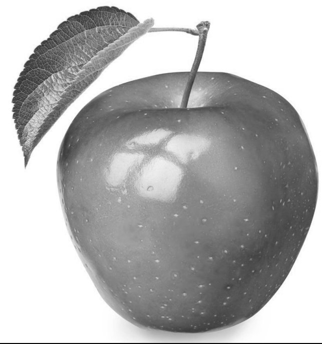
Vision and Language

Rodolfo (Rudy) Corona

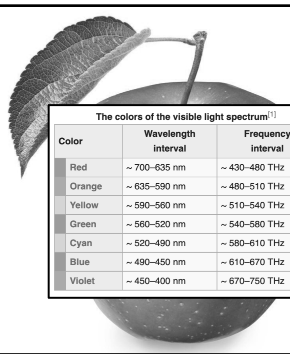
with thanks to Daniel Fried

CS 288, 4/12/2022

1



2



The colors of the visible light spectrum[1]		
Color	Wavelength interval	Frequency interval
Red	~ 700–635 nm	~ 430–480 THz
Orange	~ 635–590 nm	~ 480–510 THz
Yellow	~ 590–560 nm	~ 510–540 THz
Green	~ 560–520 nm	~ 540–580 THz
Cyan	~ 520–490 nm	~ 580–610 THz
Blue	~ 490–450 nm	~ 610–670 THz
Violet	~ 450–400 nm	~ 670–750 THz

Source: "Color" in Wikipedia

3

4

"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 …"

5

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*.

6

Grounding

- (Some) possible things to ground into:

7

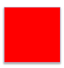
Grounding

- (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…

8

Grounding

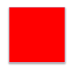
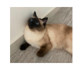
- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…
  - **High-level percepts**: *cat* means this type of pattern




9

Grounding

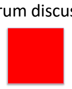
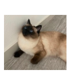
- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…
  - **High-level percepts**: *cat* means this type of pattern
  - **Embodiment (effects on the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


10

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

11

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]

12

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

13

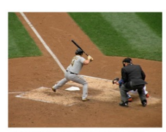
CS294-43: VISION AND LANGUAGE AI SEMINAR

14

A Gallery of Tasks

15

Image Captioning



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

16

Visual Question Answering

VQA 2.0: Goyal et al. 2017

17

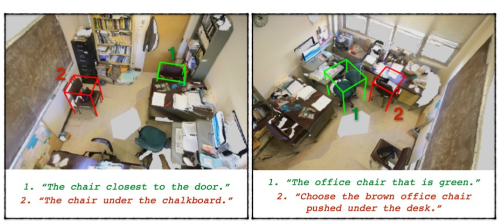
Object Detection (2D)

(a) Query: "street lamp" **(b)** Query: "major league logo" **(c)** Query: "zebras on savanna"

MDETR: Kamath et al. 2021

18

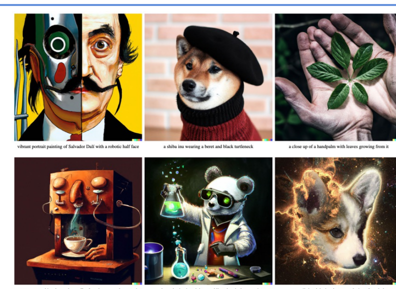
Object Detection (3D)



ReferIt3D: Achlioptas et al. 2020

19

Conditional Generation (2D)



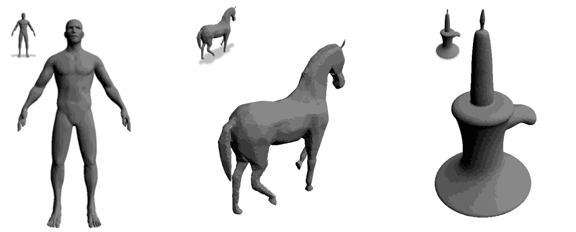
vibrant portrait painting of Salvador Dalí with a robotic half face | a shiba inu wearing a beret and black turtleneck | a close up of a handpalm with leaves growing from it

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

20

Conditional Generation (3D)



"Iron Man" "Astronaut Horse" "Colorful Crochet Candle"

Text2Mesh: Michel et al. 2021

21

Vision and Language Navigation

"Place a clean ladle on a counter"

ALFRED: Shridhar et al. 2020

22

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.

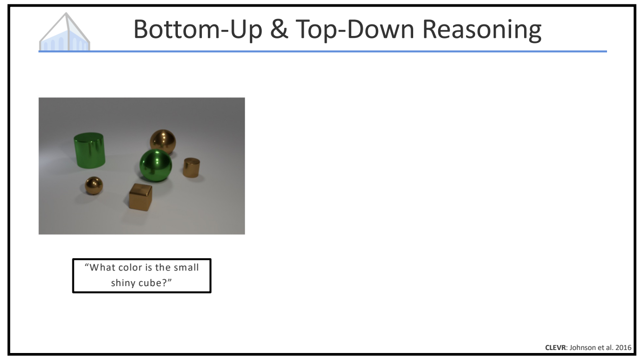
23

Source: Robohub.org

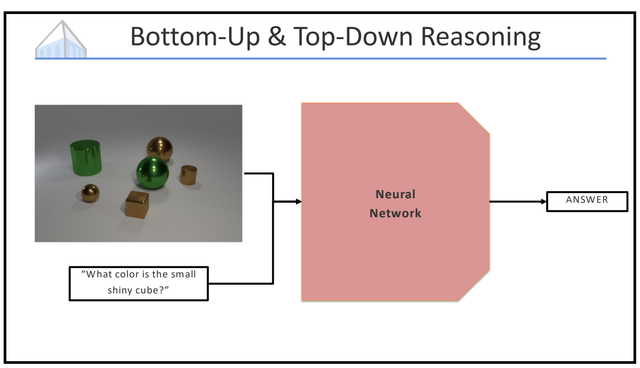
24


iRobot
Source: cnn.com

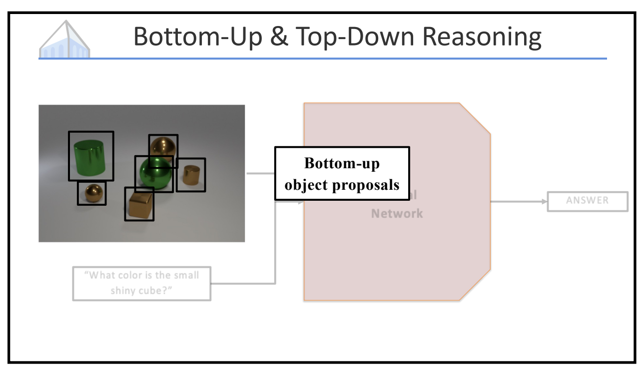
25


Bottom-Up & Top-Down Reasoning
“What color is the small shiny cube?”
CLEVR: Johnson et al. 2016

26

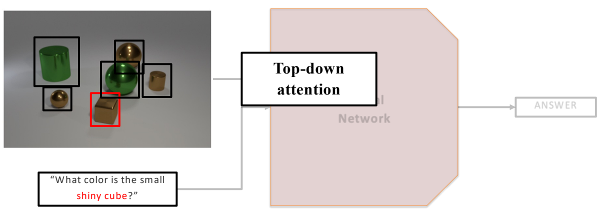

Bottom-Up & Top-Down Reasoning
“What color is the small shiny cube?”
Neural Network
ANSWER

27


Bottom-Up & Top-Down Reasoning
Bottom-up object proposals
Network
“What color is the small shiny cube?”
ANSWER

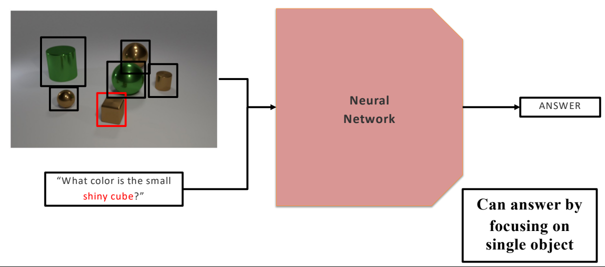
28

Bottom-Up & Top-Down Reasoning



29

Bottom-Up & Top-Down Reasoning



30

Bottom-Up & Top-Down Reasoning

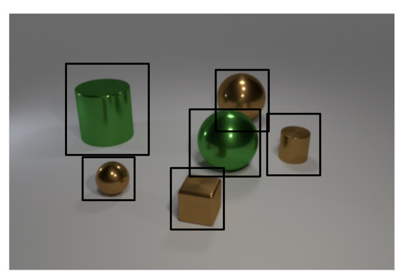
	Yes/No	Number	Other	Overall
Ours: ResNet (1×1)	76.0	36.5	46.8	56.3
Ours: ResNet (14×14)	76.6	36.2	49.5	57.9
Ours: ResNet (7×7)	77.6	37.7	51.5	59.4
Ours: Up-Down	**80.3**	**42.8**	**55.8**	**63.2**
Relative Improvement	3%	14%	8%	6%

Provides inductive bias in both directions!

