The colors of the visible light spectrum[^1]

<table>
<thead>
<tr>
<th>Color</th>
<th>Wavelength interval</th>
<th>Frequency interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>~ 700–635 nm</td>
<td>~ 430–480 THz</td>
</tr>
<tr>
<td>Orange</td>
<td>~ 635–590 nm</td>
<td>~ 480–510 THz</td>
</tr>
<tr>
<td>Yellow</td>
<td>~ 590–560 nm</td>
<td>~ 510–540 THz</td>
</tr>
<tr>
<td>Green</td>
<td>~ 560–520 nm</td>
<td>~ 540–580 THz</td>
</tr>
<tr>
<td>Cyan</td>
<td>~ 520–490 nm</td>
<td>~ 580–610 THz</td>
</tr>
<tr>
<td>Blue</td>
<td>~ 490–450 nm</td>
<td>~ 610–670 THz</td>
</tr>
<tr>
<td>Violet</td>
<td>~ 450–400 nm</td>
<td>~ 670–750 THz</td>
</tr>
</tbody>
</table>

“Apples are red”

“The numbers this month are in the red”

“Red has a wavelength between 635-700nm”

...

“Pixel (1,1) has R=240, pixel (1,2) has ...”
What is Language Grounding?

‣ Tying language to non-linguistic things (e.g. a database in semantic parsing)

‣ The world only looks like a database some of the time!

‣ Some settings depend on grounding into perceptual or physical environments:

“Add the tomatoes and mix”  “Take me to the shop on the corner”

‣ Focus today: Grounding language to visual perception.
Grounding

- (Some) possible things to ground into:
(Some) possible things to ground into:

- **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...
(Some) possible things to ground into:

- **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...

- **High-level percepts**: *cat* means this type of pattern
(Some) possible things to ground into:

- **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...

- **High-level percepts**: *cat* means this type of pattern

- **Embodiment (effects on the world)**: *go left* means the robot turns left, *speed up* means increasing actuation
(Some) possible things to ground into:

- **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...
- **High-level percepts**: *cat* means this type of pattern
- **Embodiment (effects on the world)**: *go left* means the robot turns left, *speed up* means increasing actuation
- **Social (effects on others)**: polite language is correlated with longer forum discussions
Grounding

- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...
  - **High-level percepts**: *cat* means this type of pattern
  - **Embodiment (effects on the world)**: *go left* means the robot turns left, *speed up* means increasing actuation
  - **Social (effects on others)**: polite language is correlated with longer forum discussions

For a nice taxonomy, related work, and examples, see *Experience Grounds Language* [Bisk et al. 2020]
Grounding

• (Some) key problems:
  • **Representation**: matching low-level percepts to high-level language (pixels vs cat)
  • **Abstraction and Composition**: meaning as a combination of parts
  • **Alignment**: aligning parts of language and parts of the world
  • **Content Selection and Context**: what are the important parts of the environment?
  • **Balance**: it’s easy for multi-modal models to “cheat”, rely on imperfect heuristics, or ignore important parts of the input
  • **Generalization**: to novel world contexts / input combinations
CS294-43: VISION AND LANGUAGE AI SEMINAR
A Gallery of Tasks
The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

Microsoft COCO Captions: Chen et al. 2015
Visual Question Answering

What is the dog wearing?
- life jacket
- collar

How many skiers are there?
- 2
- 1

What number is on the train?
- 7907
- 8551

What is sitting in the window?
- bird
- clock

VQA 2.0: Goyal et al. 2017
Object Detection (2D)

(a) Query: “street lamp”

(b) Query: “major league logo”

(c) Query: “zebras on savanna”
1. “The chair closest to the door.”
2. “The chair under the chalkboard.”

1. “The office chair that is green.”
2. “Choose the brown office chair pushed under the desk.”

