Embodied Language: Perception

Berkeley

N L P

EECS 183/283a: Natural Language Processing

Embodiment

- So far, we've almost exclusively covered the problem of building **language models**, which tell us:
  - How to compute the probability of a piece of text (or speech data)
  - How to compute the probability of a piece of text (or speech data), conditioned on some other text, including an instruction
- But language is produced *in context of an interaction between two or more people*
- What might that context include?

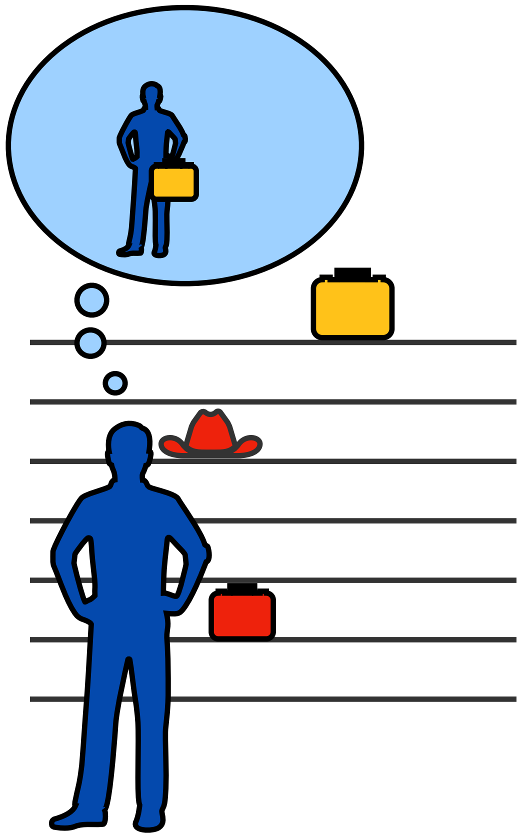
Context

Perception and Action

The “natural habitat” of language is *embodiment within* some environment comprising at least two language users, where each language user can:

- Perceive some features that the other participant(s) can also (at least partially) perceive
- Take actions that influence the state of the environment, thus influence the perception of all of the participants

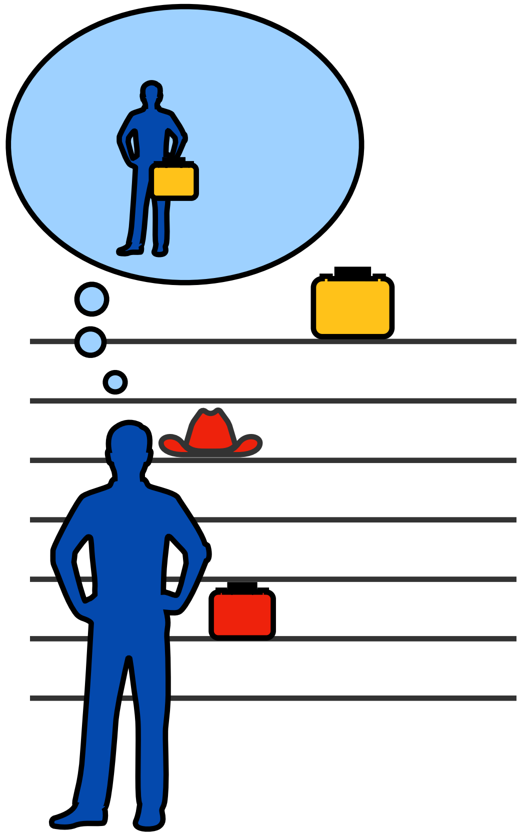
Perception and Action



<u>My goal:</u>
Get the suitcase

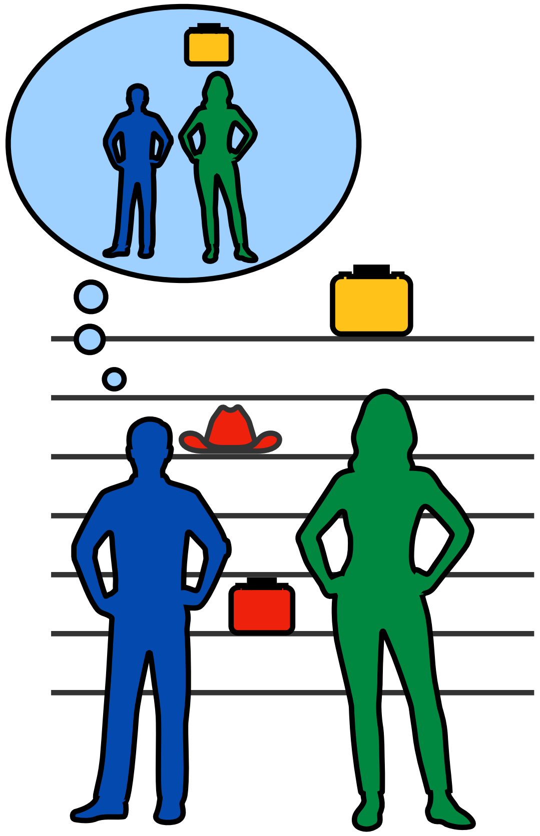
but… I'm too short

Perception and Action



<u>My goal:</u>
Get the suitcase

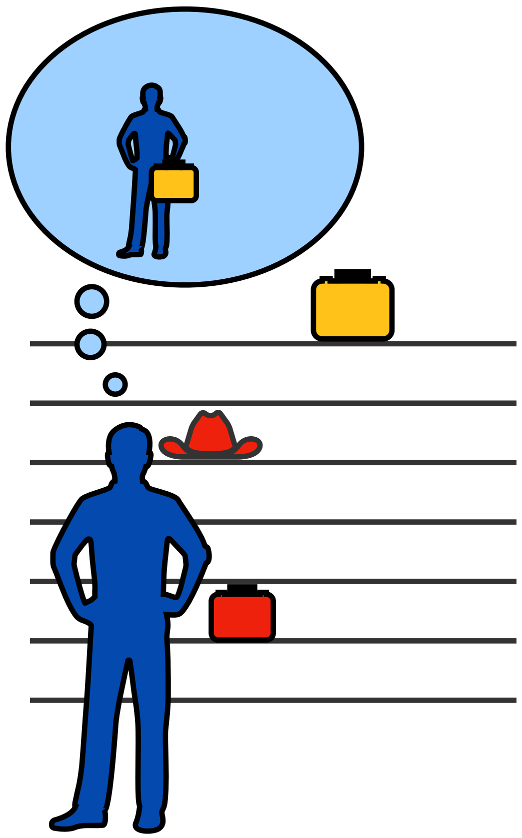
but… I'm too short



<u>Observation:</u>
Green person can reach it

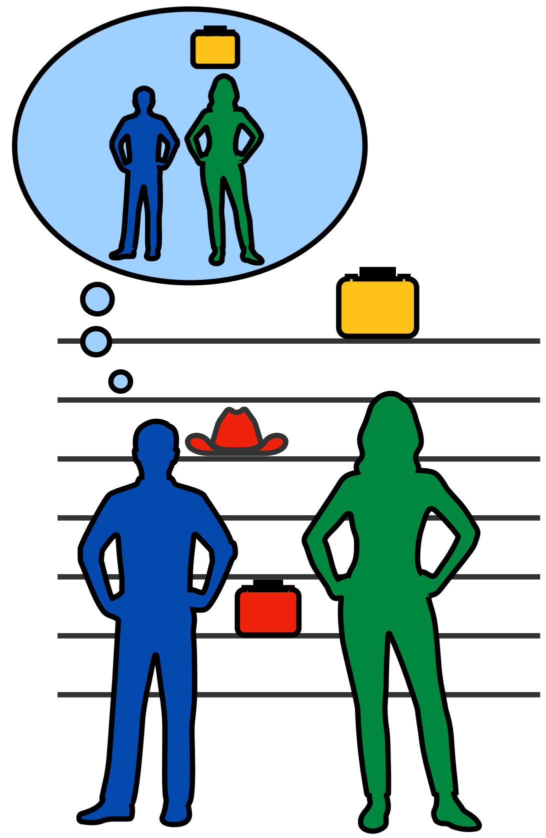
<u>Optimal action:</u>
Green person gives it to me

Perception and Action



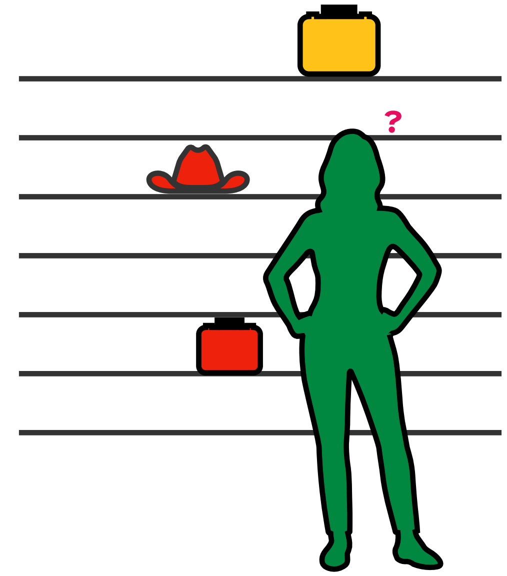
<u>My goal:</u>
Get the suitcase

but… I'm too short



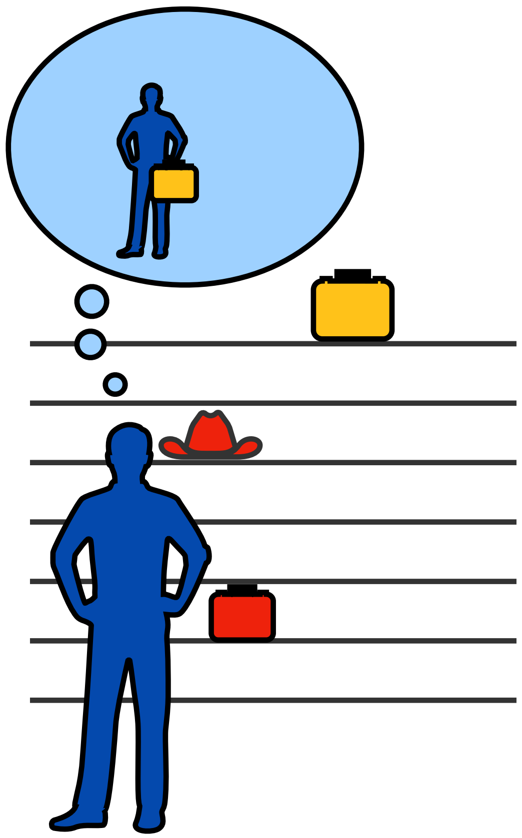
<u>Observation:</u>
Green person can reach it

<u>Optimal action:</u>
Green person gives it to me



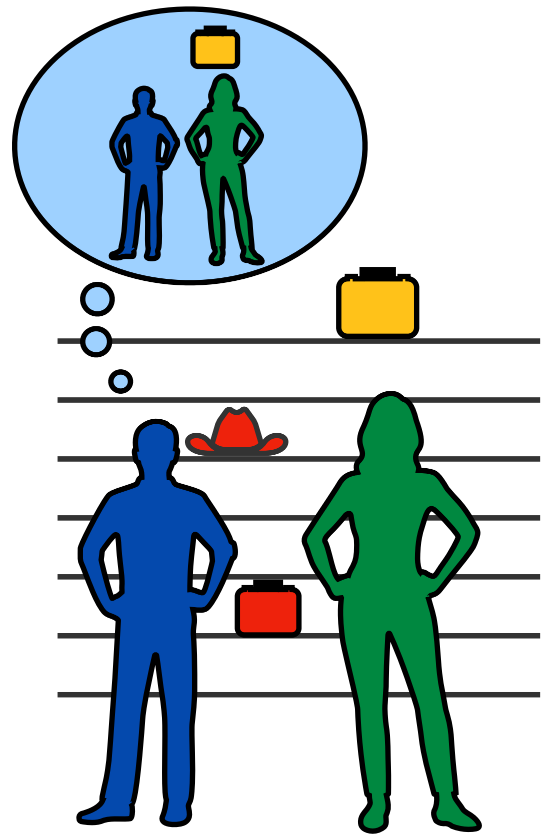
<u>Belief:</u>
Green person doesn't know my goal

