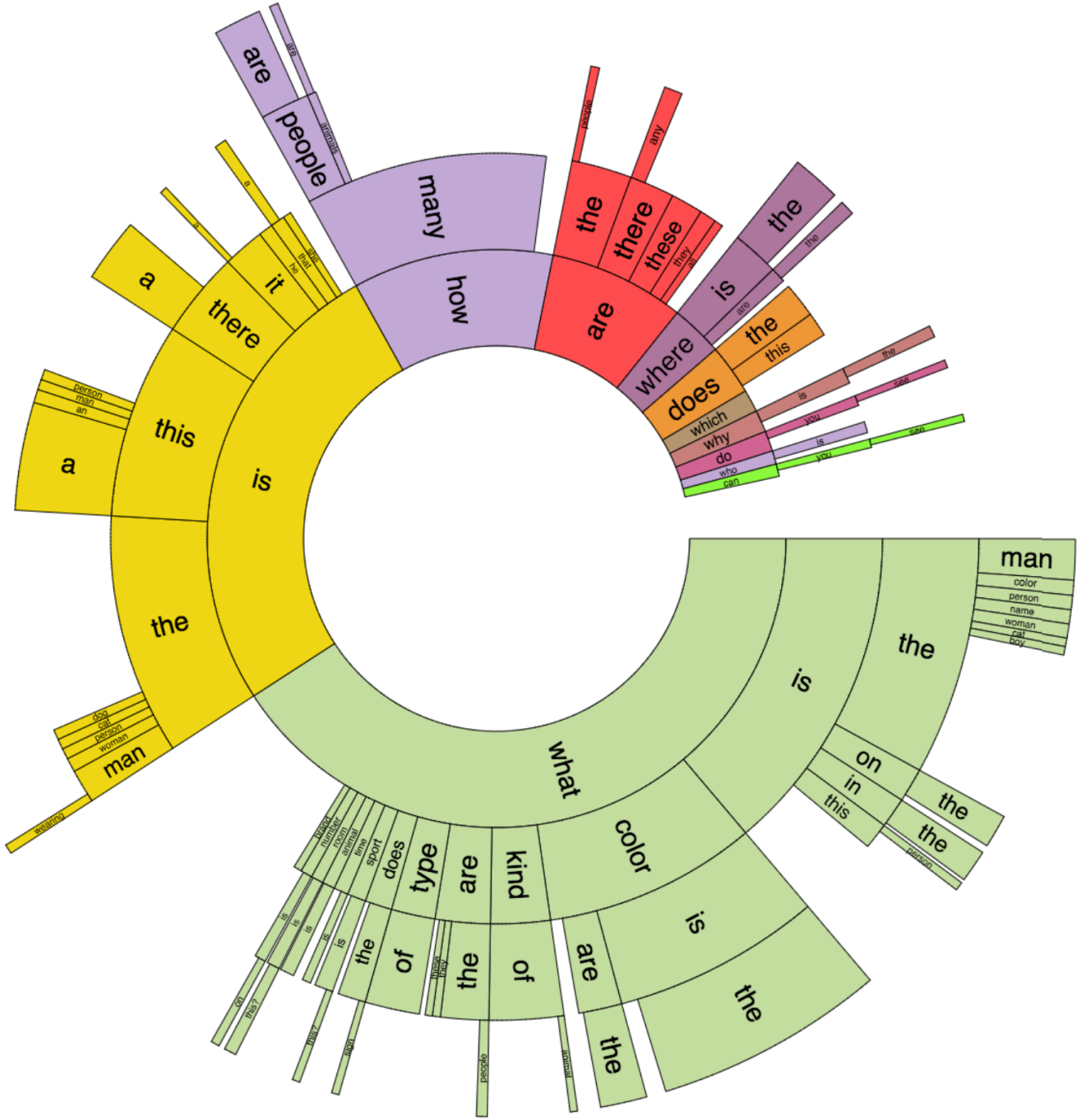


Questions

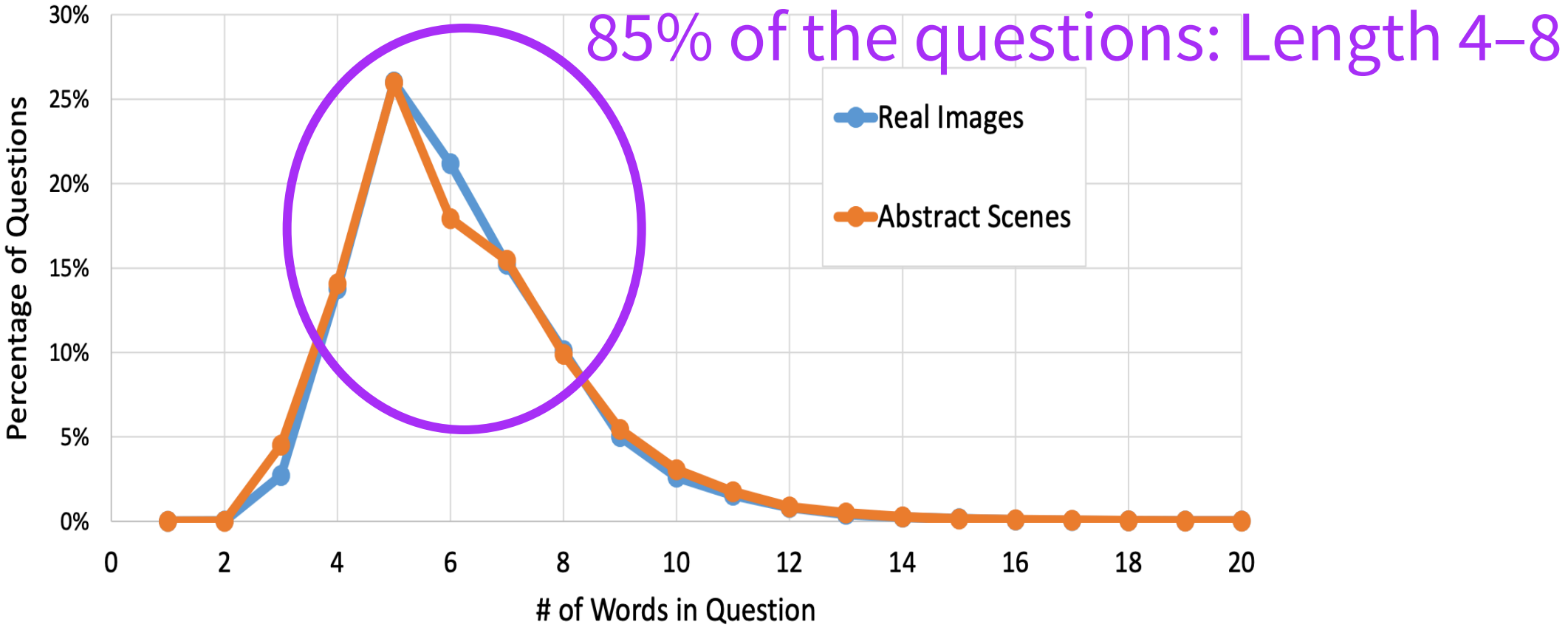
- What color is the …?
- Is this a …?
- What is the man Ving?
- How many people are …?

It’s the mirror of the answers:

- 114,000 “Is …” yes/no Qs
- 49,000 What color questions



Questions: Nothing but simple facts!



Questions: Short; no complications

What is she holding?

What color is the napkin?

Is the man wearing formal clothes?

How many people are in the picture?

Is there a street light in the photo?

——

Is there a fountain in front of the building?

Simple questions about objects and attributes fail to test compositional language understanding and reasoning

Questions: The easy direction to expand

Maybe the first place to push is more complex relational questions

And then towards more complex and interesting visual scenes

Another good direction? NLVR2

- Goals are very similar
  - Go beyond a focus on objects, properties, and a few spatial relations to considering compositional language
- Yes/No answers, but **hard**
  - Like Natural Language Inference (NLI)
- A little quirky?
  - Two picture format is a little weird
  - Fairly unnatural reasoning tasks

Both images show a silver pail being used as a flower vase

What’s still missing?

Deep understanding, coming from being able to reason in an abstract space

V Abstraction: Towards a Language of Thought

We see and reason with concepts, not visual details, 99% of the time

"Scene gists"

- A man
  - A cyclist
    - Wearing glasses, gloves, watch
- A cow
- Grassland
- Sky … clouds

V Abstraction: Towards a Language of Thought

- We use **concepts** to organize our sensory experience
- We build semantic **world models** relating concepts to represent our environment