ReferIt3D: Achlioptas et al. 2020
Conditional Generation (2D)

vibrant portrait painting of Salvador Dali with a robotic half face  
a shiba inu wearing a beret and black turtleneck  
a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation  
panda mad scientist mixing sparkling chemicals, artstation  
a corgi’s head depicted as an explosion of a nebula

DALL-E 2: Ramesh et al. 2022
Conditional Generation (3D)

“Iron Man”

“Astronaut Horse”

“Colorful Crochet Candle”
“Place a clean ladle on a counter”
Why Grounded Language?

• Much language refers to the world.
• Convenient medium to communicate with machines!
• For many tasks, agents will need perceptual understanding and motor control for this interaction.
“What color is the small shiny cube?”
“What color is the small shiny cube?”
Bottom-Up & Top-Down Reasoning

Bottom-up object proposals

“What color is the small shiny cube?”
“What color is the small shiny cube?”
“What color is the small shiny cube?”

Neural Network

Can answer by focusing on single object
### Bottom-Up & Top-Down Reasoning

<table>
<thead>
<tr>
<th>Model</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: ResNet (1×1)</td>
<td>76.0</td>
<td>36.5</td>
<td>46.8</td>
<td>56.3</td>
</tr>
<tr>
<td>Ours: ResNet (14×14)</td>
<td>76.6</td>
<td>36.2</td>
<td>49.5</td>
<td>57.9</td>
</tr>
<tr>
<td>Ours: ResNet (7×7)</td>
<td>77.6</td>
<td>37.7</td>
<td>51.5</td>
<td>59.4</td>
</tr>
<tr>
<td>Ours: Up-Down</td>
<td>80.3</td>
<td>42.8</td>
<td>55.8</td>
<td>63.2</td>
</tr>
<tr>
<td>Relative Improvement</td>
<td>3%</td>
<td>14%</td>
<td>8%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Provides inductive bias in both directions!
Bottom-Up
**Vision**: David Marr 1982

**Viewer centred**

- **Input Image**
  - Perceived intensities
  - Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries.

- **Primal Sketch**
  - Local surface orientation and discontinuities in depth and in surface orientation

**Object centred**

- **3-D Model Representation**
  - 3-D models hierarchically organised in terms of surface and volumetric primitives
Intrinsic Images

(a) ORIGINAL SCENE

(b) DISTANCE

(c) REFLECTANCE

(d) ORIENTATION (VECTOR)

(e) ILLUMINATION

Recovering Intrinsic Scene Characteristics from Images: Barrow and Tenenbaum 1978
“Solved” Perception

**Question:** Where is the object outlined in red?

**Answer:** The object outlined in red is

- left of
- right of ✓
- above
- below
- in front of
- behind
- inside of
- on
- under
- across from
“Solved” Perception

Question: Where is the object outlined in red?
Answer: The object outlined in red is

Task: Describe target object unambiguously

A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010
“Solved” Perception

Question: Where is the object outlined in red?
Answer: The object outlined in red is right of O1, above O3.

Relationships between objects known

left of
right of
above
below
in front of
behind
inside of
on
under
across from

A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010
“Solved” Perception

Problem reduced to pragmatic reasoning

\[ S(L)(o) = \arg\max_w p_L(o|w) \]

A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010
“Solved” Perception

“Go to the last butterfly on the right”

[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]
“Solved” Perception

“Go to the last butterfly on the right”

What annotators see

[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]
“Go to the last butterfly on the right”

What agent sees

[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]
“Solved” Perception

“Go to the last butterfly on the right”

Reduced to structured prediction problem

[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]
“Solved” Perception

• **Pro:** In early days of vision and language, assuming sub-problems provided traction.

• **Con:** Strong assumptions that don’t hold in real world.
Intermediate Representations

Input Image

1) Object(s)/Stuff
   a) dog
   b) person
   c) sofa

2) Attributes
   a) brown 0.01
   b) striped 0.16
   c) furry .26
   d) wooden 2
   e) feathered .06
   ...