Perception and Action



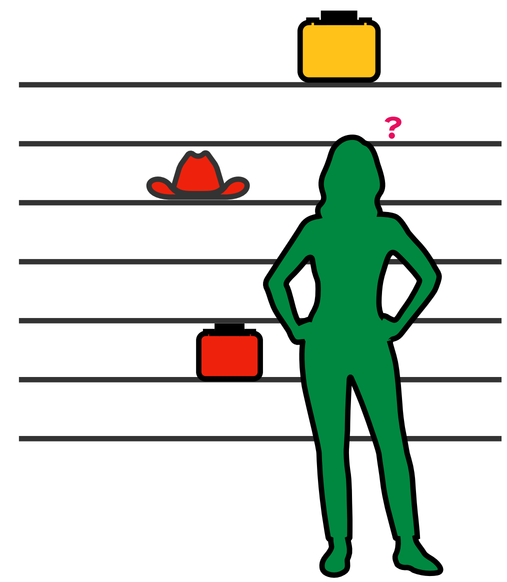
<u>My goal:</u>
Get the suitcase

but… I'm too short



<u>Observation:</u>
Green person can reach it

<u>Optimal action:</u>
Green person gives it to me

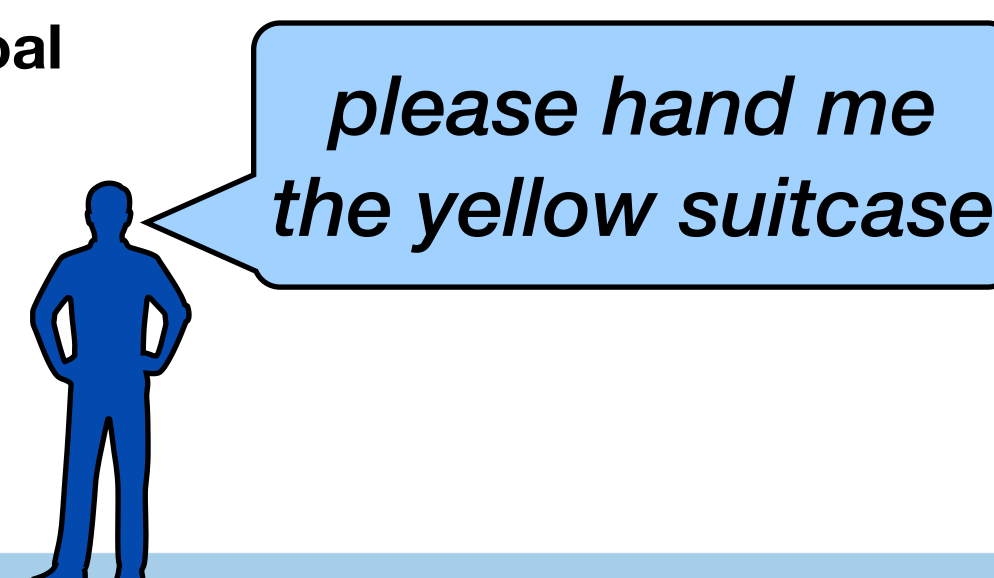


<u>Belief:</u>
Green person doesn't know my goal

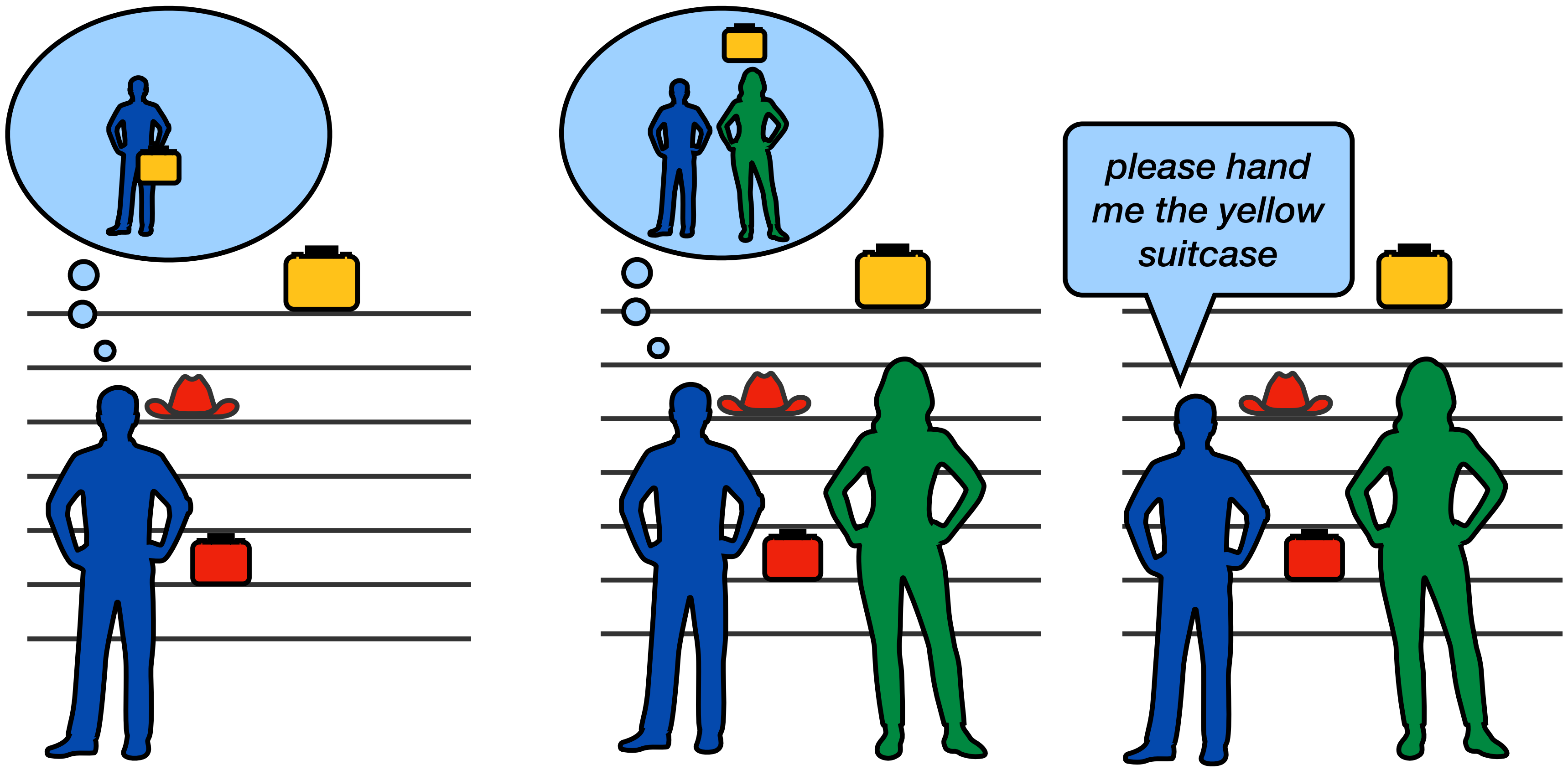
Task:
Language Generation

Input:

Output:



Perception and Action



My goal:
Get the suitcase

but… I'm too short

Observation:
Green person can reach it

Optimal action:
Green person gives it to me

Belief:
Green person doesn't know my goal

Task: Language Understanding

Input:

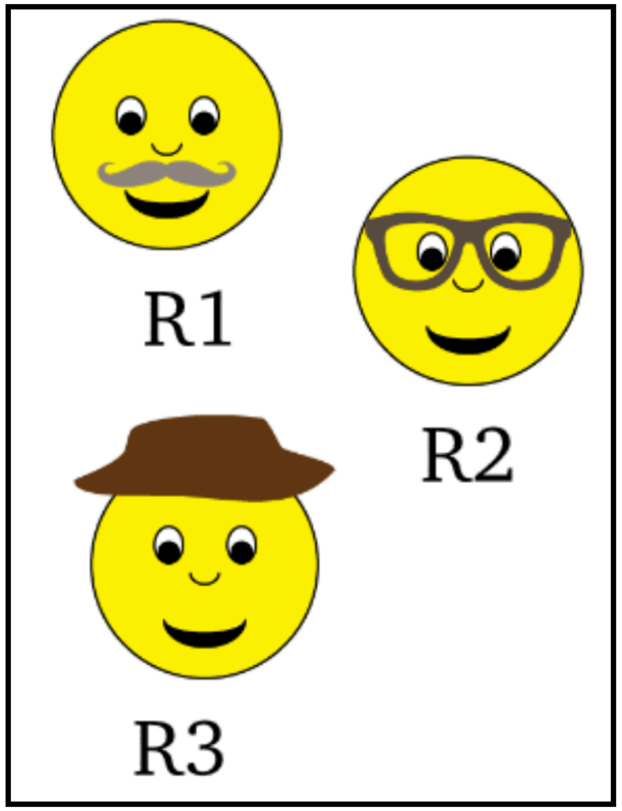
Output:

Perception

Core problems:

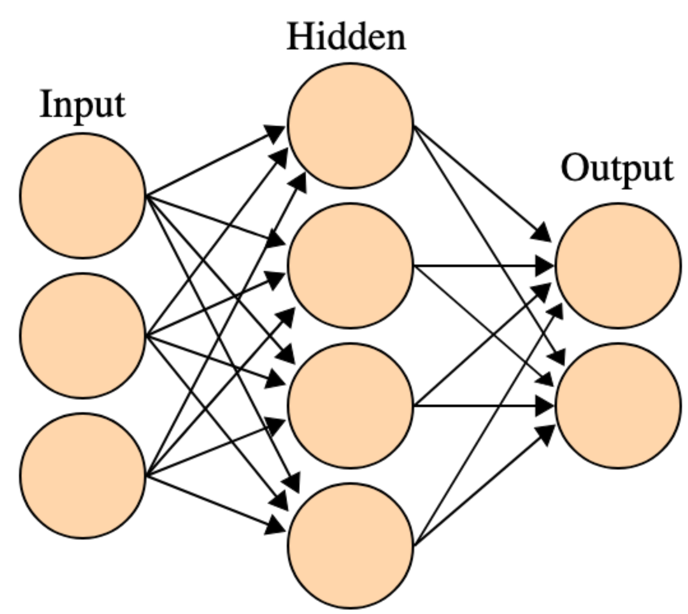
- How do we represent non-language observations of the world?
- How do we connect those representations to linguistic representations, to solve the problems of
  - language generation?
  - language understanding?

World Representations



	mustache?	glasses?	hat?
R1	yes	no	no
R2	no	yes	no
R3	no	no	yes

Structured Representation



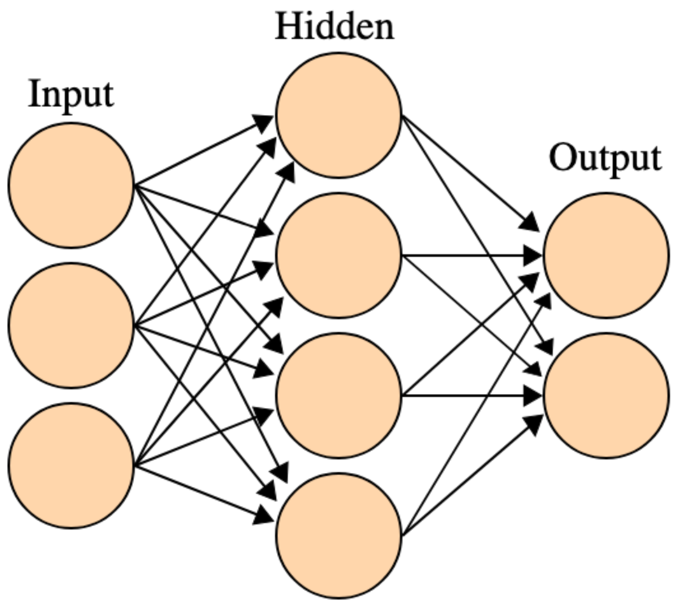
Distributed Representation

Deep Learning for Computer Vision

- **Main problem:** map from image data to some intermediate vector representation that can be used downstream to extract information from an image

- Image data: $\mathbb{R}^{3 \times h \times w}$

- What kind of information to extract?



Convolutional Neural Networks (CNN)

Convolution operation

- Takes as input some feature map $\mathbb{R}^{h \times w}$ and a filter $\mathbb{R}^{n \times n}$
- For each patch of size $n \times n$ in the input, compute dot product with filter to get a single scalar value
- For multi-channel inputs: one filter per channel, and sum over output channels

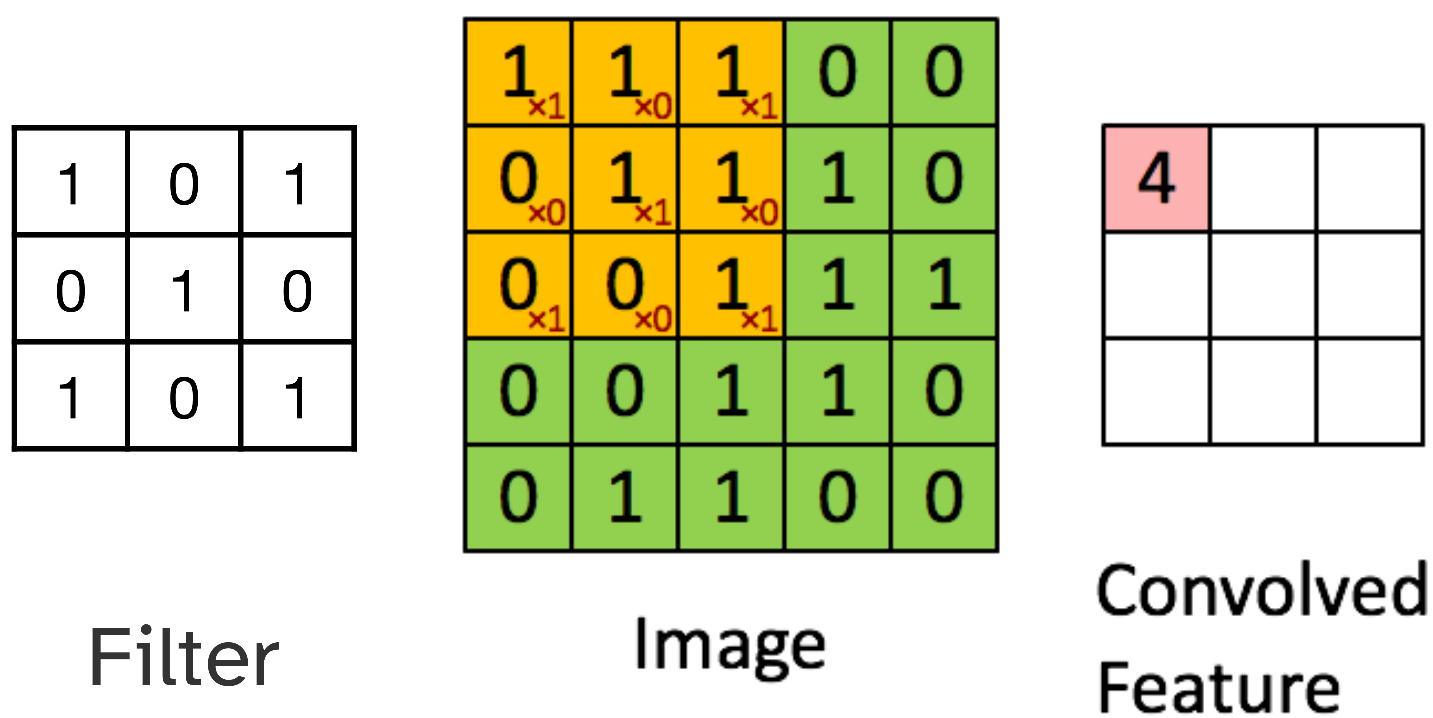
1	0	1
0	1	0
1	0	1

Filter

Convolutional Neural Networks (CNN)

Convolution operation

- Takes as input some feature map $\mathbb{R}^{h \times w}$ and a filter $\mathbb{R}^{n \times n}$
- For each patch of size $n \times n$ in the input, compute dot product with filter to get a single scalar value
- For multi-channel inputs: one filter per channel, and sum over output channels



Convolutional Neural Networks (CNN)