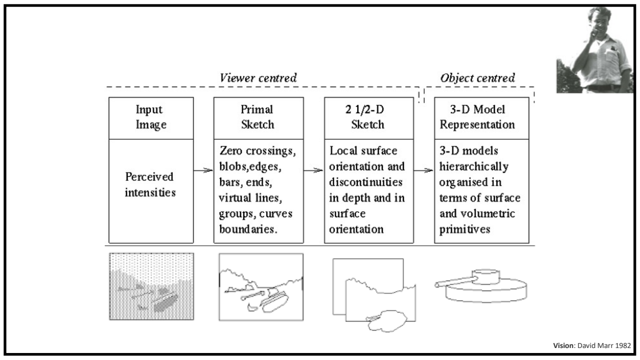
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering: Anderson et al. 2018

31

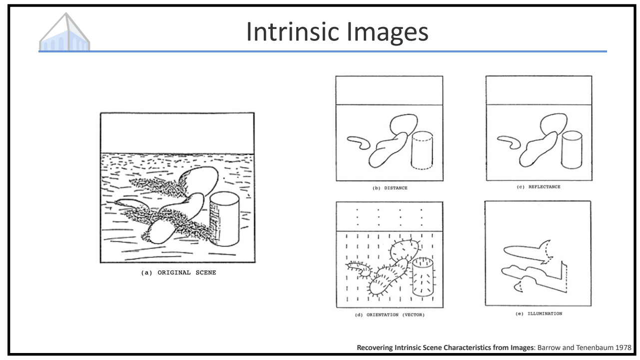
Bottom-Up



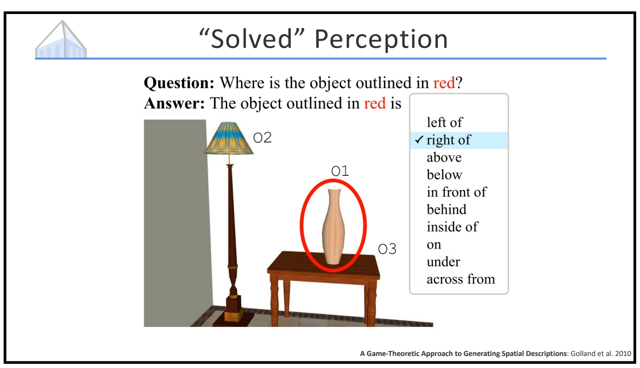
32


Viewer centred
Object centred
Input Image
Primal Sketch
2 1/2-D Sketch
3-D Model Representation
Perceived intensities
Zero crossings, blobs,edges, bars, ends, virtual lines, groups, curves boundaries.
Local surface orientation and discontinuities in depth and in surface orientation
3-D models hierarchically organised in terms of surface and volumetric primitives
Vision: David Marr 1982

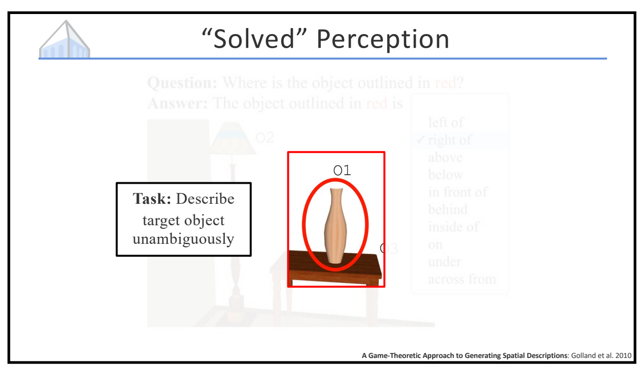
33


Intrinsic Images
(a) ORIGINAL SCENE
Recovering Intrinsic Scene Characteristics from Images: Barrow and Tenenbaum 1978

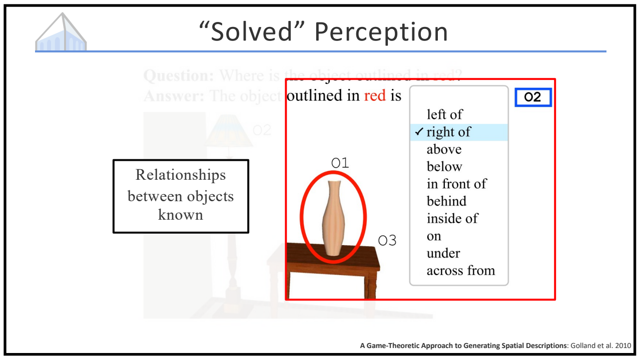
34


"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
O2
O1
O3
A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

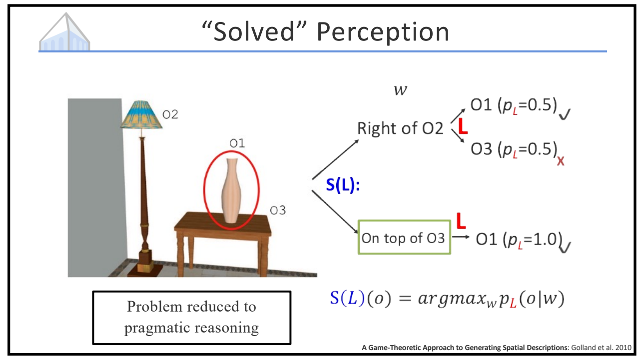
35


"Solved" Perception
Task: Describe target object unambiguously
O1
A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

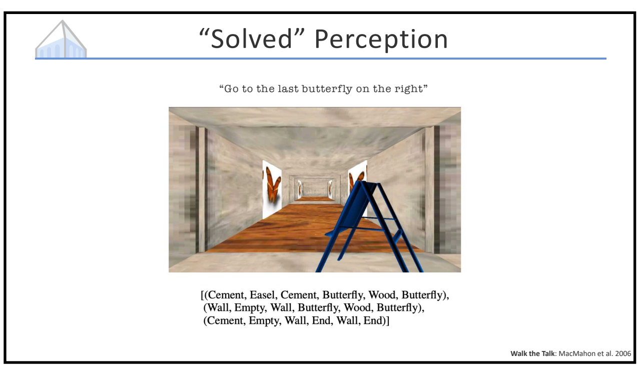
36


"Solved" Perception
outlined in red is
O2
left of
✓ right of
above
below
in front of
behind
inside of
on
under
across from
O1
O3
Relationships between objects known
A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

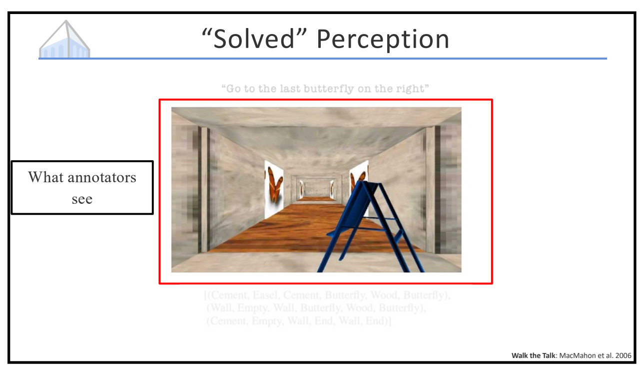
37


"Solved" Perception
O2
O1
O3
w
Right of O2
L
O1 (p_L=0.5)
O3 (p_L=0.5)
S(L):
On top of O3
L
O1 (p_L=1.0)
$S(L)(o) = argmax_w p_L(o|w)$
Problem reduced to pragmatic reasoning
A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

38


"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)]
Walk the Talk: MacMahon et al. 2006

39


"Solved" Perception
What annotators see
Walk the Talk: MacMahon et al. 2006

40

“Solved” Perception

“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)]

Walk the Talk: MacMahon et al. 2006

41

“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)]

Walk the Talk: MacMahon et al. 2006

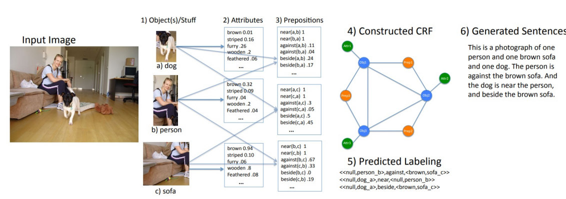
42

“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.