- Used to **generalize** from given examples to new ones
- Used to draw **inferences** from facts to conclusions

Visual Genome

[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

Visual Genome

[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

Visual Genome

Region Descriptions

1. blue ski with clouds
2. man's sunglasses
3. roaming hills
4. grassy field
5. man's biking helmet
6. a gig cow bell
7. cow laying next to the man
8. man wearing Denmark racing shirt
9. man's whites biking gloves
10. man's watch
11. Big fluffy clouds in the sky
12. A large bell hanging from a cows neck
13. A copper colored cow laying in the grass with horns
14. The horns from a cow
15. A mans head who is wearing a helmet and glasses
16. A white X on red patch
17. White gloves with red and black colors
18. Many hills in the distance
19. sunglasses on a mans face
20. A man kneeling down
21. large brown cow
22. large metal bell on cow's collar
23. short horn's on cow's head
24. man wearing bicycling gear
25. silver bike helmet
26. white rimmed sunglasses
27. red and white cross on man's shoulder
28. rolling green hills
29. wrist watch on man's arm
30. white cumulus clouds in a blue sky
31. a biking helmet on head of man
32. a glove on man's hand
33. a black watch on man's wrist
34. man with knees bent on grass
35. shirt on man with flag and symbols
36. sport sunglasses on man's head
37. face of man smiling
38. a cow laying down resting on grass
39. horns on head of cow
40. cow with light brown fur
41. there are mountains in background
42. clouds in the sky
43. man is smiling
44. the shirt is red white and blue
45. he has on blue shorts
46. animal has a collar around his neck
47. the grass is green
48. Man is wearing glasses
49. there is a cross on the arm
50. this animal has horns