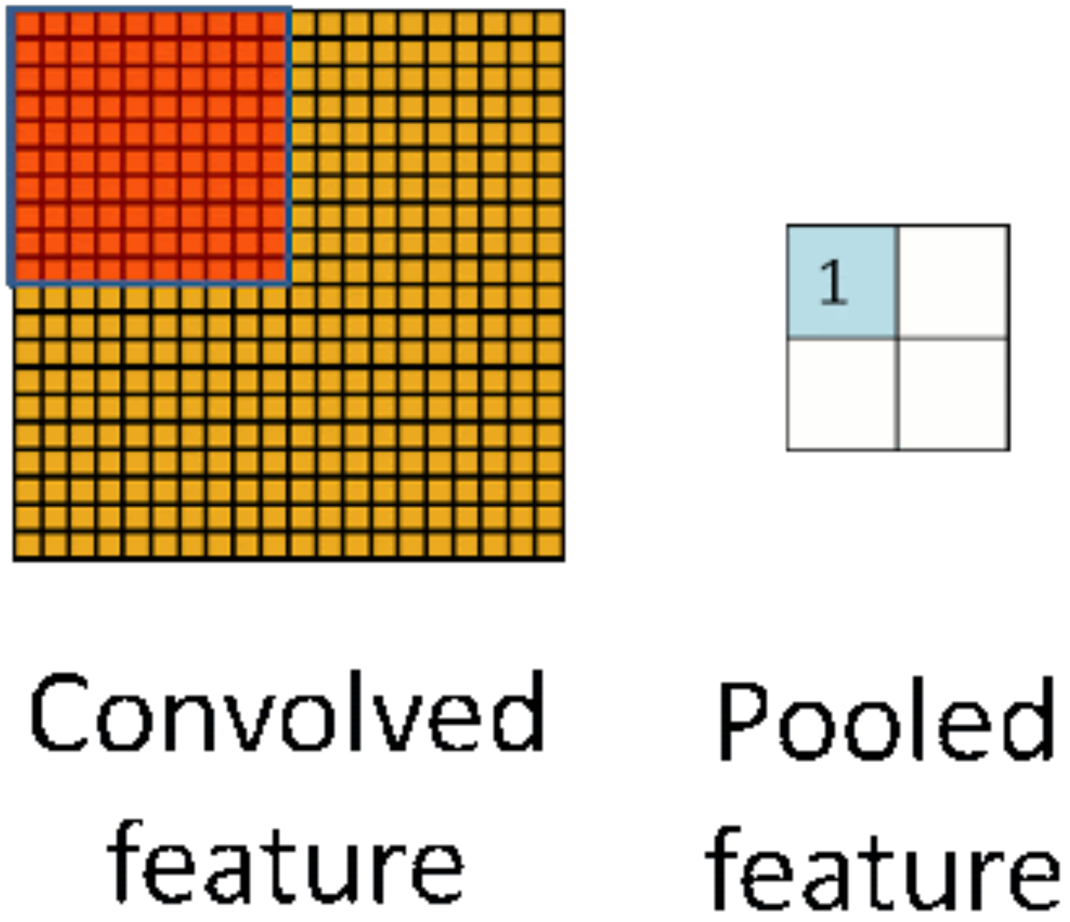
Pooling operation

- Takes as input some feature map $\mathbb{R}^{h \times w}$
- For each patch of size $n \times n$ in the input, find the maximum (or average) value in that patch

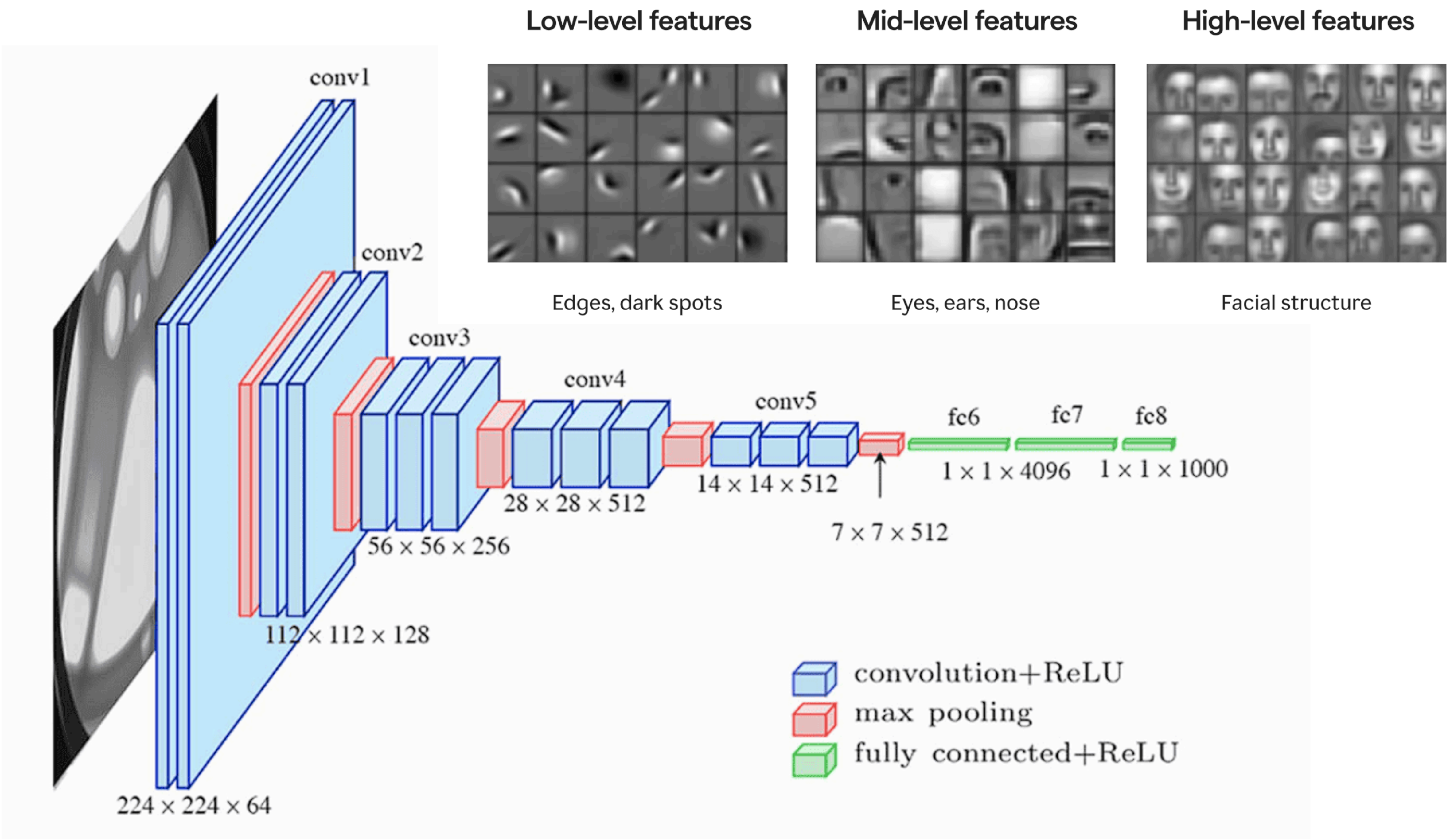
Convolutional Neural Networks (CNN)

Pooling operation

- Takes as input some feature map $\mathbb{R}^{h \times w}$
- For each patch of size $n \times n$ in the input, find the maximum (or average) value in that patch



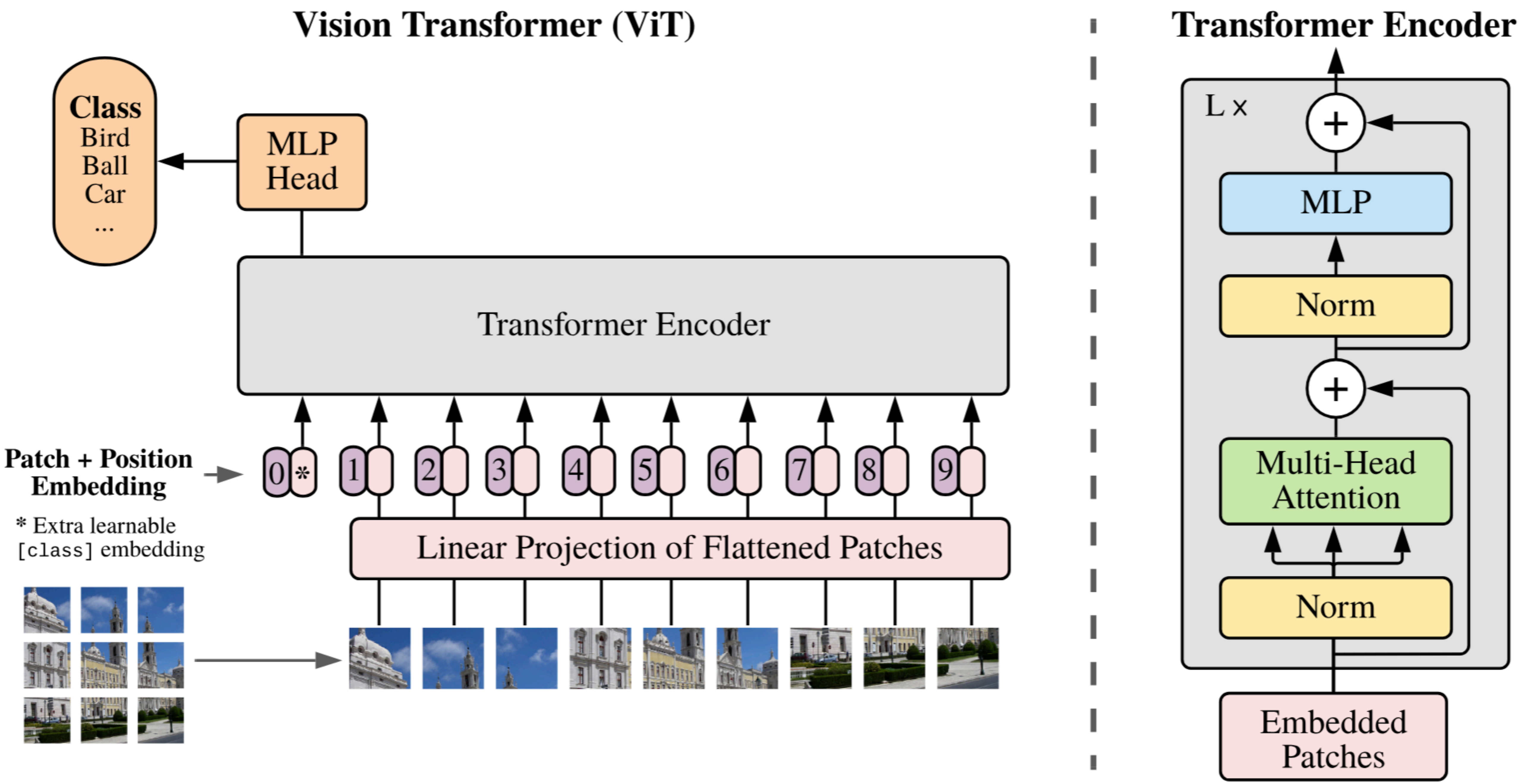
Convolutional Neural Networks (CNN)



https://poloclub.github.io/cnn-explainer/

VGGNet, Simonyan et al. 2014

Vision Transformer (ViT)



Dosovitskiy et al. 2021

From Computer Vision to Language Reasoning

- Classical computer vision tasks require “commonsense”:
  - Segmentation
  - Object recognition
  - Relative depth
  - Pose estimation

- Some of these tasks start looking a lot like language tasks…

clothing > outerwear > coat > jacket > jean jacket

Language Grounding

- How do we connect representations of non-language observations to linguistic representations, to solve the problems of
  - language generation?
  - language understanding?
- How to evaluate language grounding ability?

Image-Text Entailment

- Inputs:
  - visual observation
  - natural language statement
- Output: true or false (binary classification)

The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

[NLVR2, Suhr et al. 2019]

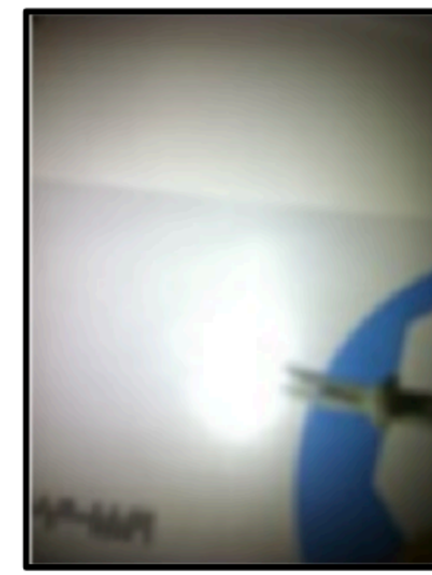
右图中的人在发球，左图中的人在接球。

[MaRVL, Liu Fangyu et al. 2021]

Visual Question Answering

- Inputs:
  - visual observation
  - natural language question
- Output: natural language response

Is this a vegetarian pizza?
[VQA, Antol et al. 2015]



Who is this mail for?
[VizWiz, Gurari et al. 2018]

Image Captioning

- Input: visual observation
- Output: natural language statement

Concadia, Kreiss et al. 2023

Image Captioning

- Input: visual observation + context of caption's purpose?
- Output: natural language statement

Concadia, Kreiss et al. 2023

Image Captioning

- Input: visual observation + context of caption's purpose?
- Output: natural language statement

Concadia, Kreiss et al. 2023

Purpose: replacement for image (e.g. for visually impaired users, or if the image doesn't load)

Grocery store photo of several bunches of bananas.

Image Captioning

- Input: visual observation + context of caption's purpose?
- Output: natural language statement

Concadia, Kreiss et al. 2023

Purpose: as an illustrative example of bananas in the Wikipedia article about bananas

Cavendish bananas are the main commercial banana cultivars sold in the world market.

Image Captioning

- Input: visual observation + context of caption's purpose?
- Output: natural language statement

Concadia, Kreiss et al. 2023

Purpose: as an illustrative example of bananas in the Wikipedia article about the color yellow

Bananas, like autumn leaves, canaries, and egg yolks, get their yellow color from natural pigments called carotenoids.