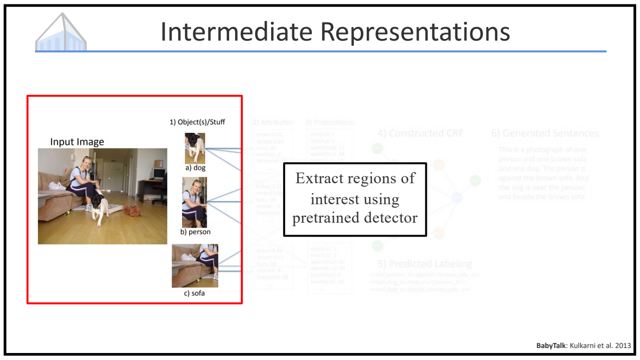
43

Intermediate Representations

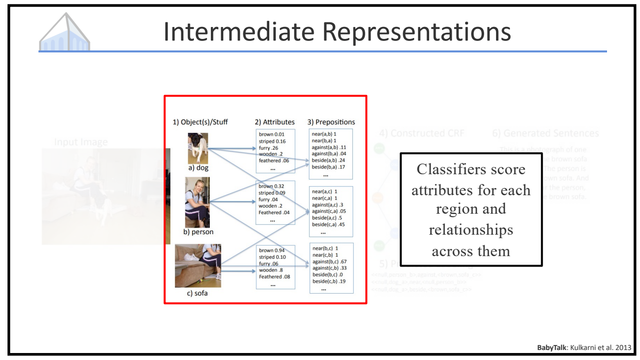


BabyTalk: Kulkarni et al. 2013

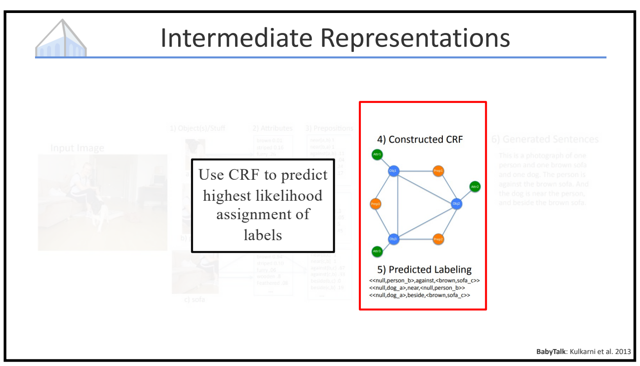
44


Intermediate Representations
Input Image
1) Object(s)/Stuff
a) dog
b) person
c) sofa
Extract regions of interest using pretrained detector
BabyTalk: Kulkarni et al. 2013

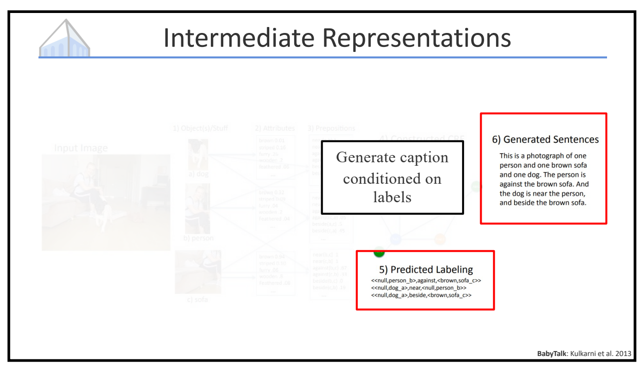
45


Intermediate Representations
1) Object(s)/Stuff
2) Attributes
3) Prepositions
a) dog
b) person
c) sofa
Classifiers score attributes for each region and relationships across them
BabyTalk: Kulkarni et al. 2013

46


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

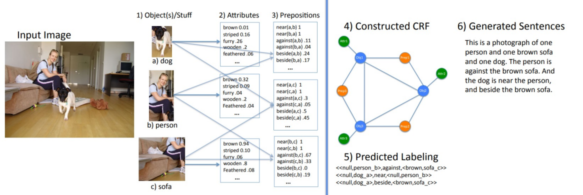
47


Intermediate Representations
Generate caption conditioned on labels
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>>
BabyTalk: Kulkarni et al. 2013

48

Intermediate Representations

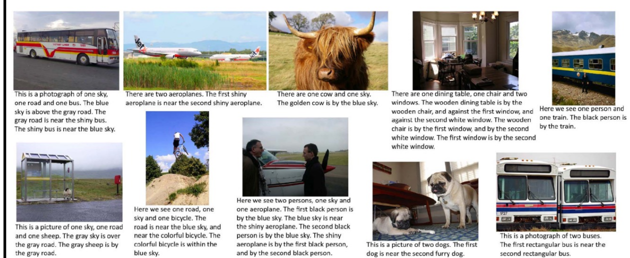
Language model never sees pixels!



BabyTalk: Kulkarni et al. 2013

49

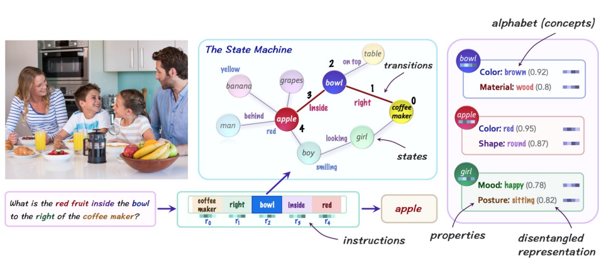
Intermediate Representations



BabyTalk: Kulkarni et al. 2013

50

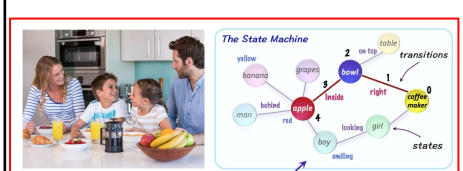
Intermediate Representations



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

51

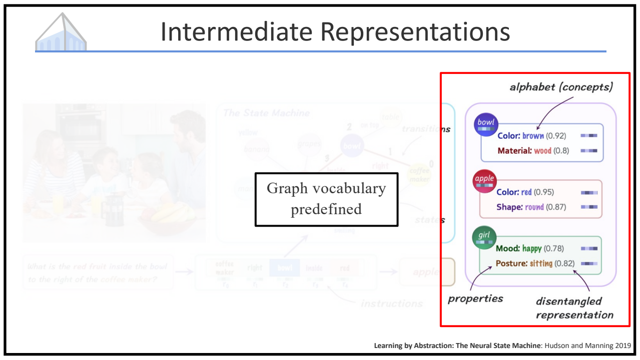
Intermediate Representations



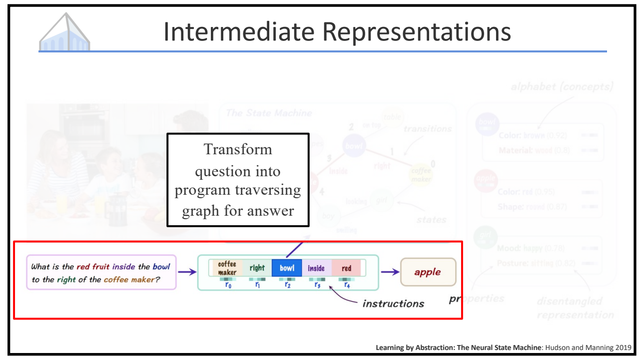
Generate scene graph from image

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

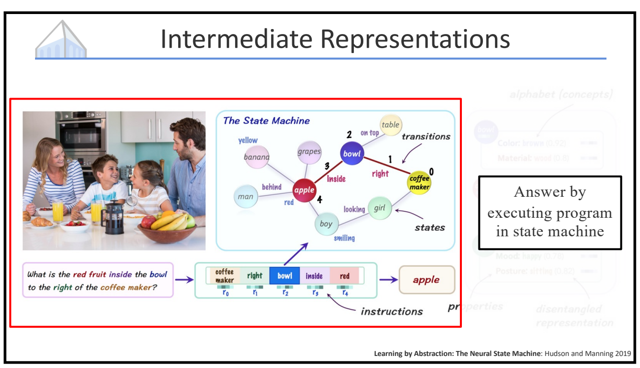
52


Intermediate Representations
Graph vocabulary predefined
alphabet (concepts)
Color: brown (0.92)
Material: wood (0.8)
Color: red (0.95)
Shape: round (0.87)
Mood: happy (0.78)
Posture: sitting (0.82)
properties
disentangled representation
Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

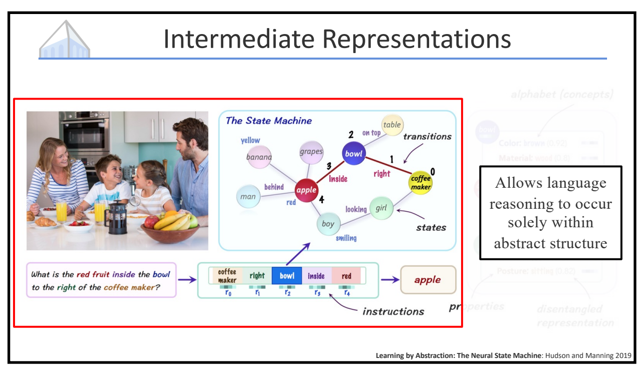
53


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?
coffee maker
right
bowl
inside
red
apple
instructions
Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

54


Intermediate Representations
The State Machine
transitions
states
Answer by executing program in state machine
What is the red fruit inside the bowl to the right of the coffee maker?
apple
instructions
Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

55


Intermediate Representations
The State Machine
transitions
states
Allows language reasoning to occur solely within abstract structure
What is the red fruit inside the bowl to the right of the coffee maker?
apple
instructions
Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

56

Intermediate Representations

Table 4: GQA generalization

Model	Content	Structure
Global Prior	8.51	14.64
Local Prior	12.14	18.21
Vision	17.51	18.68
Language	21.14	32.88
Lang+Vis	24.95	36.51
BottomUp [5]	29.72	41.83
MAC [40]	31.12	47.27
NSM	**40.24**	**55.72**

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

57

TEXT PROMPT

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.