3) Prepositions
   a) near(a,b) 1
   b) near(b,a) 1
   c) against(a,b) 1
   d) against(b,a) 1
   e) beside(a,b) 24
   f) beside(b,a) .17
   ...

4) Constructed CRF

5) Predicted Labeling
   - <null_person_b>, against, <brown_sofa_c>
   - <null_dog_a>, near, <null_person_b>
   - <null_dog_a>, beside, <brown_sofa_c>

6) Generated Sentences
   This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

BabyTalk: Kulkarni et al. 2013
Intermediate Representations

1) Object(s)/Stuff
- a) dog
- b) person
- c) sofa

2) Attributes
- brown 0.01
- striped 0.16
- furry 0.26
- wooden 0.2
- feathered 0.3

3) Prepositions
- near(a,b) 1
- near(b,a) 1
- against(b,a) 0.11
- against(a,b) 0.04

4) Constructed CRF

5) Predicted Labeling
- <<null_person_b>>, against, <<brown_sofa_c>>
- <<null_dog_a>>, <<null_person_b>>
- <<null_dog_a>>, beside, <<brown_sofa_c>>

6) Generated Sentences
This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

Extract regions of interest using pretrained detector

BabyTalk: Kulkarni et al. 2013
Intermediate Representations

Classifiers score attributes for each region and relationships across them

BabyTalk: Kulkarni et al. 2013
Intermediate Representations

Use CRF to predict highest likelihood assignment of labels

4) Constructed CRF

5) Predicted Labeling

<<null,person_b>>, against, <<brown, sofa_c>>
<<null,dog_a>>, near, <<null,person_b>>
<<null,dog_a>>, beside, <<brown, sofa_c>>

BabyTalk: Kulkarni et al. 2013
Intermediate Representations

6) Generated Sentences
This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

5) Predicted Labeling

- <<null,person_b>,against,<brown,sofa_c>>
- <<null,dog_a>,near,<null,person_b>>
- <<null,dog_a>,beside,<brown,sofa_c>>

Generate caption conditioned on labels

BabyTalk: Kulkarni et al. 2013
Intermediate Representations

Language model never sees pixels!

Input Image

1) Object(s)/Stuff
   a) dog
   b) person
   c) sofa

2) Attributes
   - brown 0.01
   - striped 0.16
   - furry .28
   - wooden .2
   - feathered .06
   - ... (and similar attributes for other objects)

3) Prepositions
   - near(a,b) 1
   - near(b,a) 1
   - against(a,b) .11
   - against(b,a) .04
   - beside(a,b) .24
   - beside(b,a) .17
   - ... (and similar prepositions for other pairs)

4) Constructed CRF

5) Predicted Labeling
   - <null,person_b>, against, <brown,sofa_c>
   - <null,dog_a>, near, <null,person_b>
   - <null,dog_a>, beside, <brown,sofa_c>

6) Generated Sentences
   This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

BabyTalk: Kulkarni et al. 2013
Intermediate Representations

This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.

There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.

There are one cow and one sky. The golden cow is by the blue sky.

There are one dining table, one chair and two windows. The wooden dining table is by the wooden chair, and against the first window, and against the second white window. The wooden chair is by the first window, and by the second white window. The first window is by the second white window.

Here we see one person and one train. The black person is by the train.

This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.

Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.

Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.

This is a picture of two dogs. The first dog is near the second furry dog.

This is a photograph of two buses. The first rectangular bus is near the second rectangular bus.

BabyTalk: Kulkarni et al. 2013
Intermediate Representations

What is the red fruit inside the bowl to the right of the coffee maker?

The State Machine

- States: apple, bowl, man, boy, girl, coffee maker
- Alphabet (concepts): bowl, apple, red, happy
- Properties: Color: brown (0.92), Material: wood (0.8), Color: red (0.95), Shape: round (0.87)
- Instructions: r0, r1, r2, r3, r4
- Disentangled representation:

Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019
Intermediate Representations

The State Machine

Generate scene graph from image

Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019
Intermediate Representations

Graph vocabulary predefined
Intermediate Representations

Transform question into program traversing graph for answer

What is the red fruit inside the bowl to the right of the coffee maker?

Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019
Intermediate Representations

Answer by executing program in state machine
Intermediate Representations

The State Machine

- Allows language reasoning to occur solely within abstract structure

**Instructions:**

- **What is the red fruit inside the bowl to the right of the coffee maker?**

**States:**
- apple
- bowl
- coffee maker

**Transitions:**
- inside
- on top
- right
- on top
## Table 4: GQA generalization

<table>
<thead>
<tr>
<th>Model</th>
<th>Content</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Prior</td>
<td>8.51</td>
<td>14.64</td>
</tr>
<tr>
<td>Local Prior</td>
<td>12.14</td>
<td>18.21</td>
</tr>
<tr>
<td>Vision</td>
<td>17.51</td>
<td>18.68</td>
</tr>
<tr>
<td>Language</td>
<td>21.14</td>
<td>32.88</td>
</tr>
<tr>
<td>Lang+Vis</td>
<td>24.95</td>
<td>36.51</td>
</tr>
<tr>
<td>BottomUp [5]</td>
<td>29.72</td>
<td>41.83</td>
</tr>
<tr>
<td>MAC [40]</td>
<td>31.12</td>
<td>47.27</td>
</tr>
<tr>
<td><strong>NSM</strong></td>
<td><strong>40.24</strong></td>
<td><strong>55.72</strong></td>
</tr>
</tbody>
</table>
a store front that has the word 'dall-e' written on it. a store front that has the word 'dall-e' written on it. a store front that has the word 'dall-e' written on it. dall-e store front.
Intermediate Representations

**Step 1**
Learn Proto-linguistic Code Book
Intermediate Representations

Step 1
Learn Proto-linguistic Code Book

Neural Discrete Representation Learning: van Oord et al. 2017

DALL-E 1: Ramesh et al. 2021
Intermediate Representations

**Step 2**
Learn Joint Language and Code Distribution

"A kitten with a pink background"

```
  1  5  2  6
  9 13 10 14
  3  7  4  8
 11 15 12 16
```
Intermediate Representations

**Step 2**
Learn Joint Language and Code Distribution

"A kitten with a pink background"

<table>
<thead>
<tr>
<th>1</th>
<th>5</th>
<th>2</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>13</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

12B Param Transformer

- "red"
- "cat"
- 32
- 16
- 72
- 91
- 8

Generating Long Sequences with Sparse Transformers: Child et al. 2019

DALL-E 1: Ramesh et al. 2021
Intermediate Representations

**Step 2**
Learn Joint Language and Code Distribution

“*A kitten with a pink background*”

```
   1 5 2 6
  9 13 10 14
  3 7 4 8
  11 15 12 16
```

Reduced to language modeling problem!

```
"red"  "cat"  32  16  72  91  8
"The"  "red"  "cat"  32  16  72  91
```

Generating Long Sequences with Sparse Transformers: Child et al. 2019
an x-ray of a capybara sitting in a forest
Anchoring to 3D

“The goal of an image understanding system is to transform two-dimensional data into a representation of the three-dimensional spatio-temporal world”
“Place a clean ladle on a counter”
Anchoring to 3D

A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021

A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021
Anchoring to 3D

Perceptual module extracts 2D semantic segmentation and depth maps
Anchoring to 3D

Maps integrated into persistent 3D map

A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021
Controller operates exclusively over 3D map
Anchoring to 3D