[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

Visual Genome Scene Graph

Objects

man

cow

helmet

watch

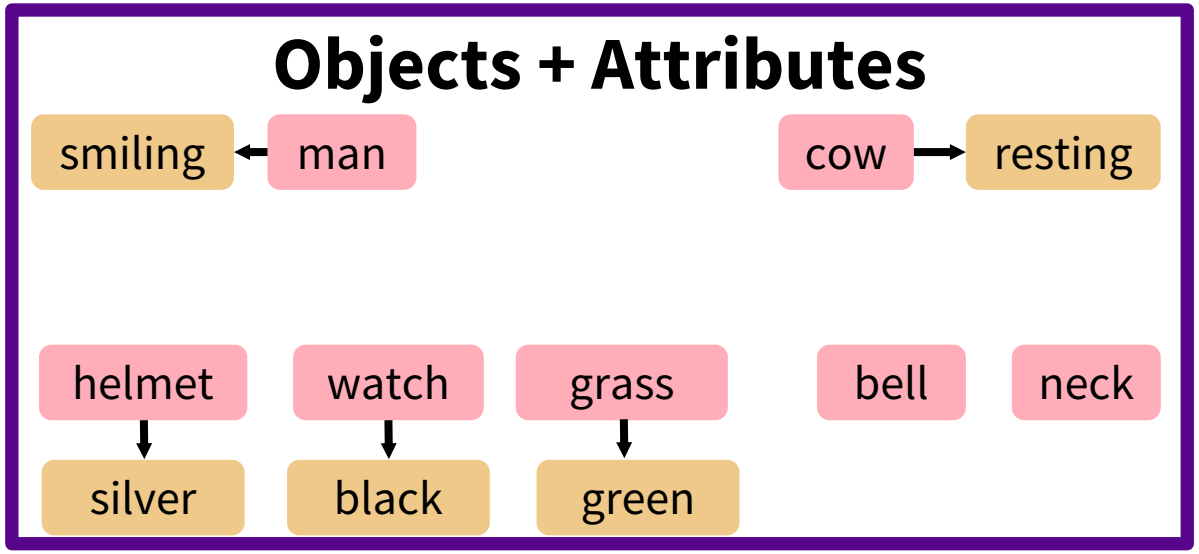
grass

bell

neck

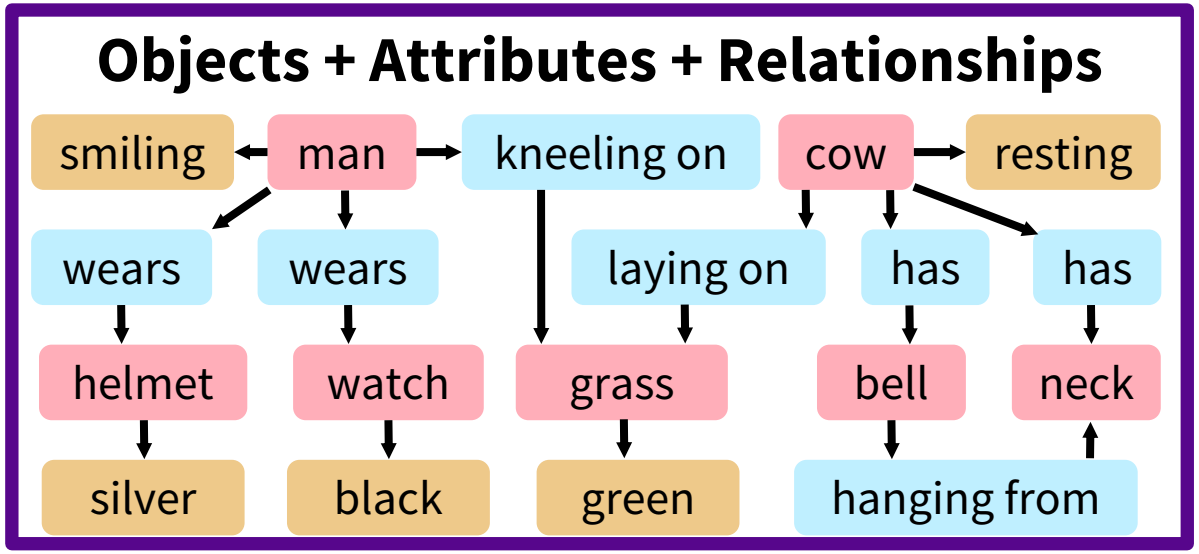
[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