AI-GENERATED IMAGES

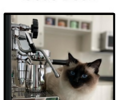
DALL-E 1: Ramesh et al. 2021

58

Intermediate Representations

Step 1

Learn Proto-linguistic Code Book



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

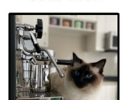
DALL-E 1: Ramesh et al. 2021

59

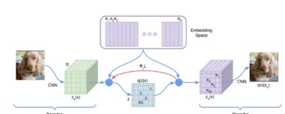
Intermediate Representations

Step 1

Learn Proto-linguistic Code Book



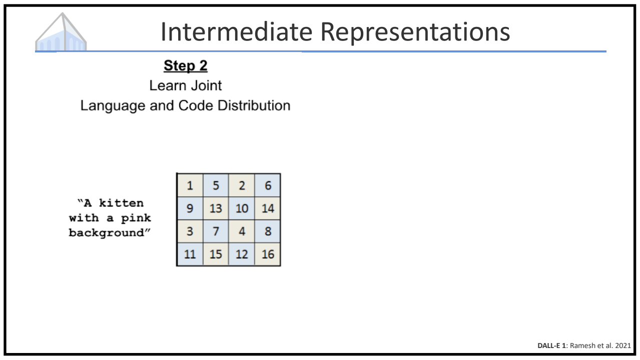
1	5	2	6
9	13	10	14
3	7	4	8
11	15	12	16



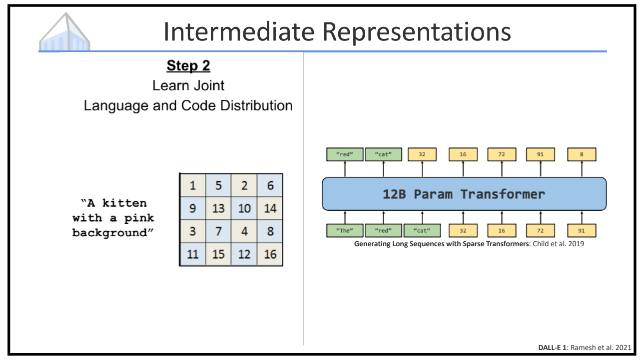
Neural Discrete Representation Learning: van Oord et al. 2017

DALL-E 1: Ramesh et al. 2021

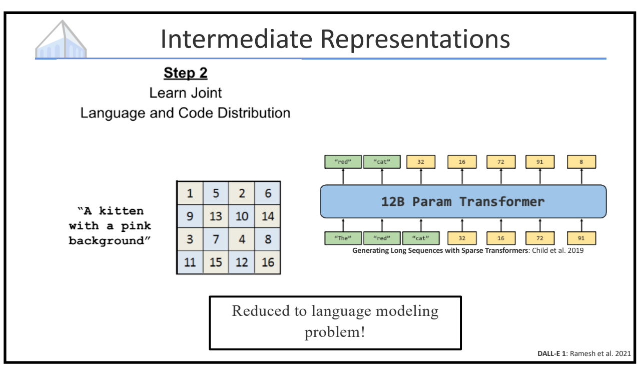
60


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
DALL-E 1: Ramesh et al. 2021

61


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
12B Param Transformer
Generating Long Sequences with Sparse Transformers: Child et al. 2019
DALL-E 1: Ramesh et al. 2021

62


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
12B Param Transformer
Generating Long Sequences with Sparse Transformers: Child et al. 2019
Reduced to language modeling problem!
DALL-E 1: Ramesh et al. 2021

63


TEXT PROMPT
an x-ray of a capybara sitting in a forest
AI-GENERATED IMAGES
DALL-E 1: Ramesh et al. 2021

64

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"

Image Understanding: John Tsotsos 1987

65

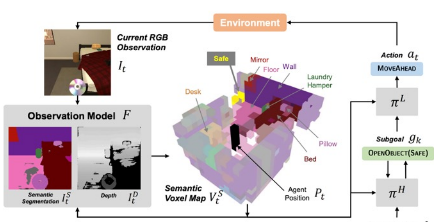
Anchoring to 3D

"Place a clean ladle on a counter"

ALFRED: Shridhar et al. 2020

66

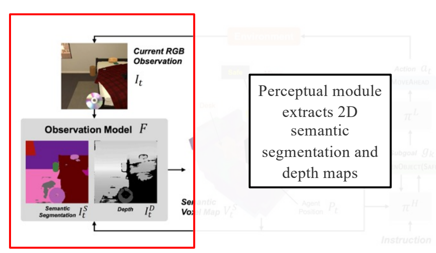
Anchoring to 3D



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

67

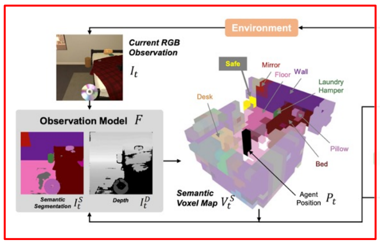
Anchoring to 3D



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

68

Anchoring to 3D



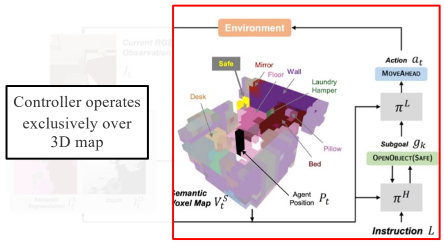
Maps integrated into persistent 3D map

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

69

Anchoring to 3D

Controller operates exclusively over 3D map



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

70

Anchoring to 3D

3D Map useful for improving performance

Method	Validation Seen SR	Validation Seen GC	Validation Unseen SR	Validation Unseen GC
HLSM	29.6	38.8	18.3	**31.2**
+ gt depth	29.6	40.5	20.1	33.7
+ gt depth, gt seg.	40.7	50.4	40.2	52.2
+ gt seg.	36.2	47.0	34.7	47.8
w/o language enc.	0.9	8.6	0.2	7.5
w/o subg. hist. enc.	29.4	38.5	16.6	29.2
w/o state repr enc.	30.0	40.6	**18.9**	30.8

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

71

Anchoring to 3D

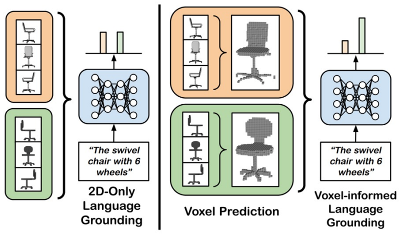
However, benefits held back by cascading errors

Method	Validation Seen SR	Validation Seen GC	Validation Unseen SR	Validation Unseen GC
HLSM	29.6	38.8	18.3	**31.2**
+ gt depth	29.6	40.5	20.1	33.7
+ gt depth, gt seg.	40.7	50.4	40.2	52.2
+ gt seg.	36.2	47.0	34.7	47.8
w/o language enc.	0.9	8.6	0.2	7.5
w/o subg. hist. enc.	29.4	38.5	16.6	29.2
w/o state repr enc.	30.0	40.6	**18.9**	30.8

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

72

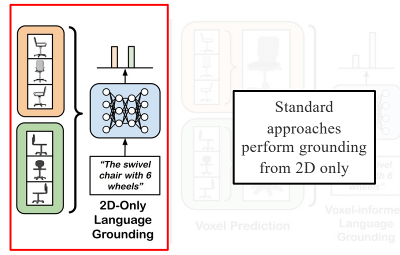
Anchoring to 3D



Voxel-informed Language Grounding: Corona et al. 2022

73

Anchoring to 3D



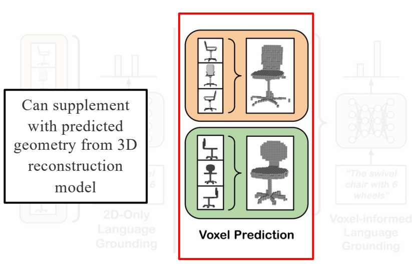
Standard approaches perform grounding from 2D only

Voxel-informed Language Grounding: Corona et al. 2022

74

Anchoring to 3D

Can supplement with predicted geometry from 3D reconstruction model



Voxel-informed Language Grounding: Corona et al. 2022

75

Anchoring to 3D

	VALIDATION			TEST		
Model	Visual	Blind	All	Visual	Blind	All
ViLBERT	89.5	76.6	83.1	80.2	**73.0**	76.6
MATCH	89.2 (0.9)	75.2 (0.7)	82.2 (0.4)	83.9 (0.5)	68.7 (0.9)	76.5 (0.5)
MATCH*	90.6 (0.4)	75.7 (1.2)	83.2 (0.8)	-	-	-
LAGOR	89.8 (0.4)	75.3 (0.7)	82.6 (0.4)	84.3 (0.4)	69.4 (0.5)	77.0 (0.5)
LAGOR*	89.8 (0.5)	75.0 (0.4)	82.5 (0.1)	-	-	-
VLG (Ours)	**91.2** (0.4)	**78.4**†(0.7)	**84.9**†(0.3)	**86.0**	71.7	**79.0**

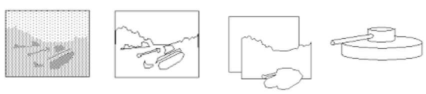
Improves performance over 2D-only approaches

Voxel-informed Language Grounding: Corona et al. 2022

76

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.