<table>
<thead>
<tr>
<th>Method</th>
<th>Validation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seen</td>
<td>Unseen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SR</td>
<td>GC</td>
<td>SR</td>
</tr>
<tr>
<td>HLSM</td>
<td>29.6</td>
<td>38.8</td>
<td>18.3</td>
</tr>
<tr>
<td>+ gt depth</td>
<td>29.6</td>
<td>40.5</td>
<td>20.1</td>
</tr>
<tr>
<td>+ gt depth, gt seg.</td>
<td>40.7</td>
<td>50.4</td>
<td>40.2</td>
</tr>
<tr>
<td>+ gt seg.</td>
<td>36.2</td>
<td>47.0</td>
<td>34.7</td>
</tr>
<tr>
<td>w/o language enc.</td>
<td>0.9</td>
<td>8.6</td>
<td>0.2</td>
</tr>
<tr>
<td>w/o subg. hist. enc.</td>
<td>29.4</td>
<td>38.5</td>
<td>16.6</td>
</tr>
<tr>
<td>w/o state repr enc.</td>
<td>30.0</td>
<td>40.6</td>
<td>18.9</td>
</tr>
</tbody>
</table>

3D Map useful for improving performance
Anchoring to 3D

<table>
<thead>
<tr>
<th>Method</th>
<th>Validation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Seen</td>
<td>Unseen</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SR</td>
<td>GC</td>
<td>SR</td>
</tr>
<tr>
<td>HLSM</td>
<td></td>
<td>29.6</td>
<td>38.8</td>
<td>18.3</td>
</tr>
<tr>
<td>+ gt depth</td>
<td></td>
<td>29.6</td>
<td>40.5</td>
<td>20.1</td>
</tr>
<tr>
<td>+ gt depth, gt seg.</td>
<td></td>
<td>40.7</td>
<td>50.4</td>
<td>40.2</td>
</tr>
<tr>
<td>+ gt seg.</td>
<td></td>
<td>36.2</td>
<td>47.0</td>
<td>34.7</td>
</tr>
<tr>
<td>w/o language enc.</td>
<td></td>
<td>0.9</td>
<td>8.6</td>
<td>0.2</td>
</tr>
<tr>
<td>w/o subg. hist. enc.</td>
<td></td>
<td>29.4</td>
<td>38.5</td>
<td>16.6</td>
</tr>
<tr>
<td>w/o state repr enc.</td>
<td></td>
<td>30.0</td>
<td>40.6</td>
<td>18.9</td>
</tr>
</tbody>
</table>

However, benefits held back by cascading errors
Anchoring to 3D

2D-Only Language Grounding

"The swivel chair with 6 wheels"

Voxel Prediction

"The swivel chair with 6 wheels"

Voxel-informed Language Grounding

Voxel-informed Language Grounding: Corona et al. 2022
Anchoring to 3D

Standard approaches perform grounding from 2D only

2D-Only Language Grounding

“The swivel chair with 6 wheels”
Anchoring to 3D

Can supplement with predicted geometry from 3D reconstruction model.
Anchoring to 3D

Improves performance over 2D-only approaches

<table>
<thead>
<tr>
<th>Model</th>
<th>Visual</th>
<th>Blind</th>
<th>All</th>
<th>Visual</th>
<th>Blind</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLBERT</td>
<td>89.5</td>
<td>76.6</td>
<td>83.1</td>
<td>80.2</td>
<td>73.0</td>
<td>76.6</td>
</tr>
<tr>
<td>MATCH</td>
<td>89.2 (0.9)</td>
<td>75.2 (0.7)</td>
<td>82.2 (0.4)</td>
<td>83.9 (0.5)</td>
<td>68.7 (0.9)</td>
<td>76.5 (0.5)</td>
</tr>
<tr>
<td>MATCH*</td>
<td>90.6 (0.4)</td>
<td>75.7 (1.2)</td>
<td>83.2 (0.8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LAGOR</td>
<td>89.8 (0.4)</td>
<td>75.3 (0.7)</td>
<td>82.6 (0.4)</td>
<td>84.3 (0.4)</td>
<td>69.4 (0.5)</td>
<td>77.0 (0.5)</td>
</tr>
<tr>
<td>LAGOR*</td>
<td>89.8 (0.5)</td>
<td>75.0 (0.4)</td>
<td>82.5 (0.1)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VLG (Ours)</td>
<td><strong>91.2 (0.4)</strong></td>
<td><strong>78.4</strong> (0.7)</td>
<td><strong>84.9</strong> (0.3)</td>
<td><strong>86.0</strong></td>
<td>71.7</td>
<td><strong>79.0</strong></td>
</tr>
</tbody>
</table>