Referring Expression (Refex) Resolution

- Inputs:
  - visual observation
  - natural language referring expression
- Output: region of the image corresponding to the refex referent

The leftmost person using sunglasses

Referring Expression (Refex) Resolution

- Inputs:
  - visual observation
  - natural language referring expression

- Output: region of the image corresponding to the refex referent

The leftmost person using sunglasses

Referring Expression (Refex) Resolution

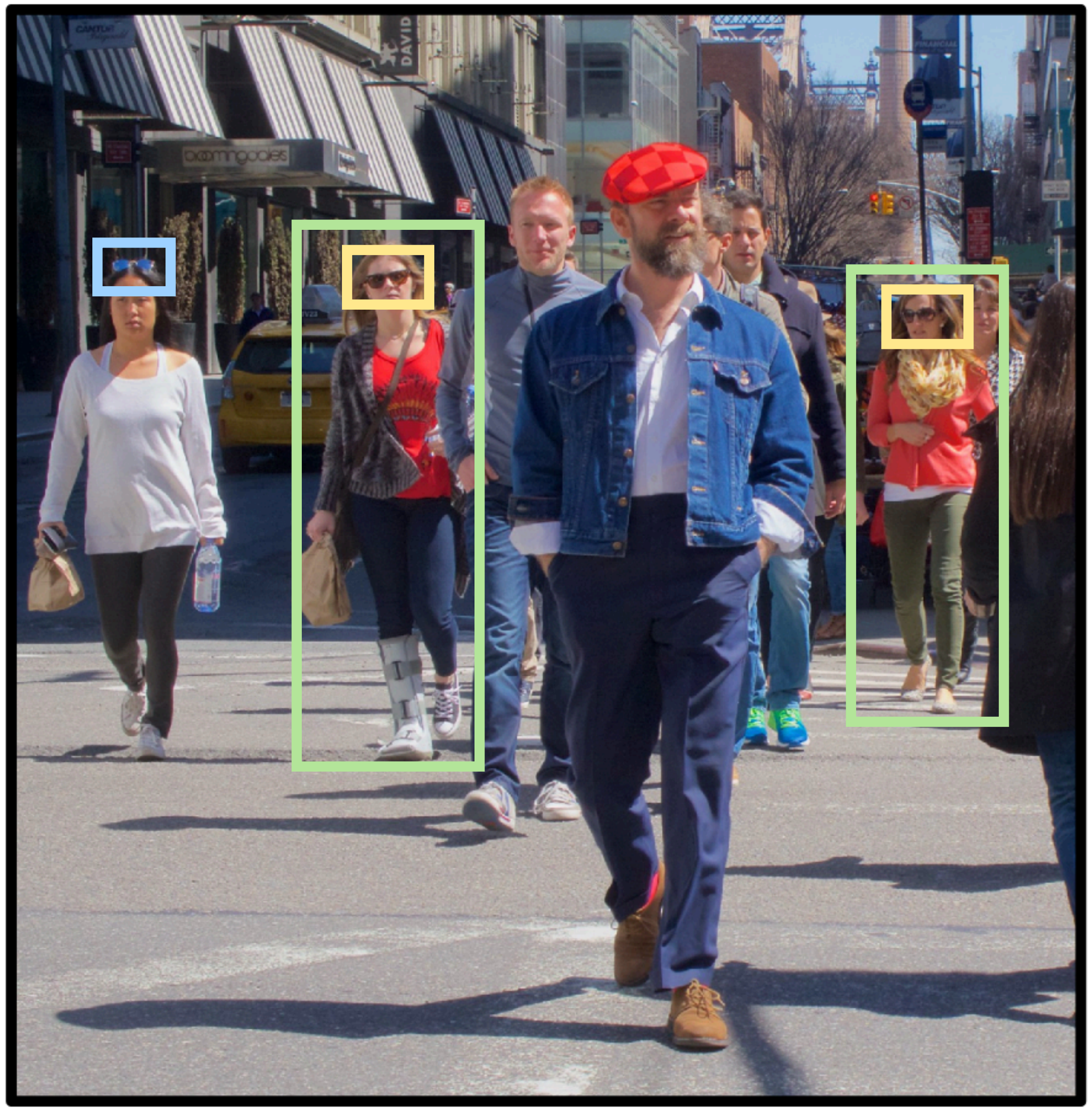
- Inputs:
  - visual observation
  - natural language referring expression
- Output: region of the image corresponding to the refex referent

The leftmost person using sunglasses

Referring Expression (Refex) Resolution

- Inputs:
  - visual observation
  - natural language referring expression
- Output: region of the image corresponding to the refex referent

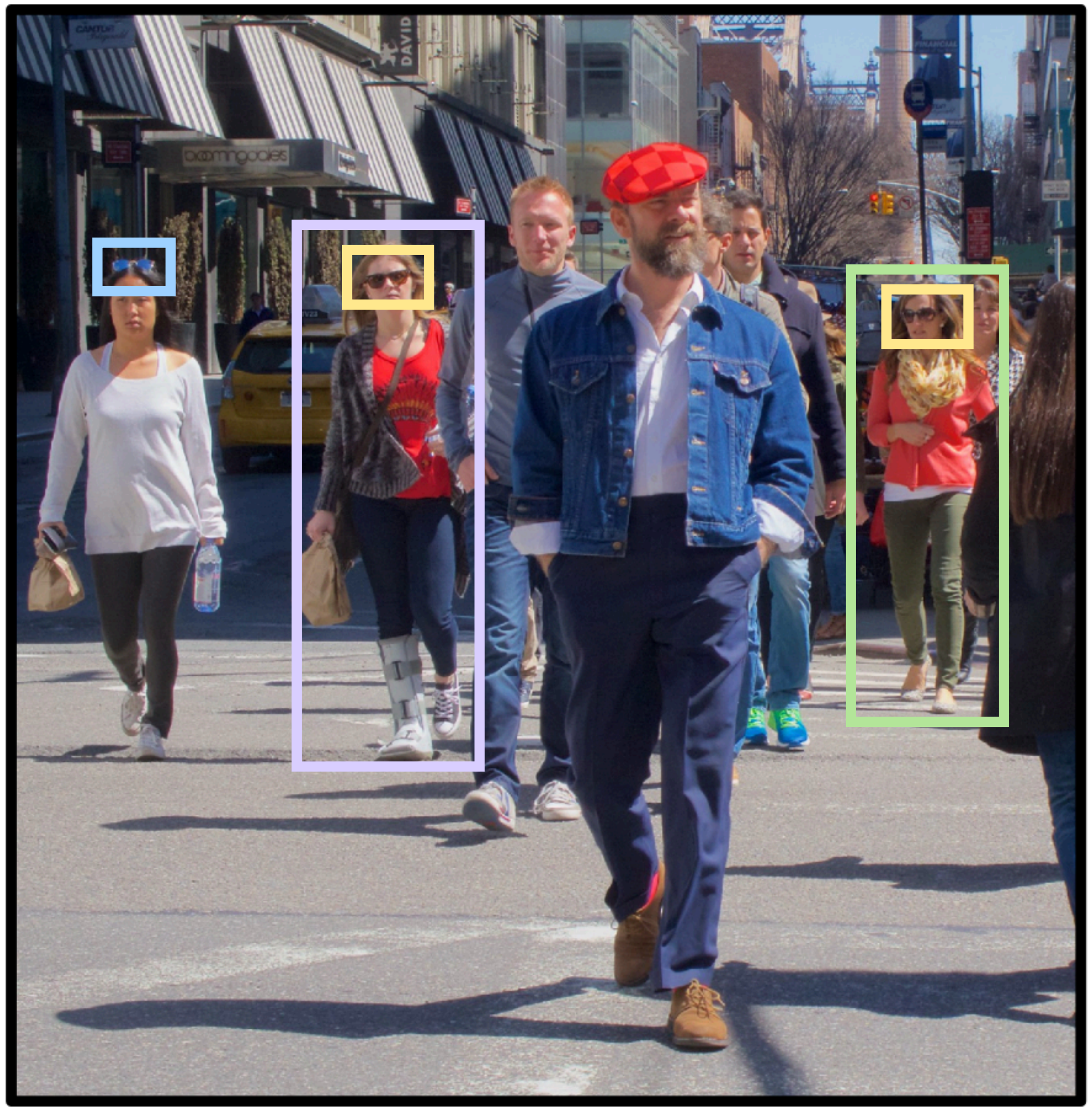
The leftmost person using sunglasses



Referring Expression (Refex) Resolution

- Inputs:
  - visual observation
  - natural language referring expression
- Output: region of the image corresponding to the refex referent

The leftmost person using sunglasses



Language Grounding Challenges

- Counting

There are <u>two</u>, and <u>only two</u>, people.

There are <u>no more than eight</u> bottles in total.

NLVR2, Suhr et al. 2019

Language Grounding Challenges

- Counting
- Compositionality
- Spatial relations

Each image contains just one bird, <u>and</u> the wires of a cage are <u>behind</u> the green bird.

Language Grounding Challenges

- Counting
- Compositionality
- Spatial relations
- Negation

A mitten is being worn in one image and the mittens are <u>not</u> being worn in the other image.

Language Grounding Challenges

- Counting
- Compositionality
- Spatial relations
- Negation
- Quantifiers

<u>Both</u> images show a silver pail being used as a flower vase.

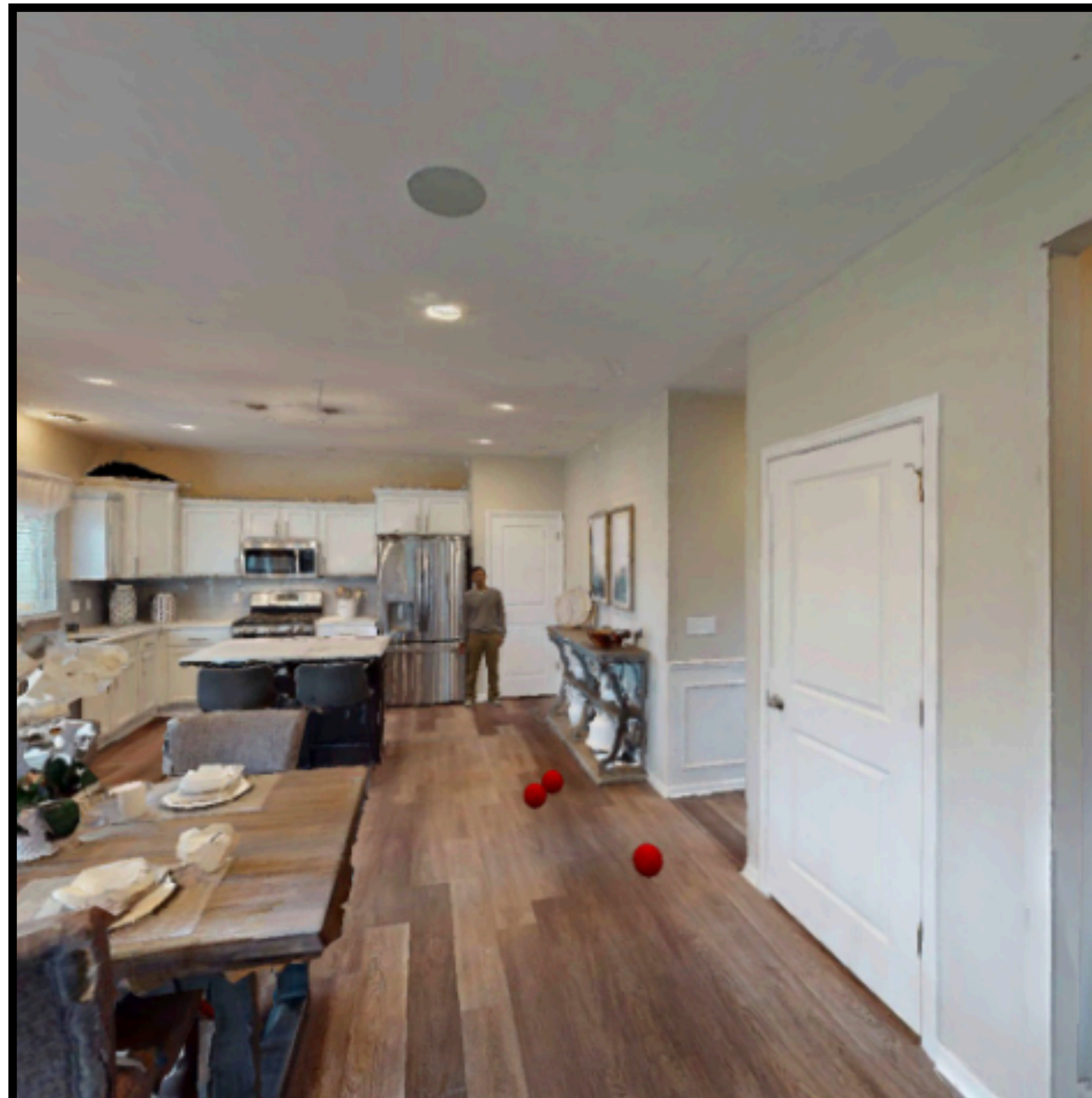
Language Grounding Challenges

- Counting
- Compositionality
- Spatial relations
- Negation
- Quantifiers
- Comparisons

The left image has 4 balloons of all <u>different</u> colors

Language Grounding Challenges

- Counting
- Compositionality
- Spatial relations
- Negation
- Quantifiers
- Comparisons
- Perspective-taking



The blue ball is near another red ball but it's further away from the painting

Vision-Language Models (VLMs)

Multimodality: model needs to be able to jointly process data in different <u>modalities</u>

- Text / language (discrete, small, information-dense)
- Visual observations (~continuous, much more data, but more self-redundant)
- Other structured data?