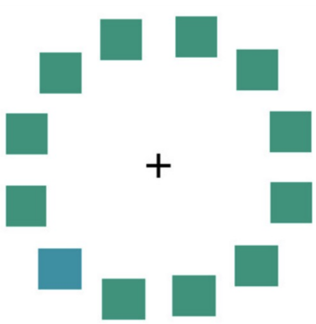


77

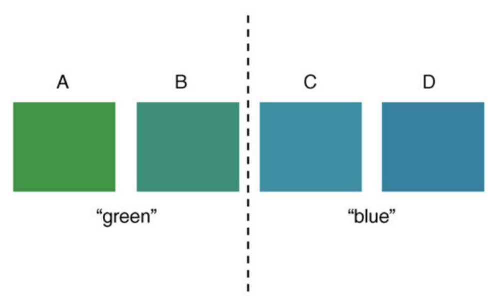
Top-Down

"What color is the small shiny cube?"

78

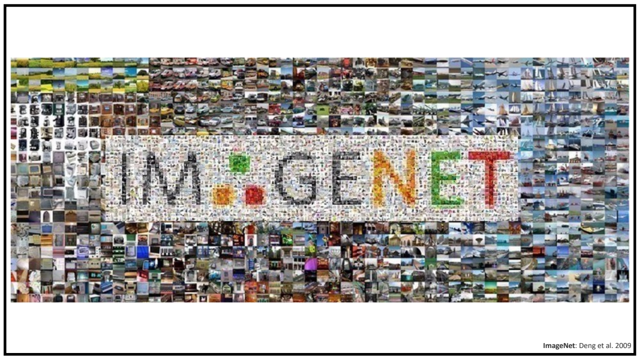


79

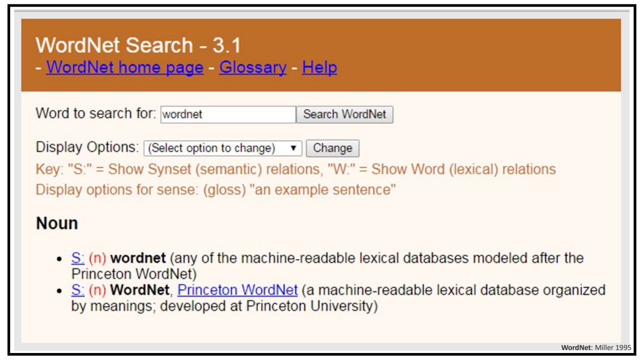


Language, thought, and color: Regier and Kay 2009

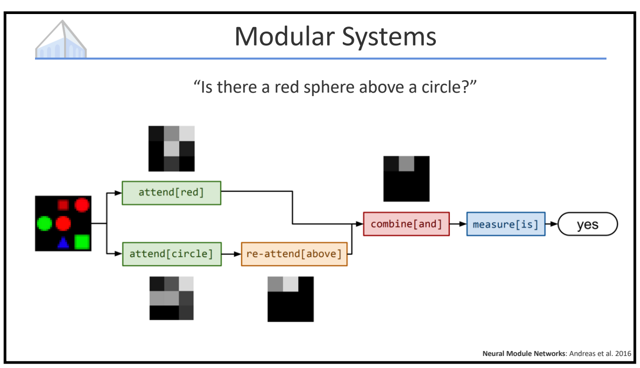
80


IM GENET
ImageNet: Deng et al. 2009

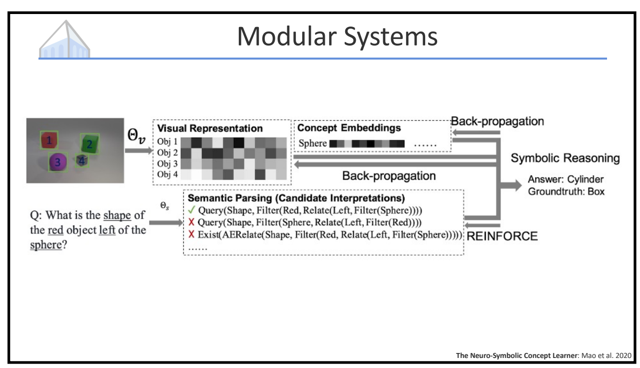
81


WordNet Search - 3.1
- WordNet home page - Glossary - Help
Word to search for: wordnet
Search WordNet
Display Options: (Select option to change)
Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"
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)
WordNet: Miller 1995

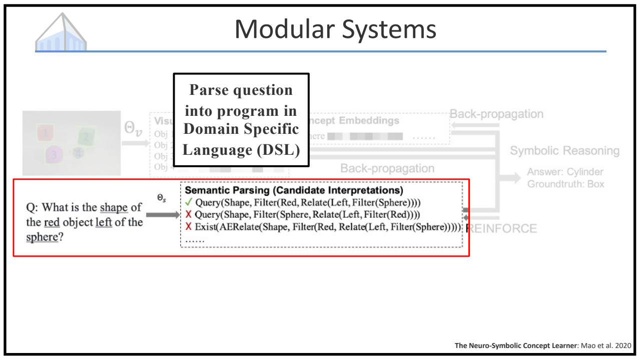
82


Modular Systems
"Is there a red sphere above a circle?"
attend[red]
attend[circle]
re-attend[above]
combine[and]
measure[is]
yes
Neural Module Networks: Andreas et al. 2016

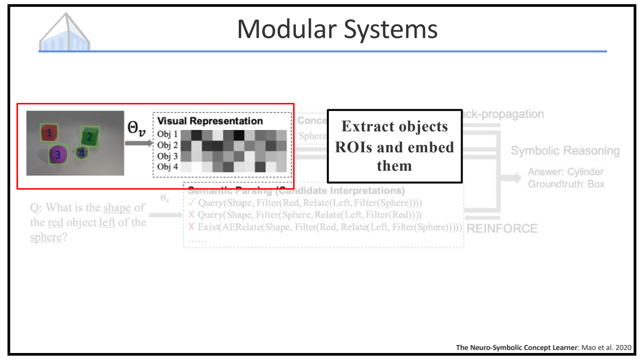
83


Modular Systems
Visual Representation
Obj 1
Obj 2
Obj 3
Obj 4
Concept Embeddings
Sphere
Back-propagation
Back-propagation
Symbolic Reasoning
Answer: Cylinder
Groundtruth: Box
Q: What is the shape of the red object left of the sphere?
Semantic Parsing (Candidate Interpretations)
✓ Query(Shape, Filter(Red, Relate(Left, Filter(Sphere))))
X Query(Shape, Filter(Sphere, Relate(Left, Filter(Red))))
X Exist(AERelate(Shape, Filter(Red, Relate(Left, Filter(Sphere)))))
REINFORCE
The Neuro-Symbolic Concept Learner: Mao et al. 2020