Visual Genome Scene Graph



[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

Visual Genome Scene Graph



[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

The hope of deep neural models is to learn higher-level abstractions

Abstractions disentangle factors of variation, improving generalization

Content-based attention over concepts

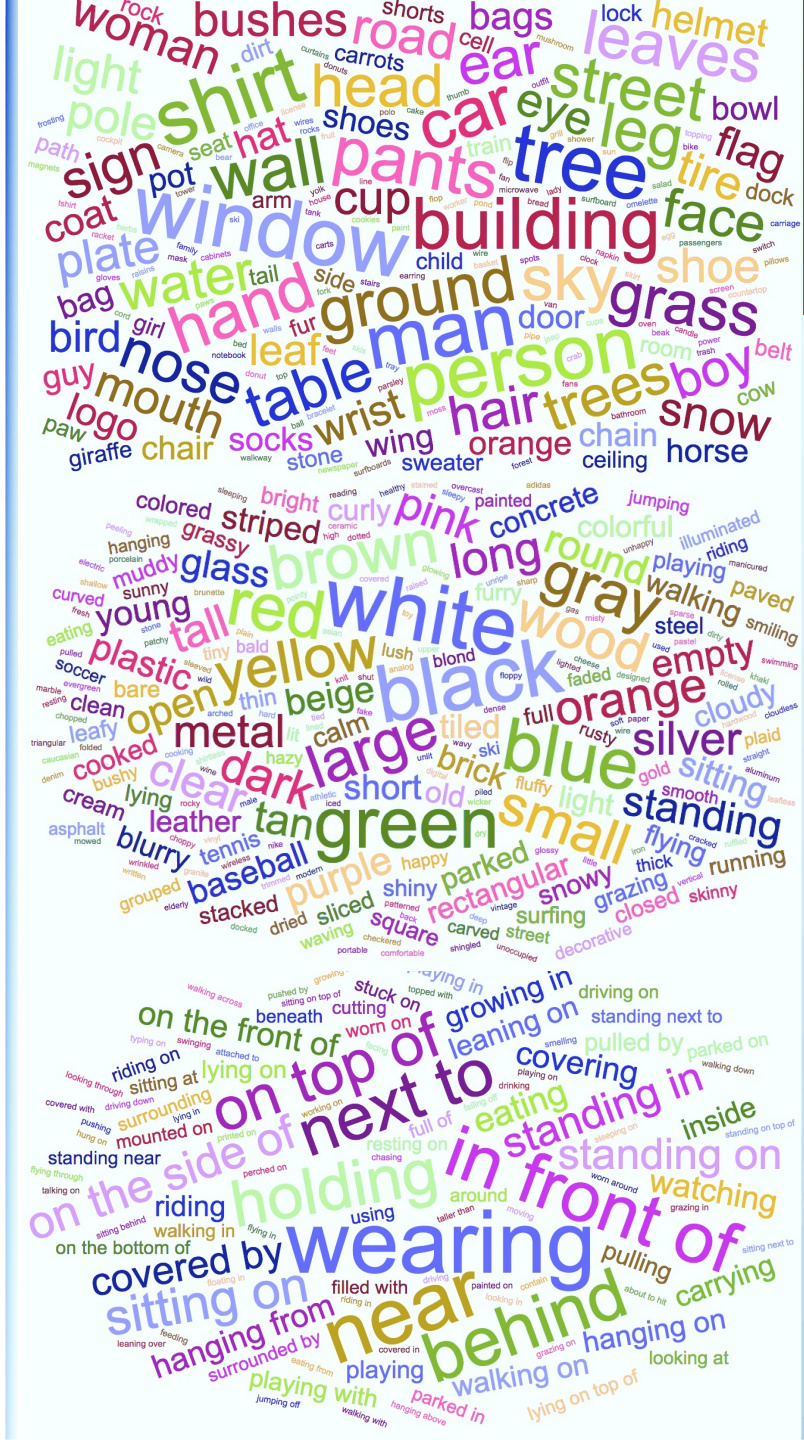
- Attention allows focus on a few elements out of a large set
- But we need attention over **concept space**, not over **pixel space**

- Cf. Yoshua Bengio's so-called "Consciousness Prior"
  - Learn a deep representation that disentangles abstract explanatory factors
  - The conscious state is then a very low-dimensional vector, an attention mechanism applied on the deep representation

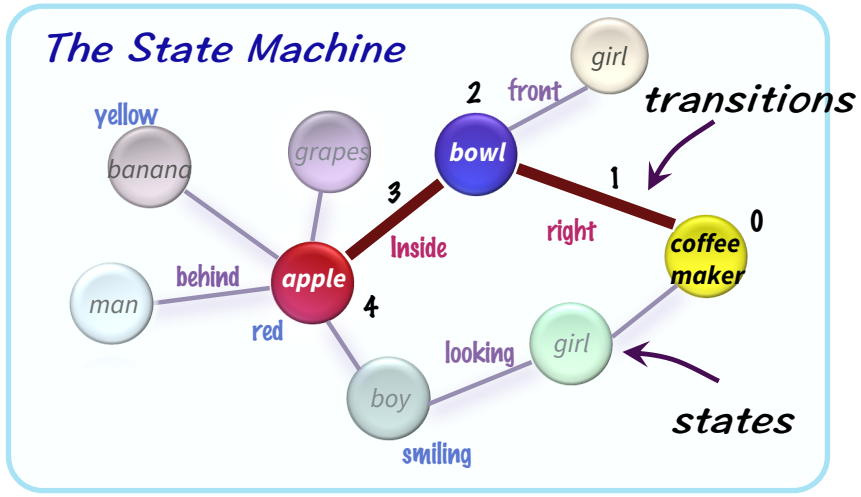
Learning by Abstraction

[Hudson and Manning submitted]