Representation Learning

- Text alone: $\phi(\overline{x})$
  - Bag-of-words
  - Bag-of-embeddings
  - Transformer
  - RNN

- Images alone: $\phi(I)$
  - Convolutional neural networks
  - Vision transformers

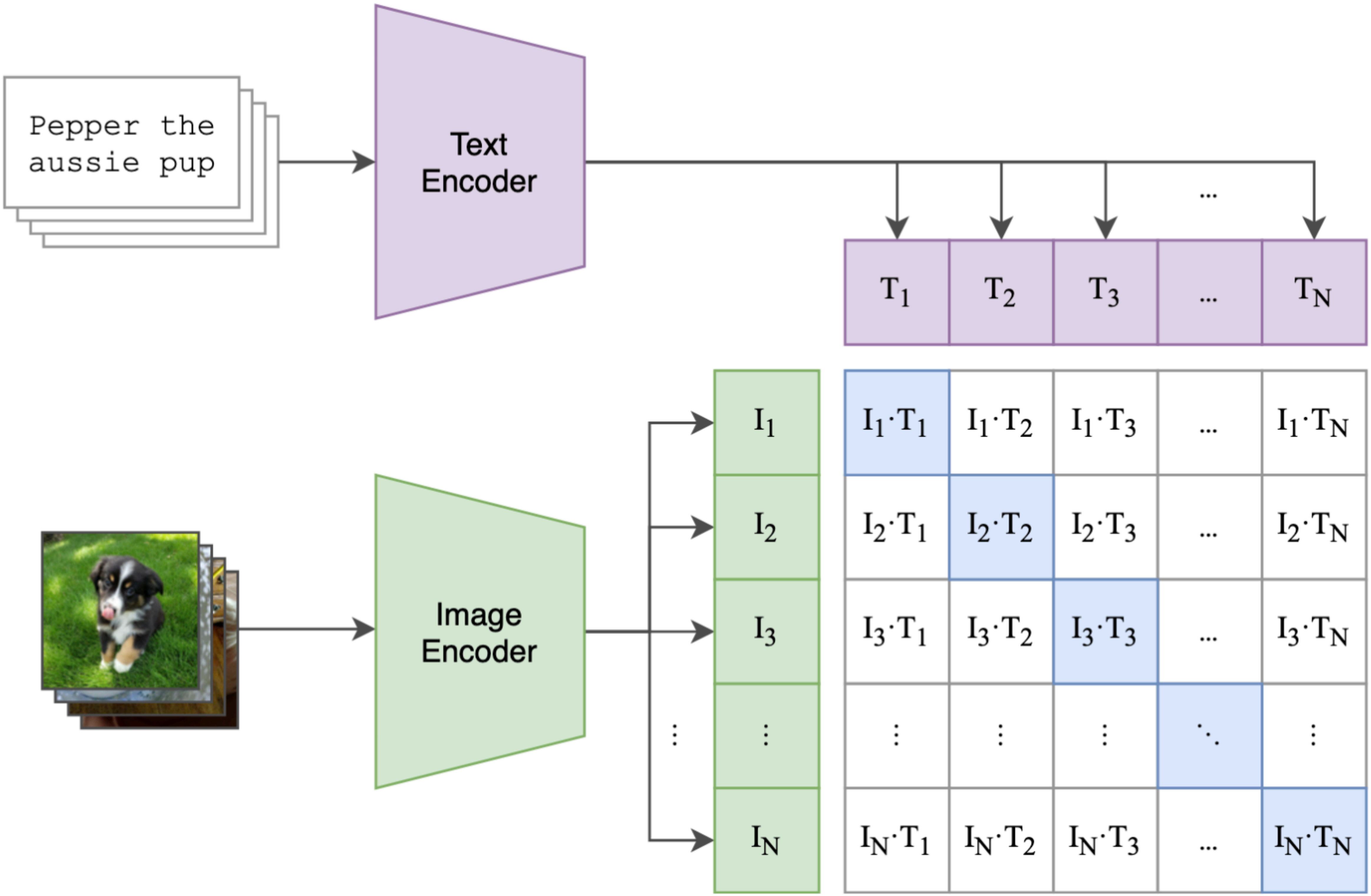
Main problem: how to learn the relationship between these two embedding spaces!

Contrastive Language-Image Pretraining (CLIP)

- **Goal:** find image embedding function $\phi(I)$ and text embedding function $\phi(\overline{x})$ such that:
  - For an image I with caption $\overline{x}$, the similarity between their embeddings is high
  - For any other image-caption pairs that are not attested, the similarity between their embeddings is low
- This results in “aligned” embedding functions: ideally, the embeddings of the image and caption for a known pair should be interchangeable

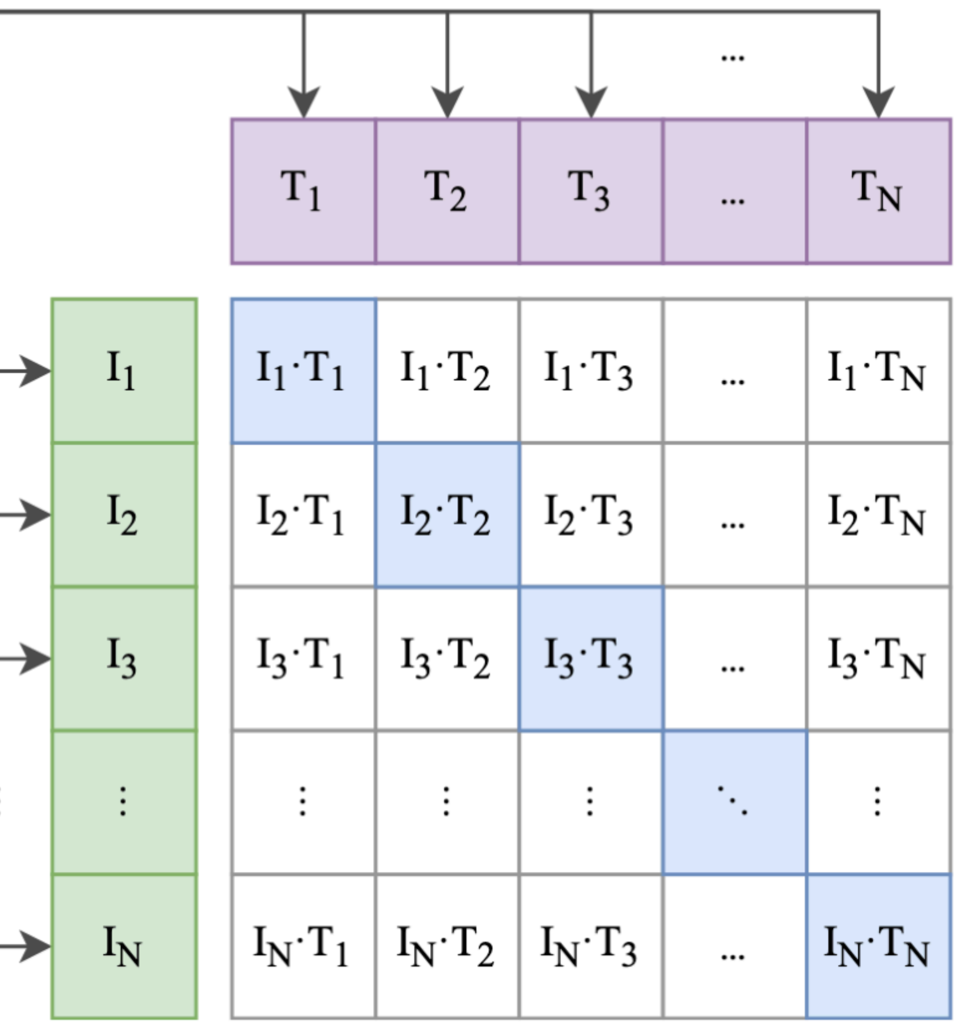
Radford et al. 2021

Contrastive Language-Image Pretraining (CLIP)



Radford et al. 2021

Contrastive Language-Image Pretraining (CLIP)