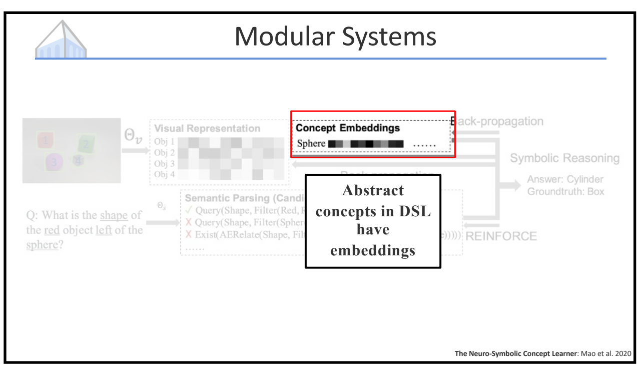
84



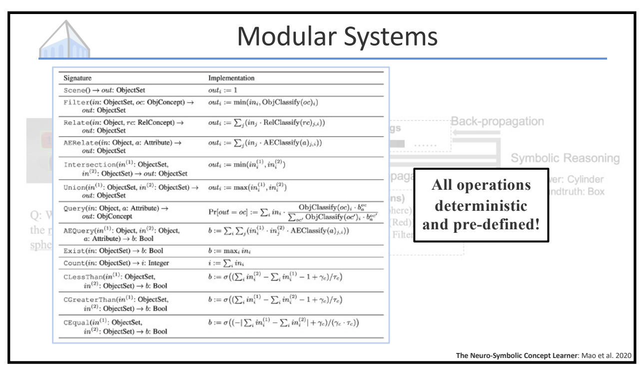
85



86



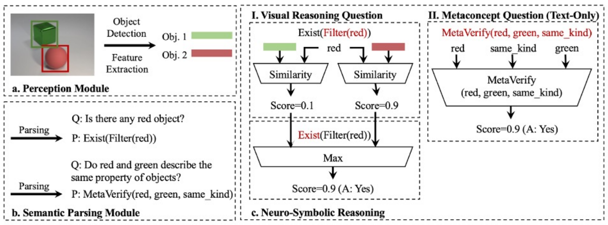
87



Signature	Implementation
Scene() → out: ObjectSet	$out_i := 1$
Filter(in: ObjectSet, oc: ObjConcept) → out: ObjectSet	$out_i := \min(in_i, \mathrm{ObjClassify}(oc)_i)$
Relate(in: Object, rc: RelConcept) → out: ObjectSet	$out_i := \sum_j (in_j \cdot \mathrm{RelClassify}(rc)_{j,i}))$
AERelate(in: Object, a: Attribute) → out: ObjectSet	$out_i := \sum_j (in_j \cdot \mathrm{AEClassify}(a)_{j,i}))$
Intersection($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) → out: ObjectSet	$out_i := \min(in_i^{(1)}, in_i^{(2)})$
Union($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) → out: ObjectSet	$out_i := \max(in_i^{(1)}, in_i^{(2)})$
Query(in: Object, a: Attribute) → out: ObjConcept	$\Pr[out = oc] := \sum_i in_i \cdot \frac{\mathrm{ObjClassify}(oc)_i \cdot b_a^{oc}}{\sum_{oc'} \mathrm{ObjClassify}(oc')_i \cdot b_a^{oc'}}$
AEQuery($in^{(1)}$: Object, $in^{(2)}$: Object, a: Attribute) → b: Bool	$b := \sum_i \sum_j (in_i^{(1)} \cdot in_j^{(2)} \cdot \mathrm{AEClassify}(a)_{j,i}))$
Exist(in: ObjectSet) → b: Bool	$b := \max_i in_i$
Count(in: ObjectSet) → i: Integer	$i := \sum_i in_i$
CLessThan($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) → b: Bool	$b := \sigma\big((\sum_i in_i^{(2)} - \sum_i in_i^{(1)} - 1 + \gamma_c)/\tau_c\big)$
CGreaterThan($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) → b: Bool	$b := \sigma\big((\sum_i in_i^{(1)} - \sum_i in_i^{(2)} - 1 + \gamma_c)/\tau_c\big)$
CEqual($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) → b: Bool	$b := \sigma\big((-\lvert\sum_i in_i^{(1)} - \sum_i in_i^{(2)}\rvert + \gamma_c)/(\gamma_c \cdot \tau_c)\big)$

88

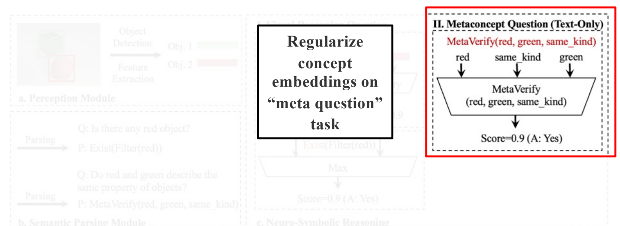
Modular Systems



Visual Concept-Metaconcept Learning: Han et al. 2019

89

Modular Systems



Visual Concept-Metaconcept Learning: Han et al. 2019

90

Modular Systems

"block" == "square"

	GRU-CNN	MAC	NS-CL	VCML
CLEVR	$50.0_{\pm 0.0}$	$68.7_{\pm 3.8}$	$80.2_{\pm 3.1}$	$\mathbf{94.1}_{\pm \mathbf{4.6}}$
GQA	$50.0_{\pm 0.0}$	$49.5_{\pm 0.2}$	$49.3_{\pm 0.6}$	$\mathbf{50.5}_{\pm \mathbf{0.1}}$

Learning *synonyms* helps zero-shot generalization

Visual Concept-Metaconcept Learning: Han et al. 2019

91

Modular Systems

 == "purple" + "square"

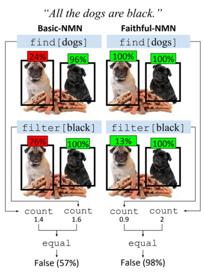
	GRU-CNN	MAC	NS-CL	VCML
CLEVR-200	$50.0_{\pm 0.0}$	$94.2_{\pm 3.3}$	$98.5_{\pm 0.3}$	$\mathbf{98.9}_{\pm \mathbf{0.2}}$
CLEVR-20	$50.0_{\pm 0.0}$	$79.7_{\pm 2.6}$	$\mathbf{95.7}_{\pm \mathbf{0.0}}$	$95.1_{\pm 1.6}$

Learning *same kind* helps compositional generalization

Visual Concept-Metaconcept Learning: Han et al. 2019

92

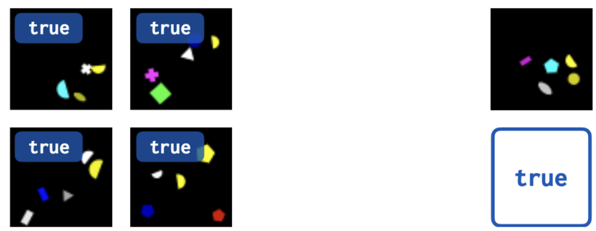
Modular Systems



Obtaining Faithful Interpretations from Compositional Neural Networks: Subramanian et al. 2020

93

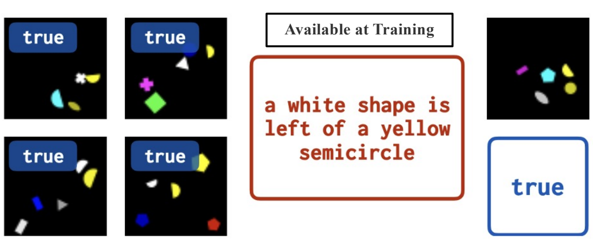
Language as Signal for Abstractions



Learning with Latent Language: Andreas et al. 2017

94

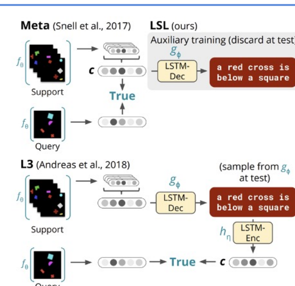
Language as Signal for Abstractions



Learning with Latent Language: Andreas et al. 2017

95

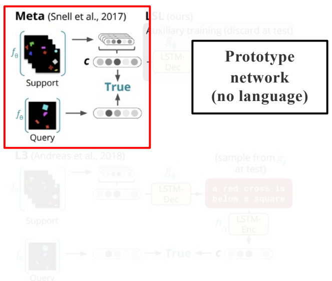
Language as Signal for Abstractions



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

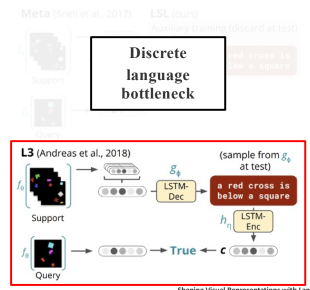
96

Language as Signal for Abstractions



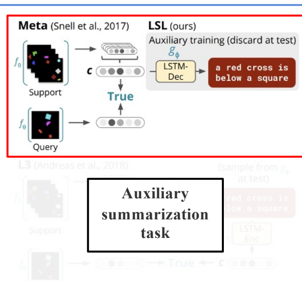
97

Language as Signal for Abstractions



98

Language as Signal for Abstractions



99

Language as Signal for Abstractions

Test Set Accuracy

	ShapeWorld	Birds ($D = 20$)
Meta	60.59 ± 1.07	57.97 ± 0.96
L3	66.60 ± 1.18	53.96 ± 1.06
LSL	$\mathbf{67.29 \pm 1.03}$	$\mathbf{61.24 \pm 0.96}$

100

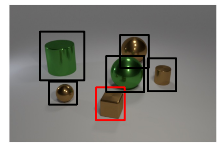
Top-Down Takeaways

- 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.

WordNet Search - 3.1
- WordNet home page - Glossary - Help

101

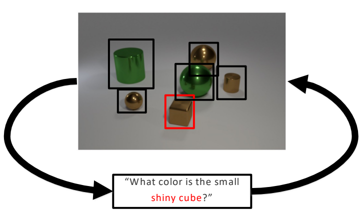
Bottom-Up & Top-Down Reasoning



"What color is the small shiny cube?"