Voxel-informed Language Grounding: Corona et al. 2022
Bottom-Up Takeaways

• Grounding to intermediate representations more tractable than grounding directly to pixels.
• Constrains the space of things to ground to.
• **Limitation:**
  – May suffer from cascading error.
  – Not always informed by language.
“What color is the small shiny cube?”
Noun

- S: (n) **wordnet** (any of the machine-readable lexical databases modeled after the Princeton WordNet)
- S: (n) **WordNet**, [Princeton WordNet](#) (a machine-readable lexical database organized by meanings; developed at Princeton University)
Modular Systems

“Is there a red sphere above a circle?”
Q: What is the shape of the red object left of the sphere?
Parse question into program in Domain Specific Language (DSL)

Q: What is the shape of the red object left of the sphere?

Semantic Parsing (Candidate Interpretations)

- Query(Shape, Filter(Left, Filter(Shape))), Filter(Sphere))
- Query(Shape, Filter(Left, Filter(Shape))), Filter(Left), Filter(Sphere))
- Query(AERelate(Shape, Filter(Left, Filter(Shape))), Filter(Sphere))
- …..

The Neuro-Symbolic Concept Learner: Mao et al. 2020
Modular Systems

Q: What is the shape of the red object to the left of the sphere?

Extract objects ROIs and embed them
Abstract concepts in DSL have embeddings
All operations deterministic and pre-defined!
Modular Systems

I. Visual Reasoning Question

Exist(Filter(red))
Score=0.1

II. Metaconcept Question (Text-Only)

MetaVerify(red, green, same_kind)
Score=0.9 (A: Yes)

a. Perception Module

Object Detection
Feature Extraction
Obj. 1
Obj. 2

Q: Is there any red object?
P: Exist(Filter(red))

Q: Do red and green describe the same property of objects?
P: MetaVerify(red, green, same_kind)

 Parsing

b. Semantic Parsing Module

c. Neuro-Symbolic Reasoning

Score=0.9 (A: Yes)
Regularize concept embeddings on “meta question” task
Modular Systems

“block” == “square”

<table>
<thead>
<tr>
<th></th>
<th>GRU-CNN</th>
<th>MAC</th>
<th>NS-CL</th>
<th>VCML</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEVR</td>
<td>50.0±0.0</td>
<td>68.7±3.8</td>
<td>80.2±3.1</td>
<td>94.1±4.6</td>
</tr>
<tr>
<td>GQA</td>
<td>50.0±0.0</td>
<td>49.5±0.2</td>
<td>49.3±0.6</td>
<td>50.5±0.1</td>
</tr>
</tbody>
</table>

Learning synonyms helps zero-shot generalization
Modular Systems

== "purple" + "square"

<table>
<thead>
<tr>
<th></th>
<th>GRU-CNN</th>
<th>MAC</th>
<th>NS-CL</th>
<th>VCML</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEVR-200</td>
<td>50.0±0.0</td>
<td>94.2±3.3</td>
<td>98.5±0.3</td>
<td>98.9±0.2</td>
</tr>
<tr>
<td>CLEVR-20</td>
<td>50.0±0.0</td>
<td>79.7±2.6</td>
<td>95.7±0.0</td>
<td>95.1±1.6</td>
</tr>
</tbody>
</table>

Learning same kind helps compositional generalization
“All the dogs are black.”