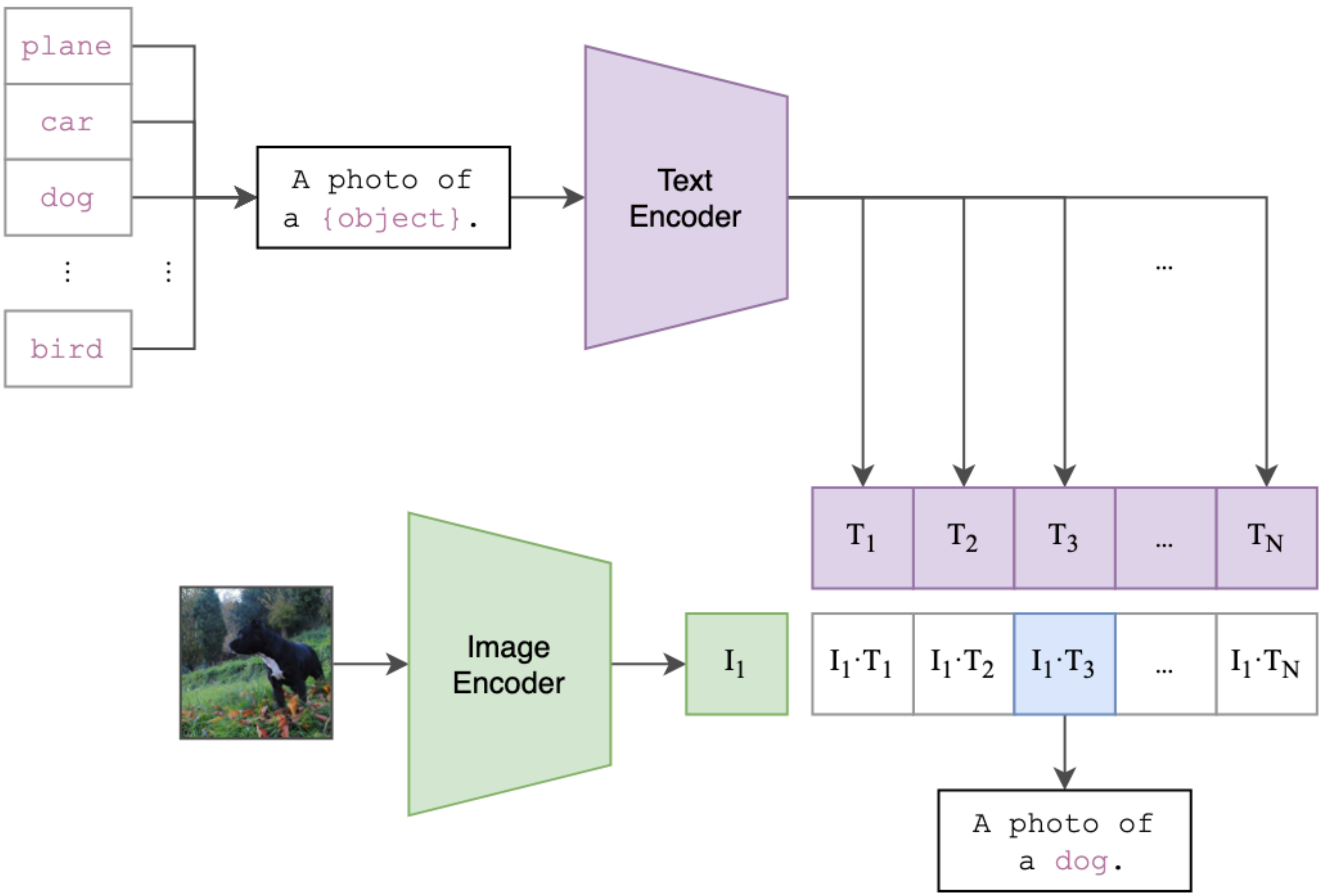



```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Radford et al. 2021

Contrastive Language-Image Pretraining (CLIP)

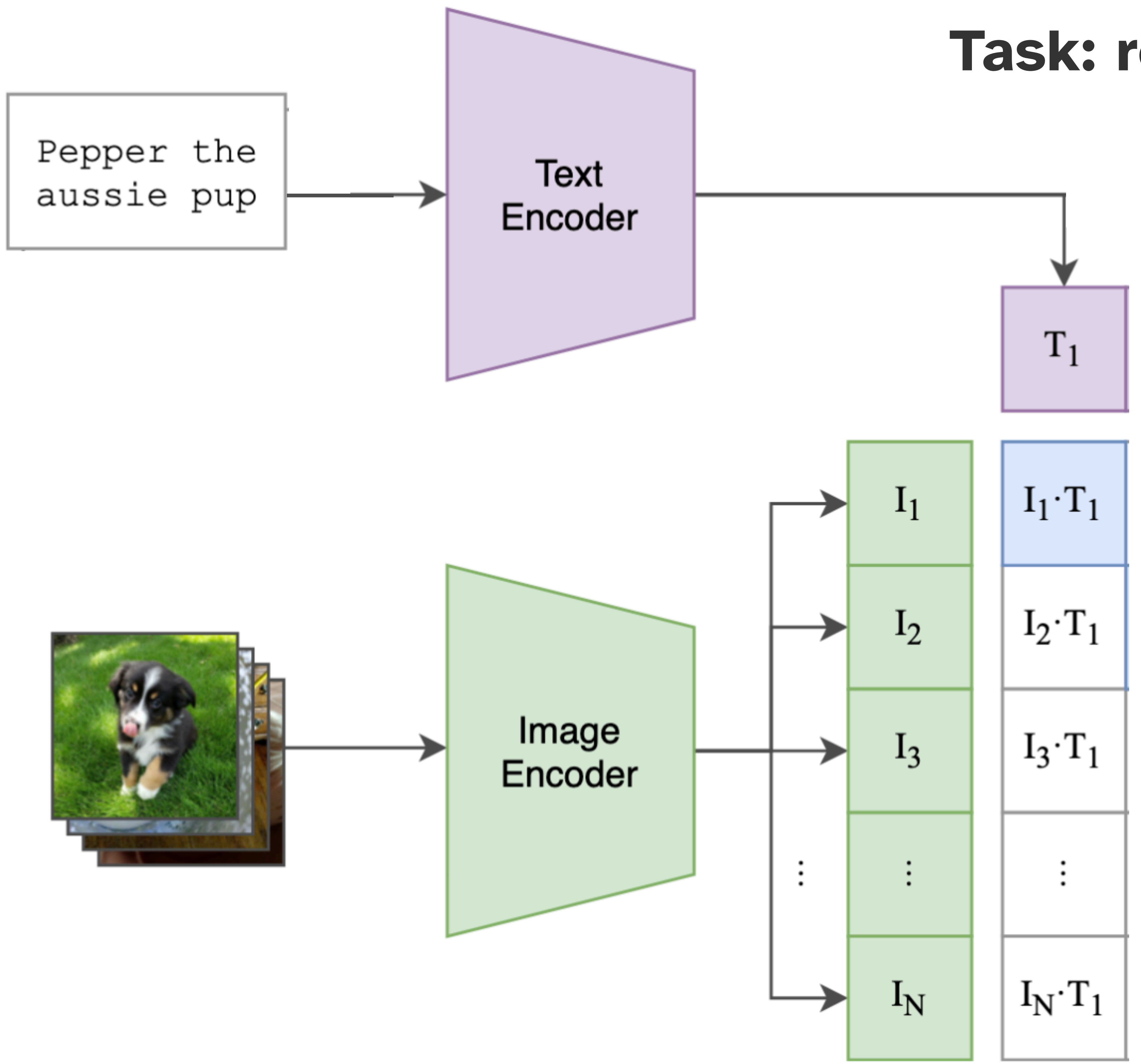
Task: classification



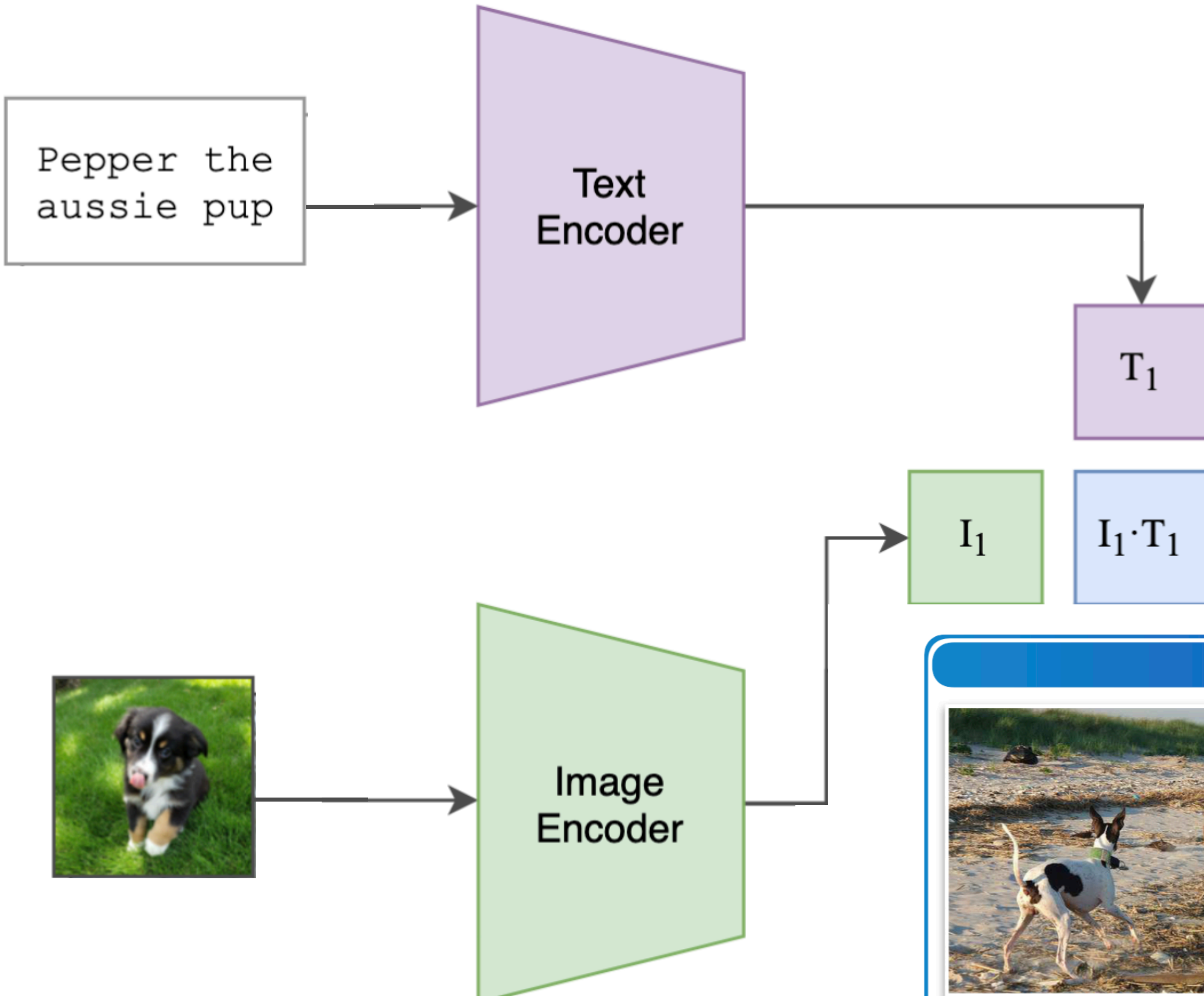
Radford et al. 2021

Contrastive Language-Image Pretraining (CLIP)

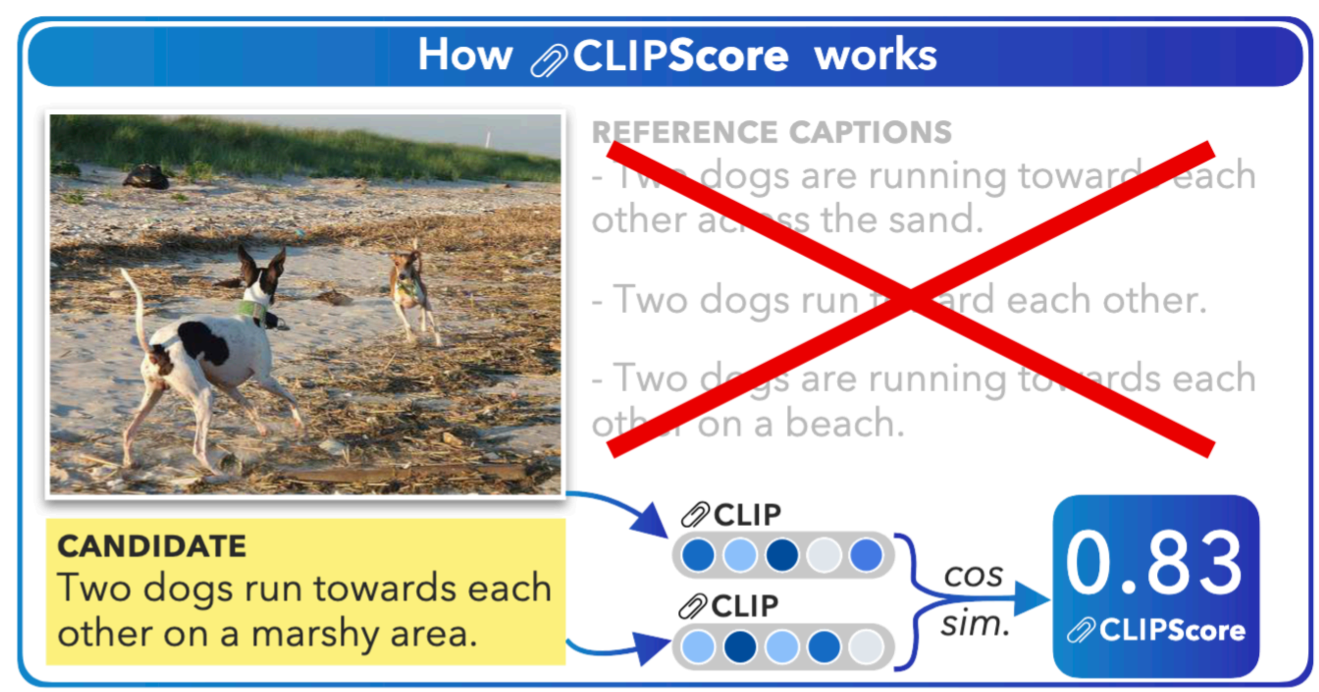
Task: retrieval



Contrastive Language-Image Pretraining (CLIP)

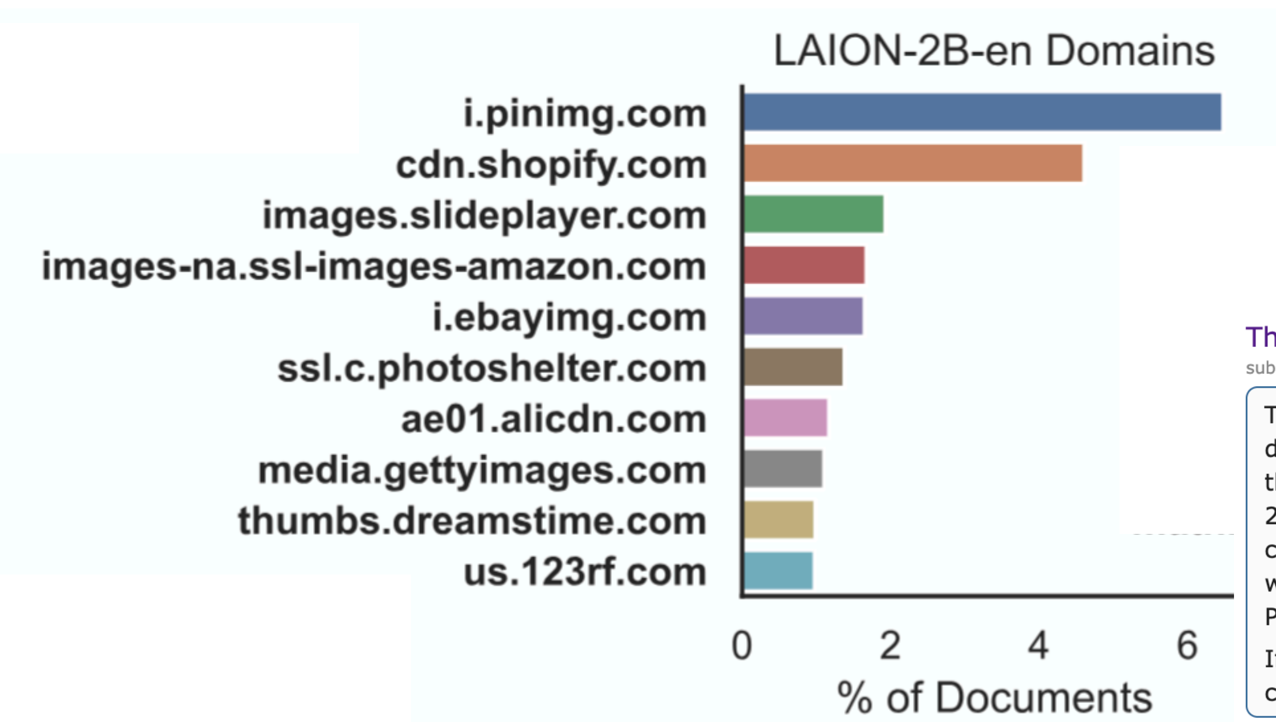


Task: image captioning evaluation



Limitations of CLIP

- Only representation learning — no parameters learned for any prediction tasks
- Representations only keep around information useful for the similarity task, and might discard:
  - Language data: word ordering
  - Image data: fine-grained details not mentioned in the text
- Training data: images on the web paired with alt text



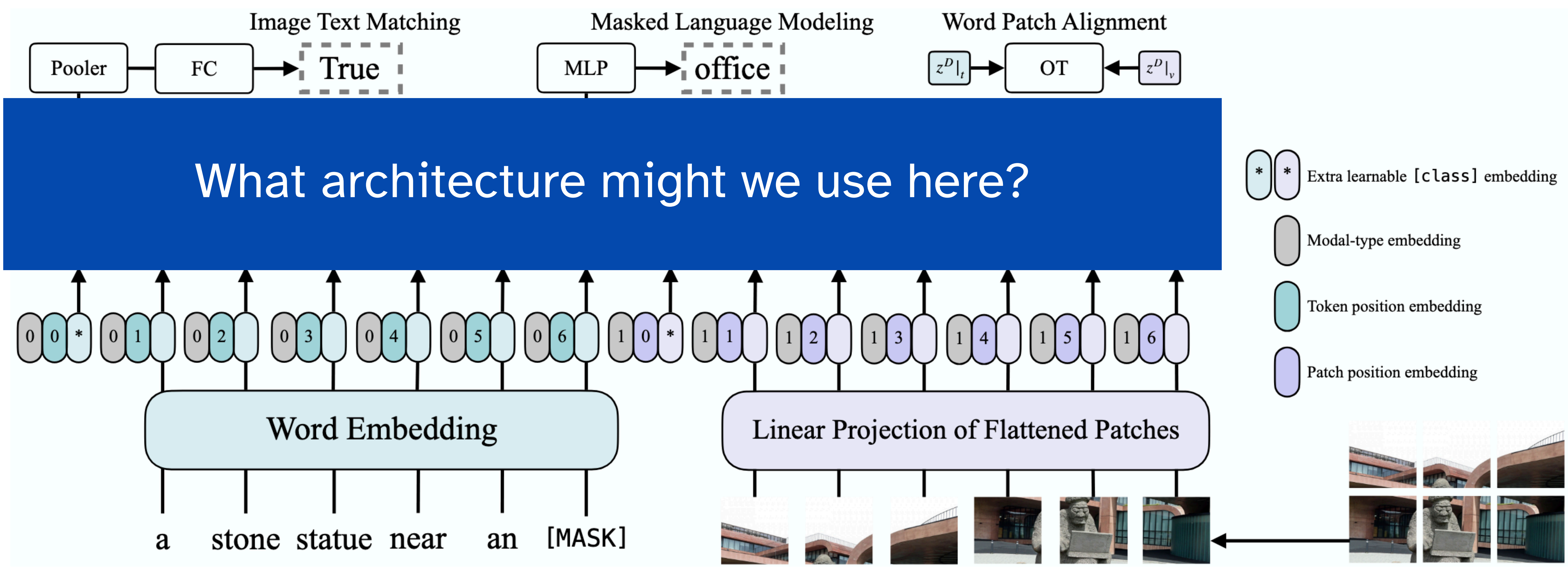

The Disappearance of Slideplayer.com [Fully Lost] Internet Media (self.lostmedia)
submitted 29 days ago by fallbunn001

This is not "lost media" in the sense of it being a single piece of media, but rather, a website that seems to have disappeared for good. Slideplayer.com was like Majhost or Brickshelf, but for PowerPoint Presentations. Any links to the website are now redirected to a crypto site called "Pied Piper." I last used the website in operating condition in late 2024. I discovered the redirect in June and immediately knew something was wrong. But I assumed it would be taken care of pretty quickly and decided to leave it be. It's almost November now. This leads me to believe the original website may have been abandoned by its owners and subsequently hacked. Or, its URL was acquired by this "Pied Piper" group, and they use it to redirect visitors to their crypto site.