102

Bottom-Up & Top-Down Reasoning



103

Galatea of the Spheres, Salvador Dali 1952

104

Extra Slides

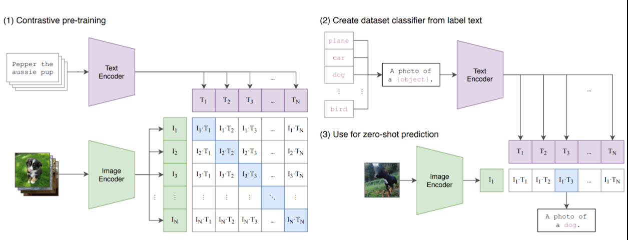
105

Open-Set Models

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

106

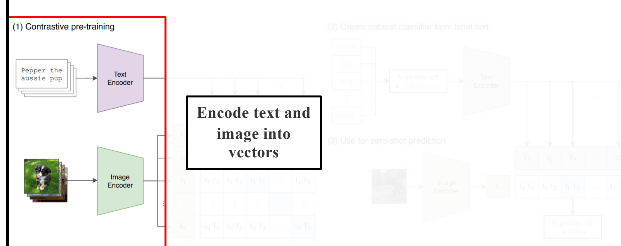
Open-Set Models



CLIP: Radford et al. 2021

107

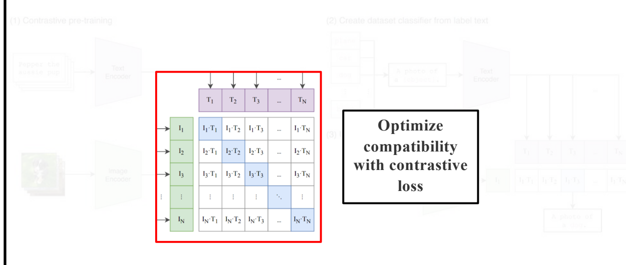
Open-Set Models



CLIP: Radford et al. 2021

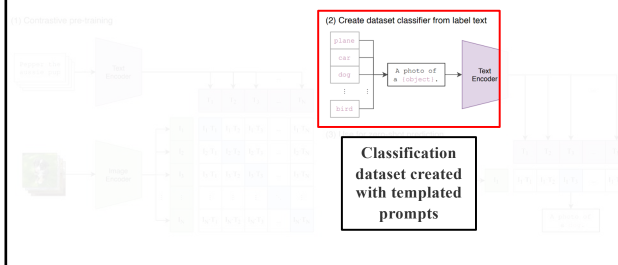
108

Open-Set Models



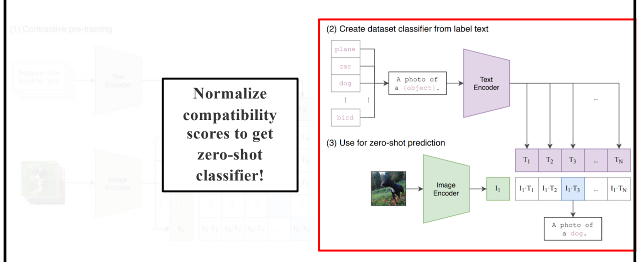
109

Open-Set Models



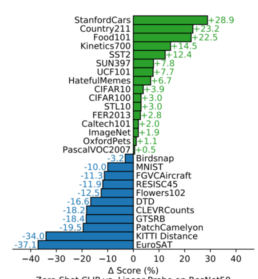
110

Open-Set Models

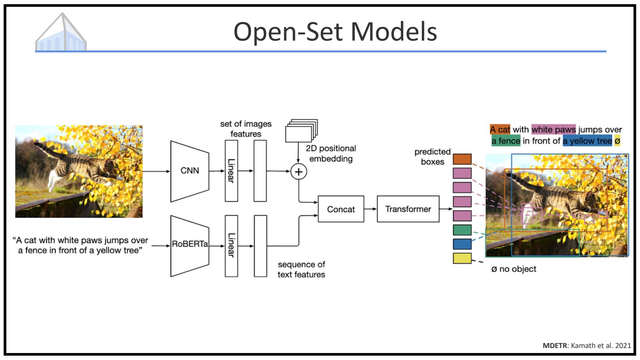


111

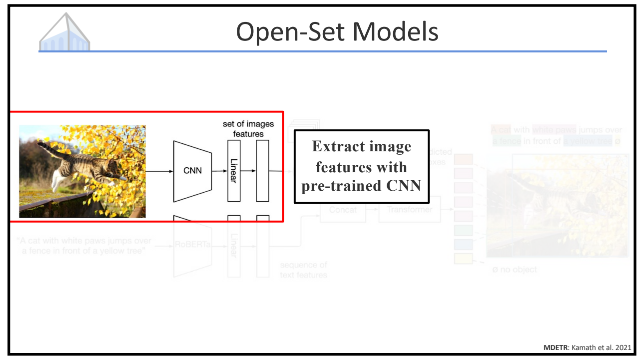
Open-Set Models



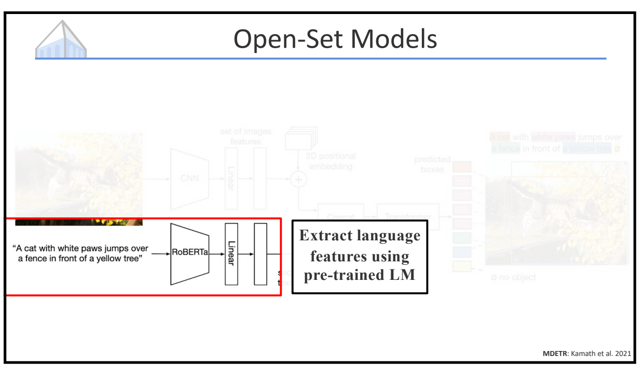
112



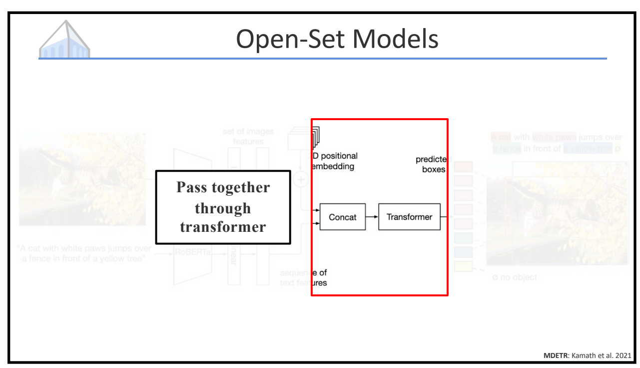
113



114

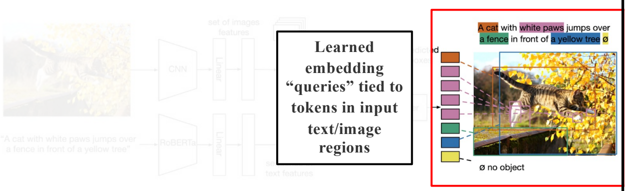


115



116

Open-Set Models



MDETR: Kamath et al. 2021

117

Open-Set Models

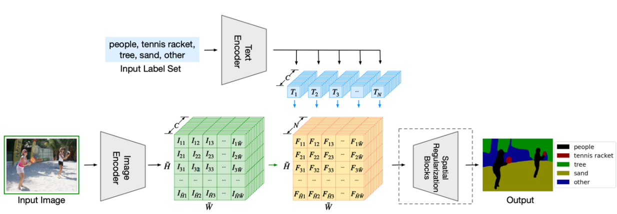
(a) "one small boy climbing a pole with the help of another boy on the ground"