- Operate over a vocabulary of **embedded concepts**, **atomic semantic units** that represent aspects of the world (cleaned up Visual Genome ontology)
- **Translate** both **modalities** (image and question) to **"speak the same language"** of concepts
  - Everything is attention over the concept vocabulary
- **Abstract** over the raw dense features
- Inspired by **concept learning and use** in humans



Reasoning with Abstractions

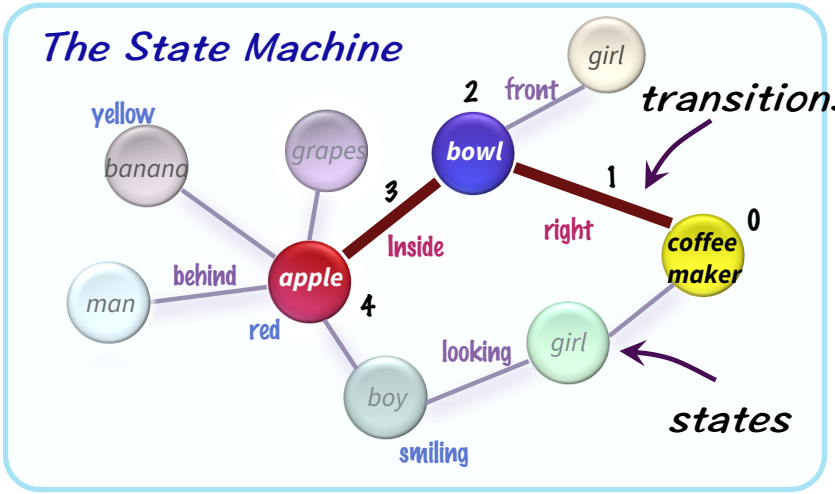


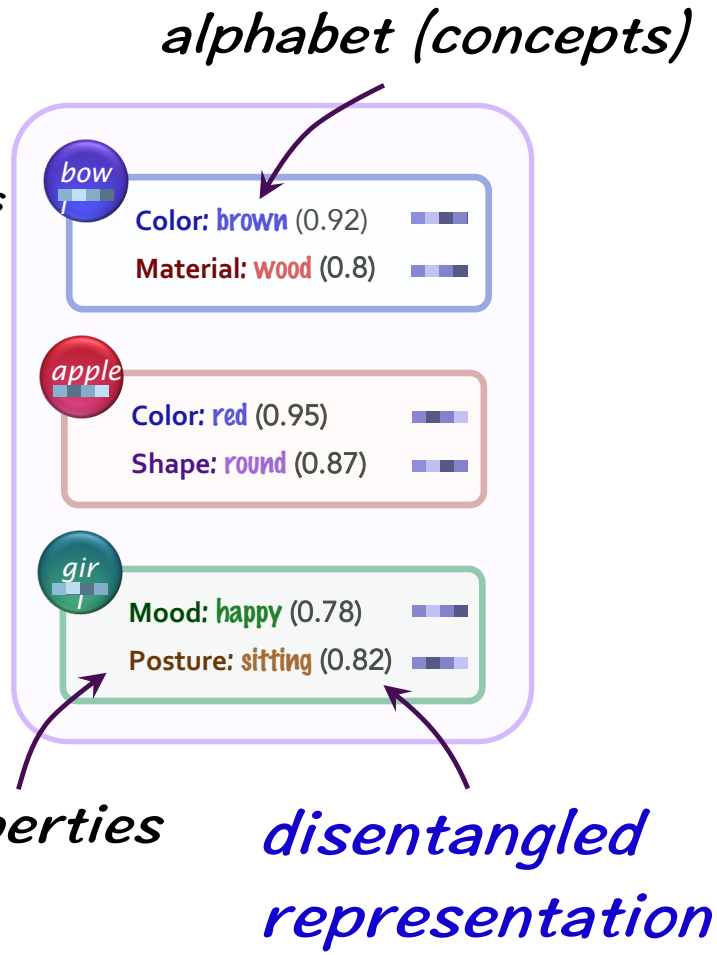
Given an **image**, we construct a **scene graph**

Treat it as a **neural state machine**, where:

- **States** correspond to ***objects***
- **Transitions** correspond to ***relations***
- States have different *(soft)* **properties** *(**attributes**)* via **attention**