Modular Systems

Obtaining Faithful Interpretations from Compositional Neural Networks: Subramanian et al. 2020
Language as Signal for Abstractions

Learning with Latent Language: Andreas et al. 2017
Language as Signal for Abstractions

Available at Training

true
true
true
true

A white shape is left of a yellow semicircle

true

Learning with Latent Language: Andreas et al. 2017
Language as Signal for Abstractions

**Meta** (Snell et al., 2017)

- Support
- Query

**LSL** (ours)

- Auxiliary training (discard at test)

**L3** (Andreas et al., 2018)

- Support
- Query

Shaping Visual Representations with Language for Few-Shot Classification: Mu et al. 2020
Language as Signal for Abstractions

Prototype network (no language)
Language as Signal for Abstractions

Auxiliary summarization task
<table>
<thead>
<tr>
<th></th>
<th>ShapeWorld</th>
<th>Birds ($D = 20$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta</td>
<td>60.59 ± 1.07</td>
<td>57.97 ± 0.96</td>
</tr>
<tr>
<td>L3</td>
<td>66.60 ± 1.18</td>
<td>53.96 ± 1.06</td>
</tr>
<tr>
<td>LSL</td>
<td>67.29 ± 1.03</td>
<td>61.24 ± 0.96</td>
</tr>
</tbody>
</table>

Shaping Visual Representations with Language for Few-Shot Classification: Mu et al. 2020
Top-Down Takeaways

- Language provides labels for supervised learning of perceptual systems.
- Can provide powerful inductive biases in computational structure if known.
- Serves as signal for useful perceptual abstractions to learn either as bottleneck or auxiliary signal.
“What color is the small shiny cube?”
Bottom-Up & Top-Down Reasoning

“What color is the small shiny cube?”
Extra Slides
Open-Set Models

Models which leverage the open-vocabulary of language to enjoy a practically open set of labels!
Open-Set Models

(1) Contrastive pre-training

Pepper the aussie pup

Text Encoder

T_1 T_2 T_3 ... T_N

I_1 \cdot T_1 I_1 \cdot T_2 I_1 \cdot T_3 ... I_1 \cdot T_N

I_2 \cdot T_1 I_2 \cdot T_2 I_2 \cdot T_3 ... I_2 \cdot T_N

I_3 \cdot T_1 I_3 \cdot T_2 I_3 \cdot T_3 ... I_3 \cdot T_N

... ... ... ... ...

I_N \cdot T_1 I_N \cdot T_2 I_N \cdot T_3 ... I_N \cdot T_N

Image Encoder

(2) Create dataset classifier from label text

Text Encoder

A photo of a {object}.

plane

car

dog

bird

(3) Use for zero-shot prediction

Text Encoder

T_1 T_2 T_3 ... T_N

Image Encoder

I_1 \cdot T_1 I_1 \cdot T_2 I_1 \cdot T_3 ... I_1 \cdot T_N

A photo of a dog.

CLIP: Radford et al. 2021
Open-Set Models

(1) Contrastive pre-training

Pepper the aussie pup

Text Encoder

Image Encoder

Encode text and image into vectors

(2) Create dataset classifier from label text

A photo of a [object].

Text Encoder

(3) Use for zero-shot prediction

CLIP: Radford et al. 2021
Open-Set Models

Optimize compatibility with contrastive loss

CLIP: Radford et al. 2021
Open-Set Models

Classification dataset created with templated prompts

CLIP: Radford et al. 2021
Normalize compatibility scores to get zero-shot classifier!
Open-Set Models

Zero-Shot CLIP vs. Linear Probe on ResNet50

- StanfordCars: +28.9
- Country211: +23.2
- Food101: +22.5
- Kinetics700: +14.5
- SST2: +12.4
- SUN397: +7.8
- UCF101: +7.7
- HatefulMemes: +6.7
- CIFAR10: +3.9
- CIFAR100: +3.0
- STL10: +3.0
- FER2013: +2.8
- Caltech101: +2.0
- ImageNet: +1.9
- OxfordPets: +1.1
- PascalVOC2007: +0.5

CLIP: Radford et al. 2021
Open-Set Models

“A cat with white paws jumps over a fence in front of a yellow tree”