If this is not the right subreddit for this type of post, I apologize. But I would greatly appreciate any insight you all could give.

Multimodal Models

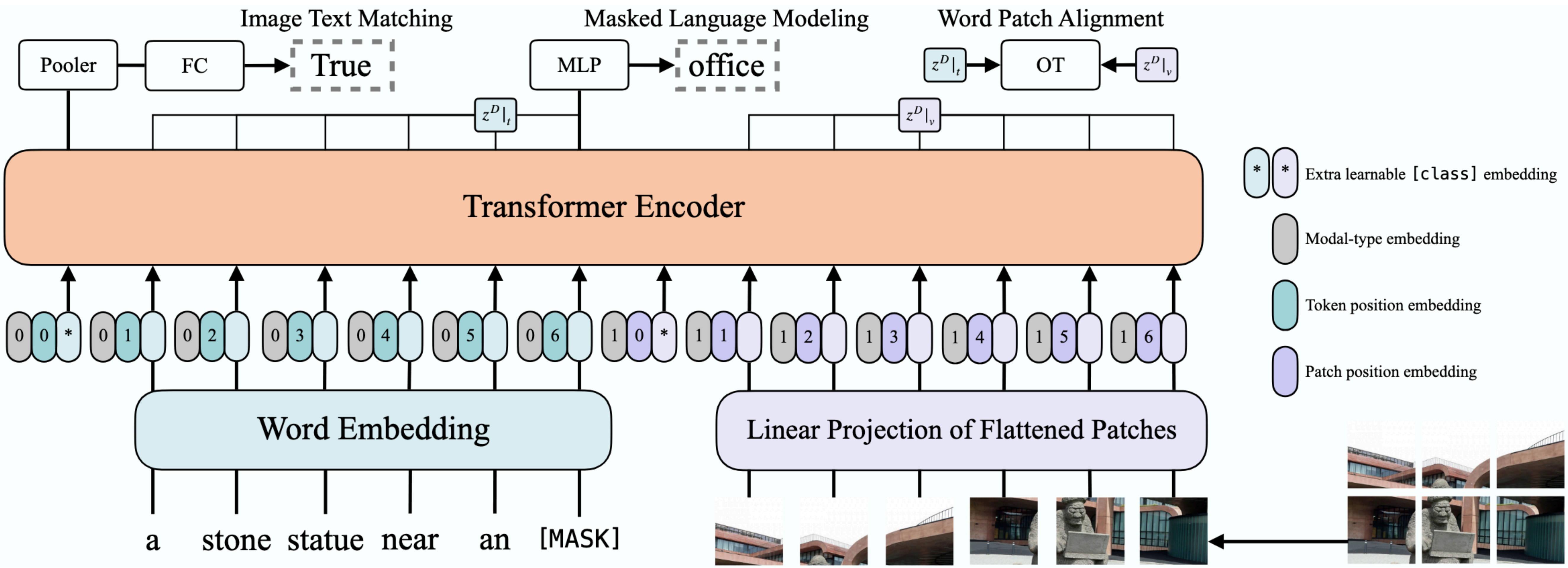
Goal: fuse representations from text and image to learn to perform language grounding tasks



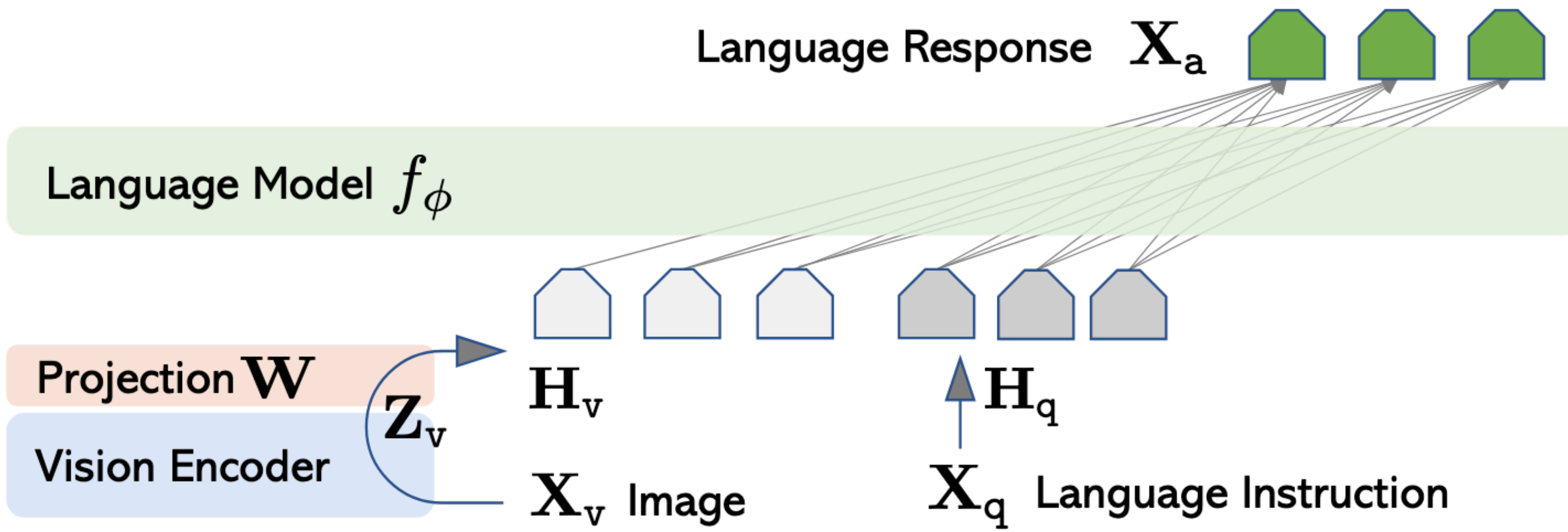
Kim et al. 2021

Vision and Language Transformer (ViLT)

Goal: fuse representations from text and image to learn to perform language grounding tasks



Visual Instruction Tuning (Llava)



Synthetic instruction-tuning data:

- Multi-turn conversations with "user" asking "assistant" questions about the image
- Question asking for a detailed description + detailed description as response
- Questions requiring in-depth reasoning + response and reasoning

Modular Approaches

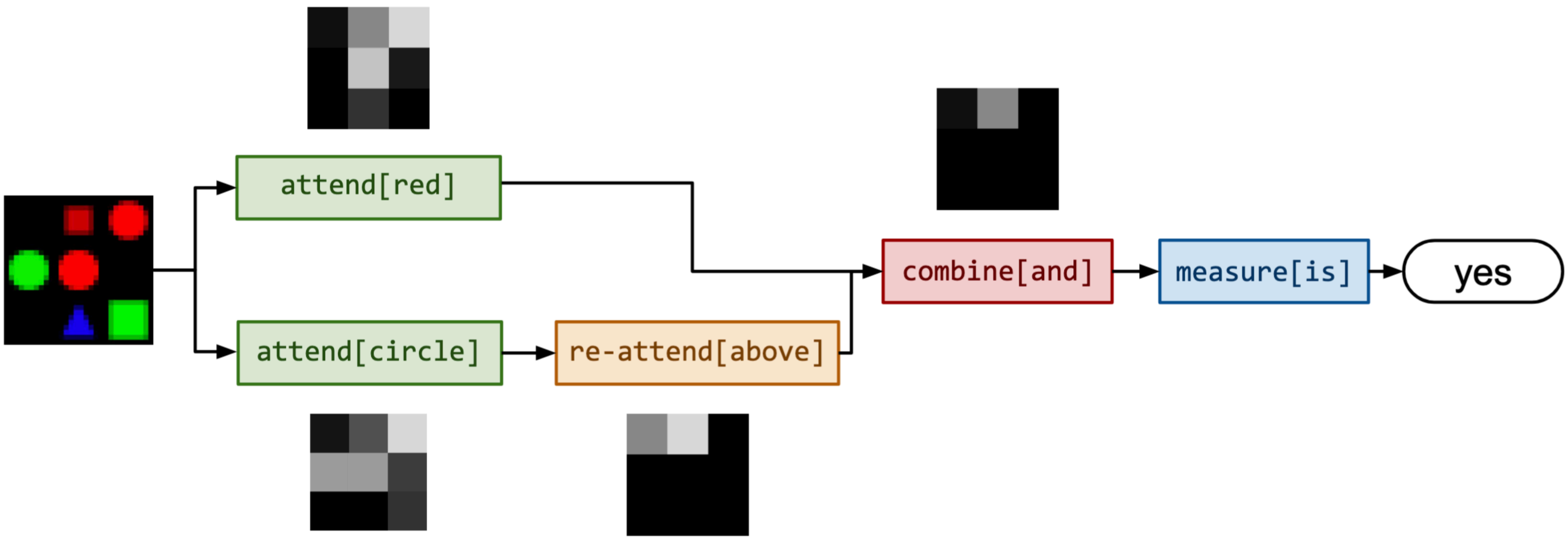
- VLMs still struggle with some grounding tasks:
  - Counting
  - Understanding spatial relations
  - Comparisons and superlatives
- But there are structured representations we can use that might give us more precise answers…
  - Language models are good at generating code
  - And we have pretty robust classical CV models, e.g. for object detection

Modular Approaches

Is there a red shape above a circle?

Modular Approaches

Is there a red shape above a circle?



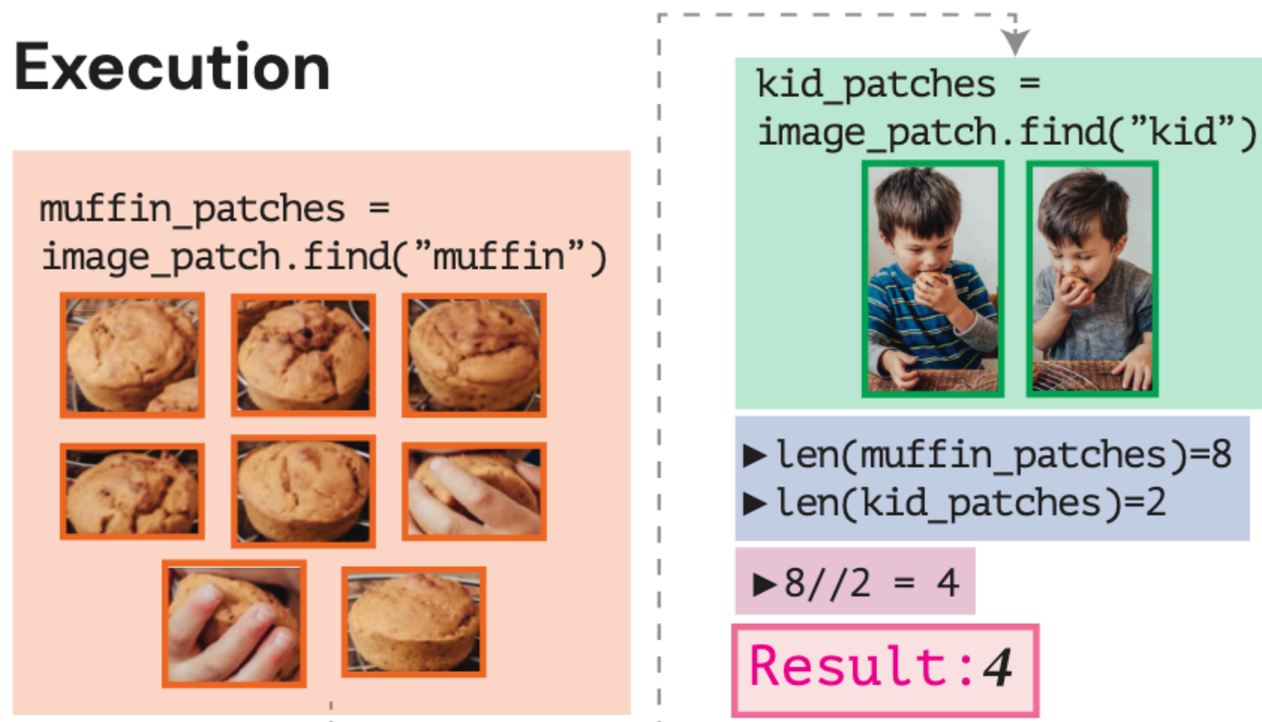
Modular Approaches

How many muffins can each kid have for it to be fair?

Generated Code