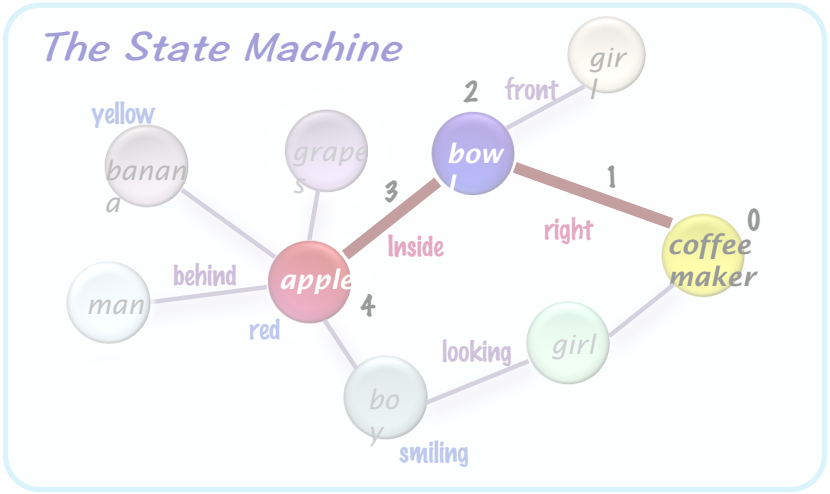
Reasoning with Abstractions

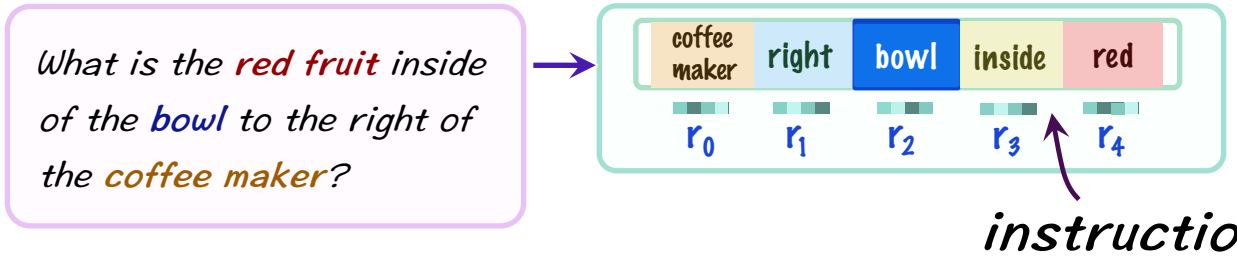




Objects are represented through a **factorized distribution** over **semantic properties** *(color, shape, material)*, defined over the **concept vocabulary**.

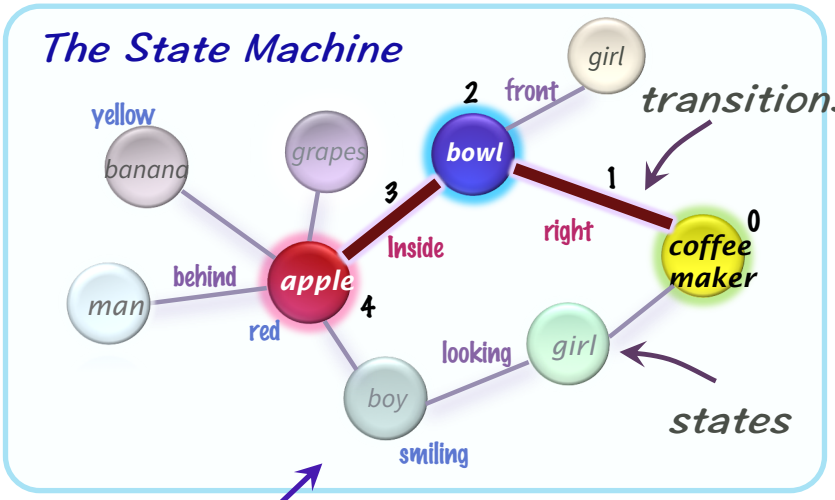
Reasoning with Abstractions

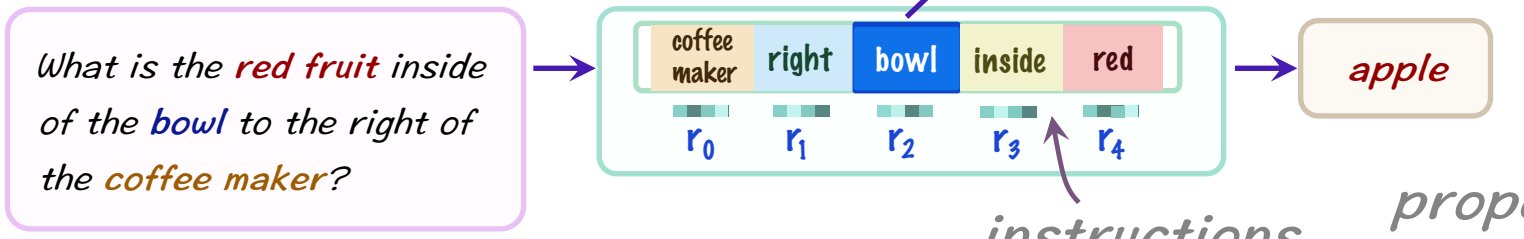




The question is translated into a **series of instructions** (with an attention-based encoder-decoder), defined over the **concepts**.