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

MDETR: Kamath et al. 2021

118

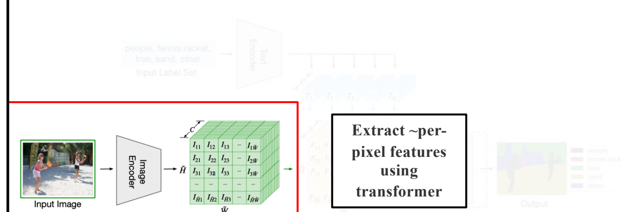
Open-Set Models



Language-Driven Semantic Segmentation: Li et al. 2022

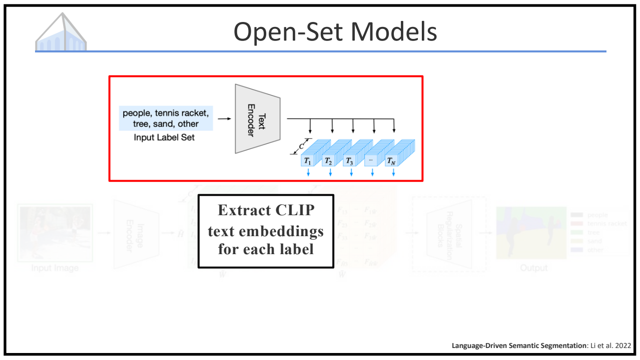
119

Open-Set Models

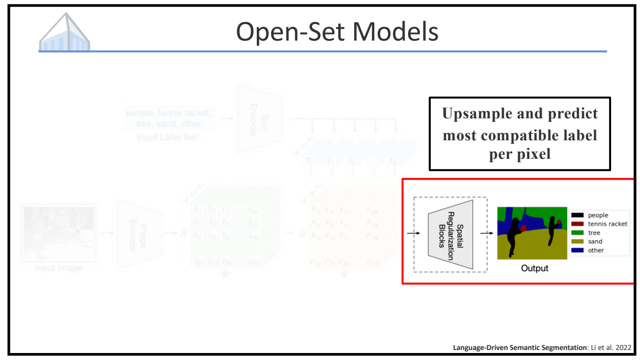


Language-Driven Semantic Segmentation: Li et al. 2022

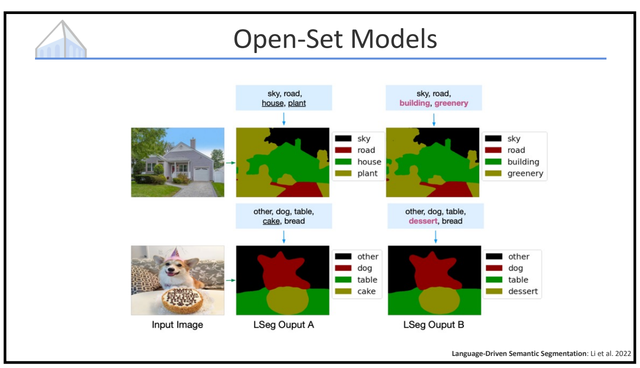
120



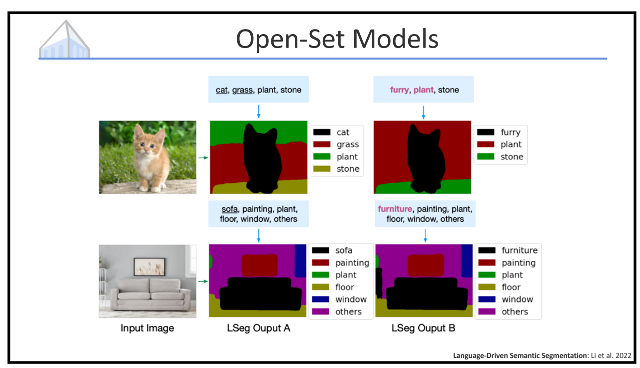
121



122

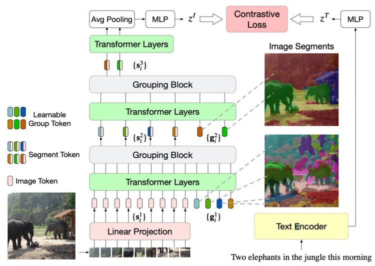


123



124

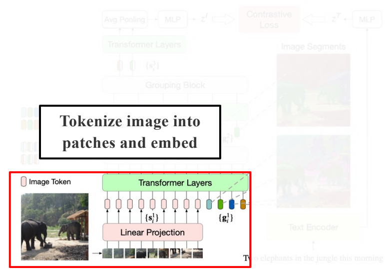
Open-Set Models



GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

125

Open-Set Models

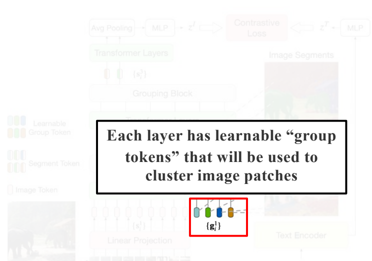


Tokenize image into patches and embed

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

126

Open-Set Models



Each layer has learnable "group tokens" that will be used to cluster image patches

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

127

Open-Set Models

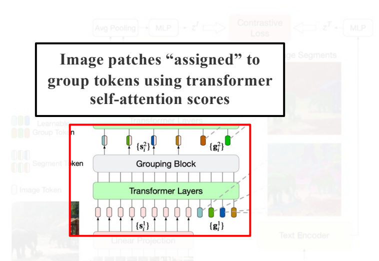
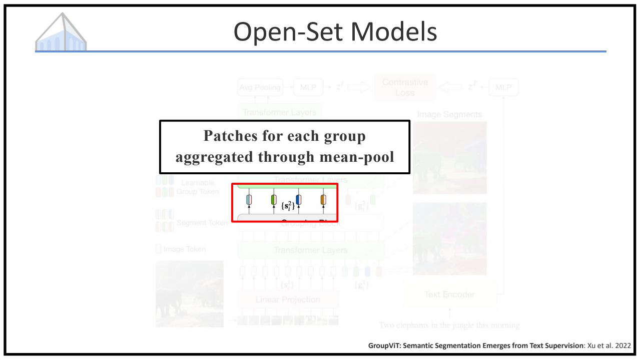


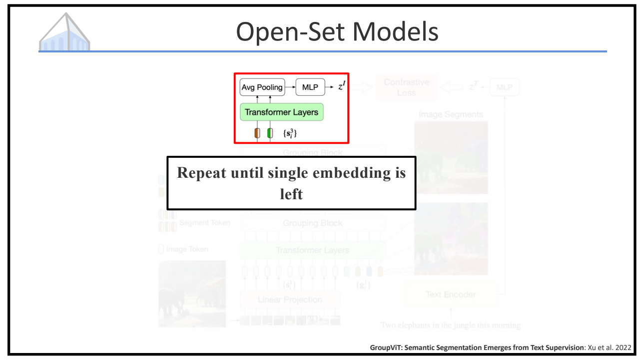
Image patches "assigned" to group tokens using transformer self-attention scores

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

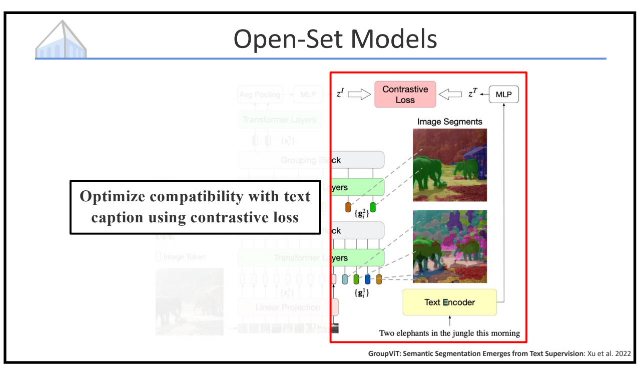
128



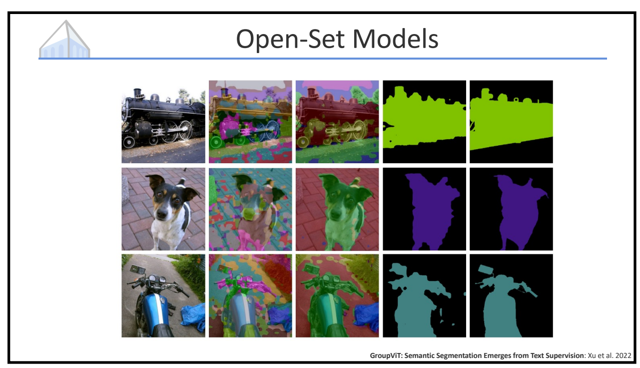
129



130



131



132

Open-Set Models

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

133

Open-Set Models

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

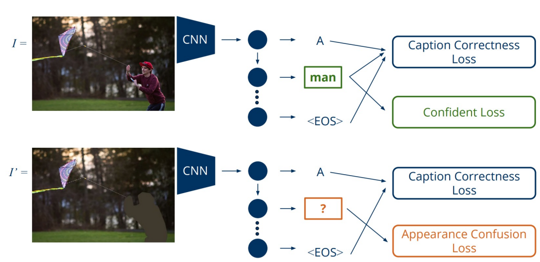
134

Bias in Vision and Language Models

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

135

Bias in Vision and Language Models



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

136

Bias in Vision and Language Models

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications: Agarwal et al. 2021

137

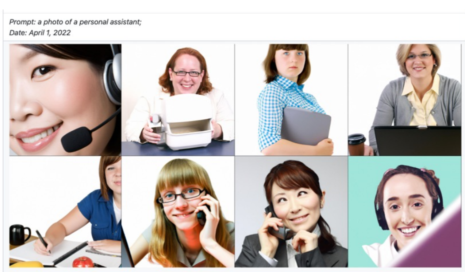
Bias in Vision and Language Models

Neurons work

Multimiodal Neurons in Artificial Neural Networks: Goh et al. 2021

138

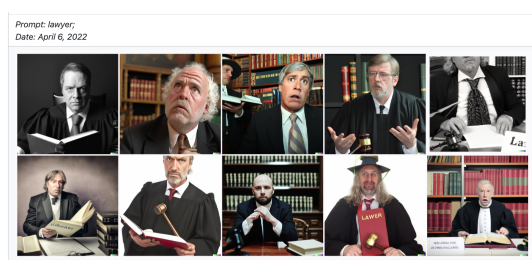
Bias in Vision and Language Models



DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022

139

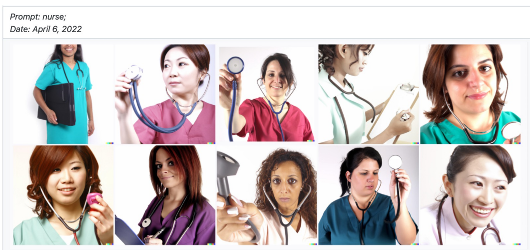
Bias in Vision and Language Models



DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022

140

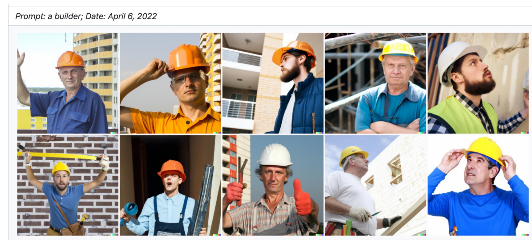
Bias in Vision and Language Models



DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022

141

Bias in Vision and Language Models



DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022

142