MDETR: Kamath et al. 2021
Open-Set Models

Extract image features with pre-trained CNN

MDETR: Kamath et al. 2021
Open-Set Models

Extract language features using pre-trained LM

“A cat with white paws jumps over a fence in front of a yellow tree”
Pass together through transformer
Open-Set Models

Learned embedding “queries” tied to tokens in input text/image regions

MDETR: Kamath et al. 2021
Open-Set Models

(a) “one small boy climbing a pole with the help of another boy on the ground”

(b) “A man talking on his cellphone next to a jewelry store”

MDETR: Kamath et al. 2021
Open-Set Models

Input Image → Image Encoder → $\mathbf{H}^I$ → Text Encoder → $\mathbf{H}$ → Spatial Regularization Blocks → Output

Input Label Set:
- people, tennis racket, tree, sand, other

Language-Driven Semantic Segmentation: Li et al. 2022
Open-Set Models

Extract ~per-pixel features using transformer

Language-Driven Semantic Segmentation: Li et al. 2022
Open-Set Models

Extract CLIP text embeddings for each label
Open-Set Models

Upsample and predict most compatible label per pixel

Language-Driven Semantic Segmentation: Li et al. 2022
Open-Set Models

Language-Driven Semantic Segmentation: Li et al. 2022
Open-Set Models

- **cat, grass, plant, stone**
- **furry, plant, stone**

**Input Image**

**LSeg Output A**

**LSeg Output B**

Language-Driven Semantic Segmentation: Li et al. 2022
Open-Set Models

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022
Tokenize image into patches and embed

Open-Set Models
Each layer has learnable “group tokens” that will be used to cluster image patches.
Image patches “assigned” to group tokens using transformer self-attention scores
Open-Set Models

Patches for each group aggregated through mean-pool
Repeat until single embedding is left

**GroupViT:** Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022
Open-Set Models

Optimize compatibility with text caption using contrastive loss

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022
Open-Set Models
Open-Set Models

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022
Open-Set Models

Stage 1
Group 5
“eye”

Stage 1
Group 36
“limb”

Stage 2
Group 6
“grass”

Stage 2
Group 4
“body”

Stage 2
Group 7
“face”

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022
Bias in Vision and Language Models

Wrong

Baseline: A man sitting at a desk with a laptop computer.

Our Model: A woman sitting in front of a laptop computer.

Right for the Right Reasons

Baseline: A man holding a tennis racquet on a tennis court.

Our Model: A man holding a tennis racquet on a tennis court.

Right for the Wrong Reasons

Women also Snowboard: Overcoming Bias in Caption Models: Burns et al. 2019
Bias in Vision and Language Models

$I = \quad \text{CNN} \quad \rightarrow \quad A \quad \rightarrow \quad \text{Caption Correctness Loss}$

$I' = \quad \text{CNN} \quad \rightarrow \quad A \quad \rightarrow \quad ? \quad \rightarrow \quad \text{Appearance Confusion Loss}$
<table>
<thead>
<tr>
<th>Category</th>
<th>Black</th>
<th>White</th>
<th>Indian</th>
<th>Latino</th>
<th>Middle Eastern</th>
<th>Southeast Asian</th>
<th>East Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime-related Categories</td>
<td>16.4</td>
<td>24.9</td>
<td>24.4</td>
<td>10.8</td>
<td>19.7</td>
<td>4.4</td>
<td>1.3</td>
</tr>
<tr>
<td>Non-human Categories</td>
<td>14.4</td>
<td>5.5</td>
<td>7.6</td>
<td>3.7</td>
<td>2.0</td>
<td>1.9</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Bias in Vision and Language Models

Neurons work
Bias in Vision and Language Models

Prompt: a photo of a personal assistant;
Date: April 1, 2022

DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022
Bias in Vision and Language Models

Prompt: lawyer;
Date: April 6, 2022

DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022
Prompt: nurse;
Date: April 6, 2022
Bias in Vision and Language Models

Prompt: a builder; Date: April 6, 2022