Reasoning with Abstractions





We **simulate** a computation as a **neural state machine**, feeding one **instruction** at a time and **traversing the states** until completion.

One more example

bed → left → tall → made

What is the **tall object** to the **left** of the **bed** made of?

Cabinet: wood (0.95), tall (0.92), shiny (0.86)

Bed: white (0.84), comfortable (0.91)

Lamp: yellow (0.92), on (0.74), thin (0.82)

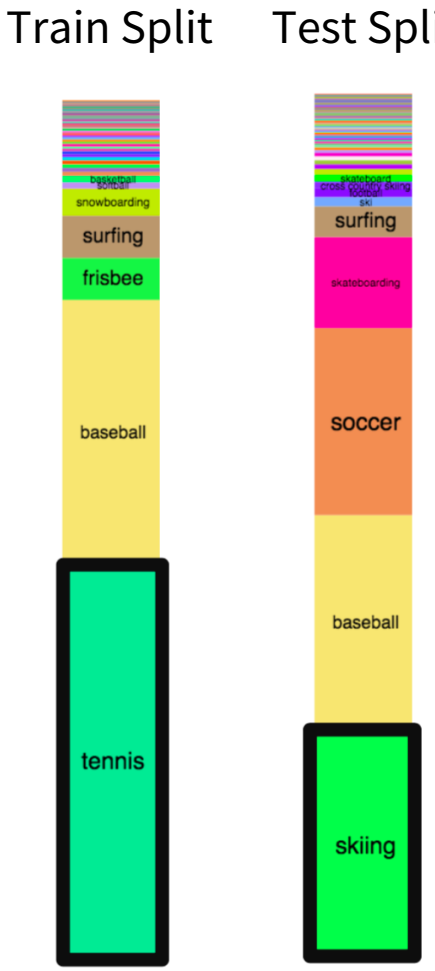
(cabinet, *left*, bed) (0.82)

(pillow, *on*, bed) (0.74)

...

→ *Wood*

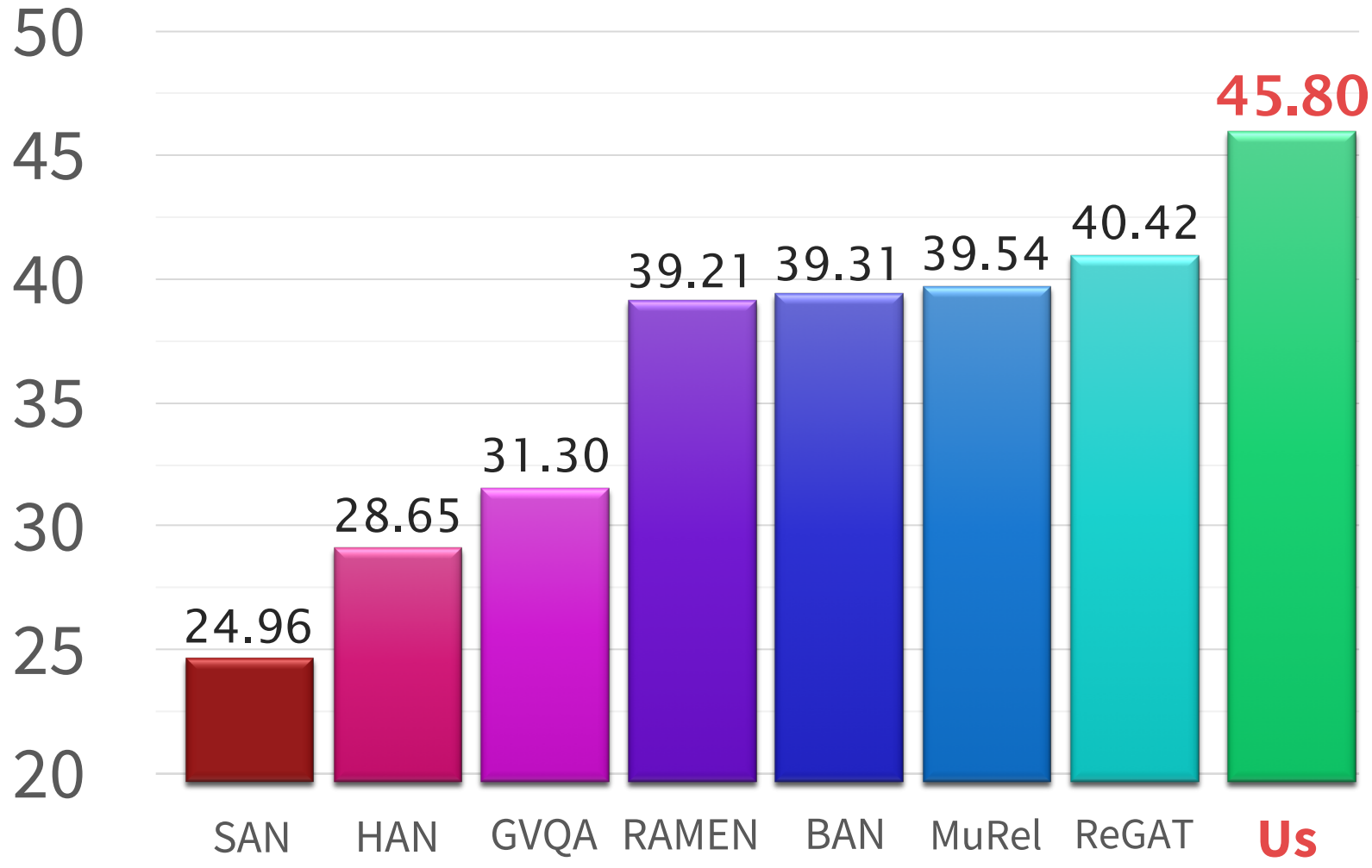
Testing Disentanglement (≈ Understanding) — VQA-CP: VQA under Changing Priors [Agrawal et al. 2017]