```
def execute_command(image):
    image_patch = ImagePatch(image)
    muffin_patches = image_patch.find("muffin")
    kid_patches = image_patch.find("kid")
    return str(len(muffin_patches) // len(kid_patches))
```

Execution



Drawback: Code Bottleneck

The potted plant is to the right of the bench.

Drawback:
Code Bottleneck

The potted plant is to the right of the bench.

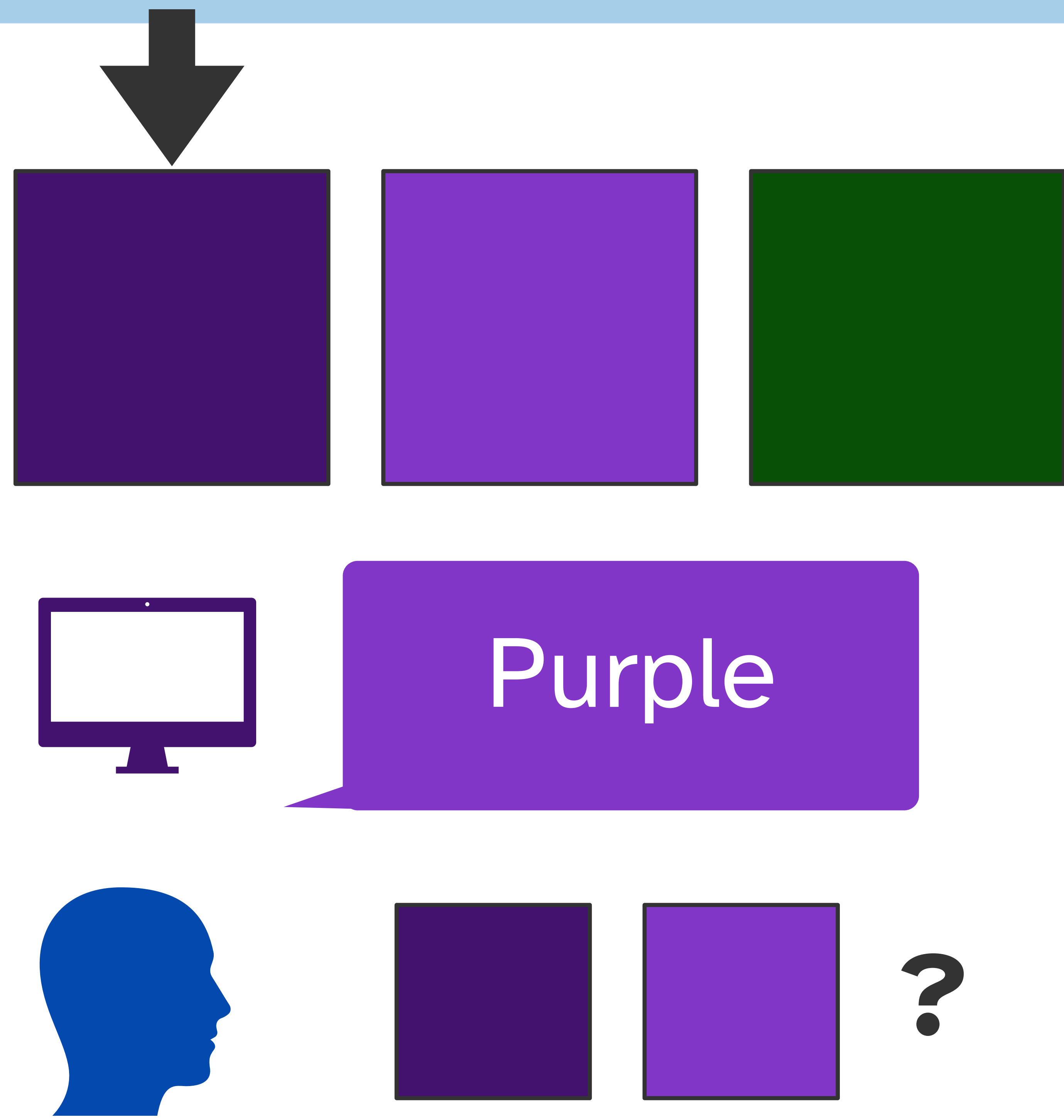
Drawback:
Code Bottleneck

The potted plant is to the right of the bench.

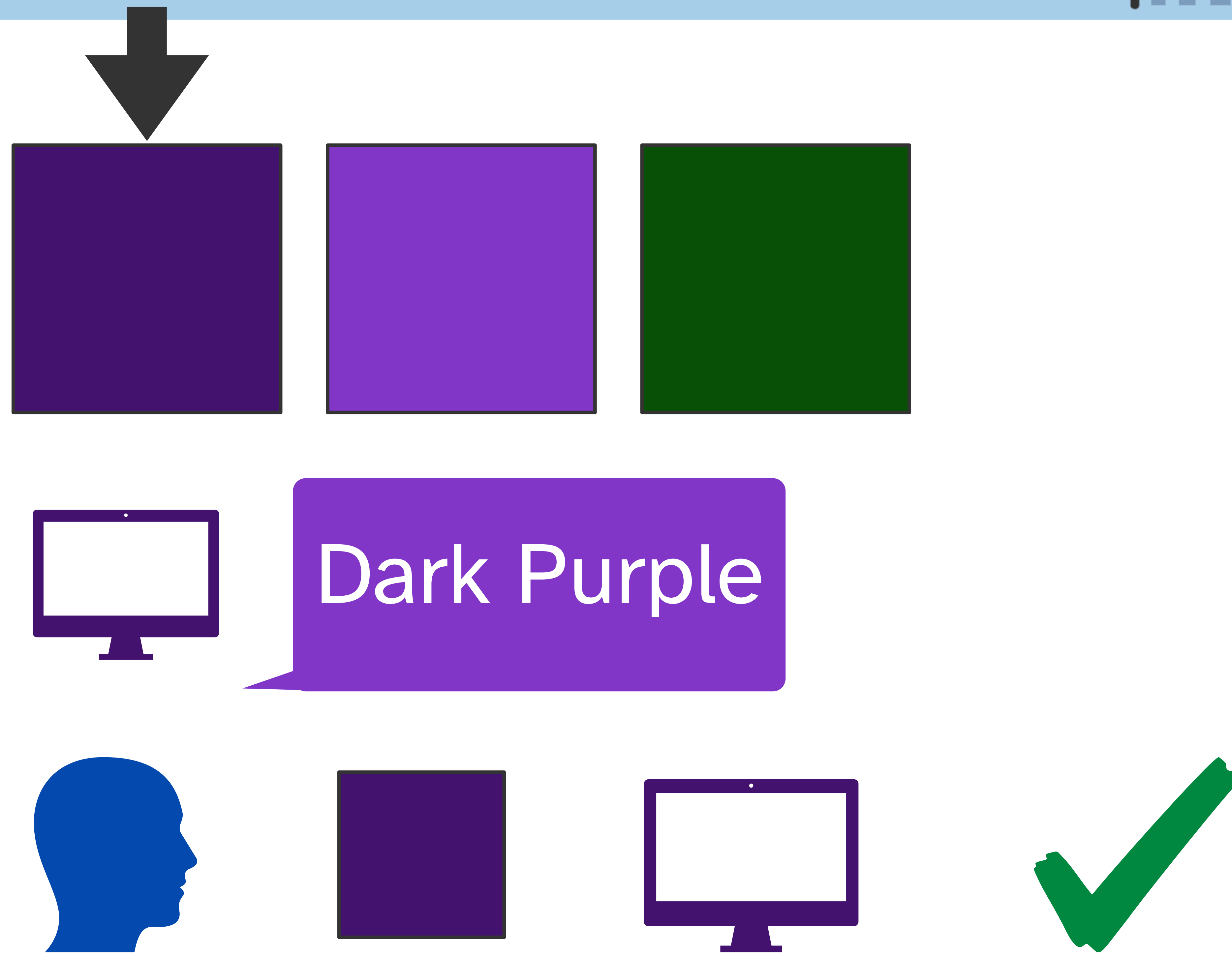
Pragmatics

- Now we have models that can do a lot of the vision-language tasks pretty well
  - Image-text entailment
  - Visual question answering
  - Image captioning
  - Referring expression resolution
- But recall: language is used in the context of other language users!

Pragmatics

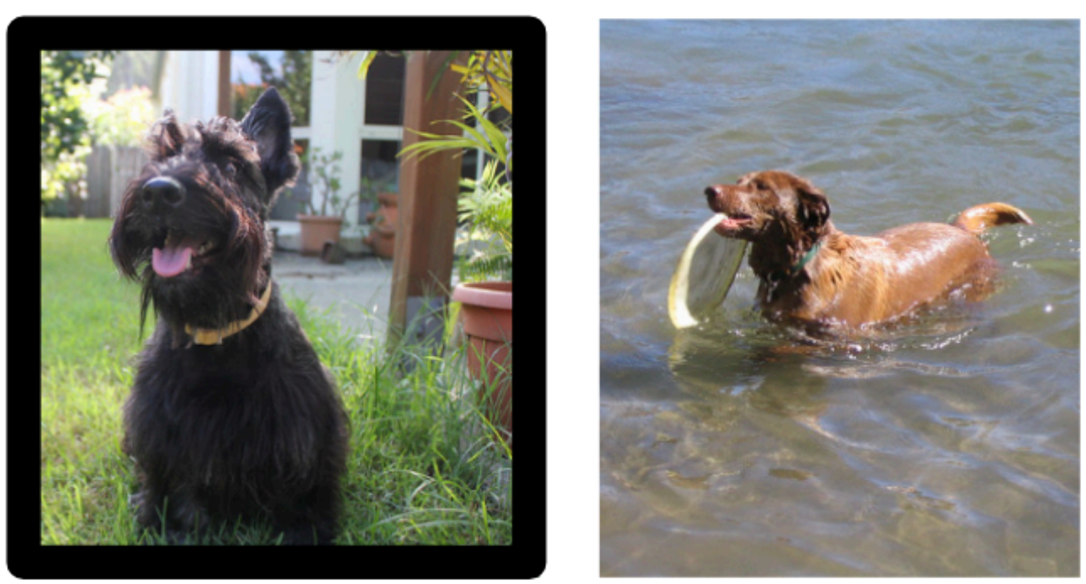


Pragmatics

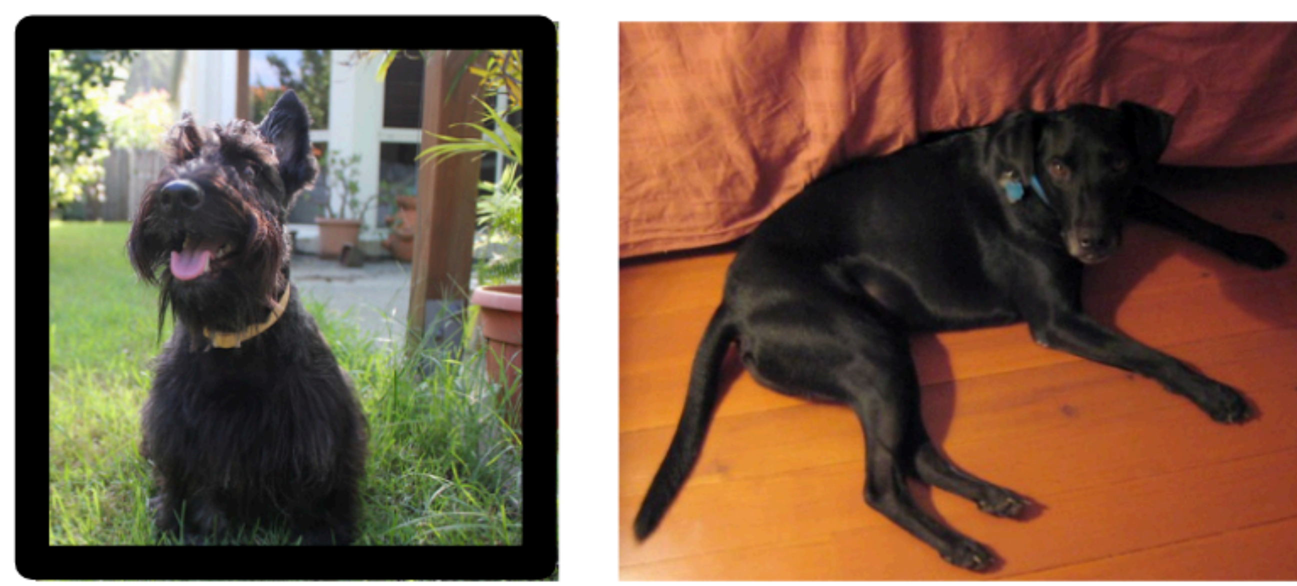


Reference Games

Ice skater

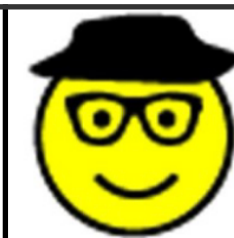


?



?

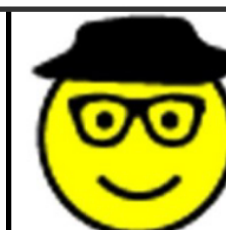
Speakers and Listeners

	 R1	 R2	 R3
[[*hat*]]			
[[*glasses*]]			
[[*mustache*]]			

Speakers and Listeners

	 R1	 R2	 R3
[[*hat*]]	0	0	1
[[*glasses*]]	0	1	1
[[*mustache*]]	0	0	0

$$p^{\text{Speaker}}_{\text{Literal}}(\cdot \mid r)$$

	 R1	 R2	 R3
hat			
glasses			
mustache			

Speakers and Listeners

	R1	R2	R3
[[*hat*]]	0	0	1
[[*glasses*]]	0	1	1
[[*mustache*]]	0	0	0

$$p^{\text{Listener}}_{\text{Literal}}(\cdot \mid x)$$

	R1	R2	R3
hat			
glasses			
mustache			

Speakers and Listeners

	R1	R2	R3
[[*hat*]]	0	0	1
[[*glasses*]]	0	1	1
[[*mustache*]]	0	0	0

$$p^{\text{Listener}}_{\text{Literal}}(\cdot \mid x)$$

	R1	R2	R3
hat			
glasses			
mustache			

Speakers and Listeners

	R1	R2	R3
[[*hat*]]	0	0	1
[[*glasses*]]	0	1	1
[[*mustache*]]	0	0	0

denotation of utterance

$$p^{\text{Listener}}_{\text{Literal}}(r \mid x) = \frac{\llbracket x \rrbracket_r}{\sum_{r' \in R} \llbracket x \rrbracket_{r'}}$$

sum over possible referents

	R1	R2	R3
hat			
glasses			
mustache			

Speakers and Listeners

	R1	R2	R3
hat	0	0	1
glasses	0	0.5	0.5
mustache	0	0	0

$$p^{\text{Listener}}_{\text{Literal}}(\cdot \mid x)$$

$$p^{\text{Speaker}}_{\text{Pragmatic}}(x \mid r) = \frac{p^{\text{Listener}}_{\text{Literal}}(r \mid x)}{\sum_{x' \in X} p^{\text{Listener}}_{\text{Literal}}(r \mid x')}$$

sum over possible utterances

	R1	R2	R3
hat			
glasses			
mustache			

Speakers and Listeners

	R1	R2	R3
hat	0	0	1
glasses	0	0.5	0.5
mustache	0	0	0

$$p^{\text{Speaker}}_{\text{Pragmatic}}(\cdot \mid r)$$

$$p^{\text{Listener}}_{\text{Pragmatic}}(r \mid x) = \frac{p^{\text{Speaker}}_{\text{Pragmatic}}(x \mid r)}{\sum_{r' \in R} p^{\text{Speaker}}_{\text{Pragmatic}}(x \mid r')}$$

sum over possible referents

	R1	R2	R3
hat			
glasses			
mustache			

Rational Speech Acts

- **Start with denotational semantics** that assigns a score to each utterance-referent pair, independent of context

Rational Speech Acts

- **Start with denotational semantics** that assigns a score to each utterance-referent pair, independent of context
- **Literal listener** uses denotational semantics to map each utterance to the probability of all referents

$$p^{\text{Listener}}_{\text{Literal}}(r \mid x) = \frac{[\![x]\!]_r}{\sum_{r' \in R} [\![x]\!]_{r'}}$$

Rational Speech Acts

- **Start with denotational semantics** that assigns a score to each utterance-referent pair, independent of context
- **Literal listener** uses denotational semantics to map each utterance to the probability of all referents

$$p^{\text{Listener}}_{\text{Literal}}(r \mid x) = \frac{[\![x]\!]_r}{\sum_{r' \in R} [\![x]\!]_{r'}}$$

- **Pragmatic speaker** re-normalizes probabilities over utterances given the literal listener's interpretations

$$p^{\text{Speaker}}_{\text{Pragmatic}}(x \mid r) = \frac{p^{\text{Listener}}_{\text{Literal}}(r \mid x)}{\sum_{x' \in X} p^{\text{Listener}}_{\text{Literal}}(r \mid x')}$$

Rational Speech Acts

- **Start with denotational semantics** that assigns a score to each utterance-referent pair, independent of context
- **Literal listener** uses denotational semantics to map each utterance to the probability of all referents

$$p_{\text{Literal}}^{\text{Listener}}(r \mid x) = \frac{[\![x]\!]_r}{\sum_{r' \in R} [\![x]\!]_{r'}}$$

- **Pragmatic speaker** re-normalizes probabilities over utterances given the literal listener's interpretations

$$p_{\text{Pragmatic}}^{\text{Speaker}}(x \mid r) = \frac{p_{\text{Literal}}^{\text{Listener}}(r \mid x)}{\sum_{x' \in X} p_{\text{Literal}}^{\text{Listener}}(r \mid x')}$$

- **Pragmatic listener** takes into account alternative utterances that the speaker *could* have used to refer to a referent, but didn't

$$p_{\text{Pragmatic}}^{\text{Listener}}(r \mid x) = \frac{p_{\text{Pragmatic}}^{\text{Speaker}}(x \mid r)}{\sum_{r' \in R} p_{\text{Pragmatic}}^{\text{Speaker}}(x \mid r')}$$