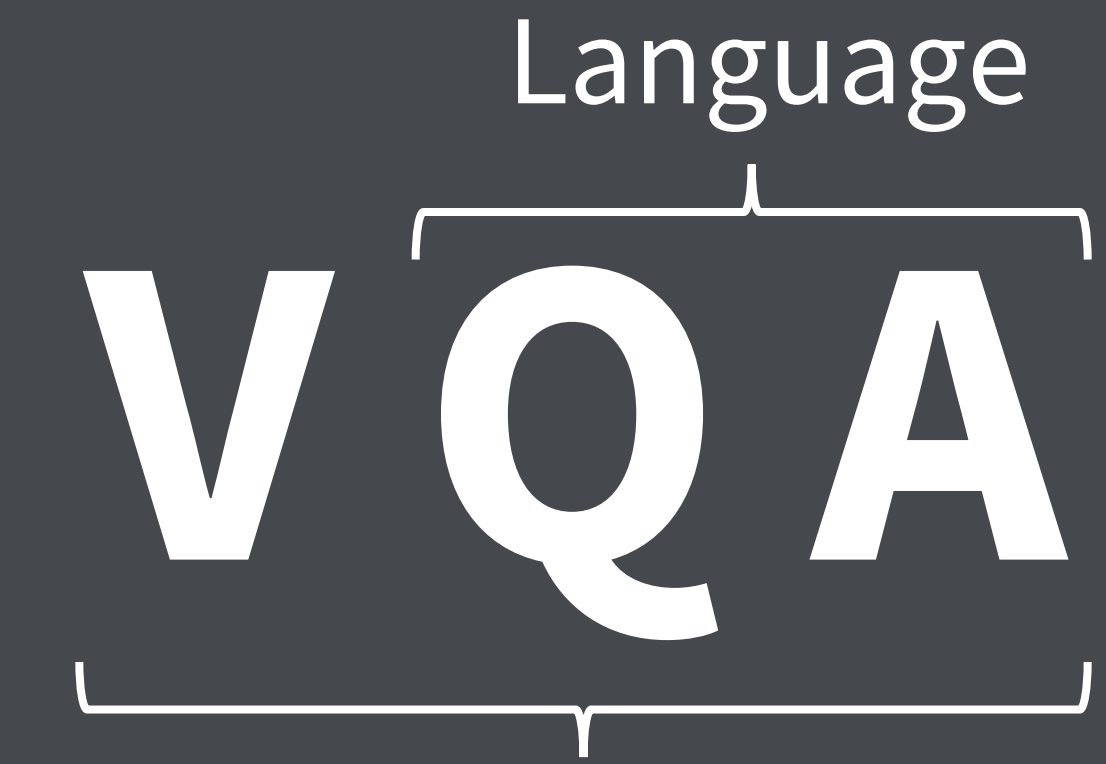


What sport …?

Model	Dataset	Overall score
d-LSTM Q + norm I	VQA v1	54.40
(Antol et al. ICCV 2015)	VQA-CP v1	23.51 –31%
NMN	VQA v1	54.83
(Andreas et al. CVPR 2016)	VQA-CP v1	29.64 –25%
SAN	VQA v1	55.86
(Yang et al. CVPR 2016)	VQA-CP v1	26.88 –29%
MCB	VQA v1	60.97
(Fukui et al. EMNLP 2016)	VQA-CP v1	34.39 –27%

Generalization of VQA-CP v2




Language
VQA
Language of Thought

We should seek tasks involving understanding and multi-step compositional reasoning

Let’s build networks that think!

By iterative attention over abstracted, disentangled concepts