CS 784: Computational Linguistics
Lecture 16: Grounded Semantics

Freda Shi

School of Computer Science, University of Waterloo
fhs@uwaterloo.ca

March 13, 2025

Meanings in the Real World

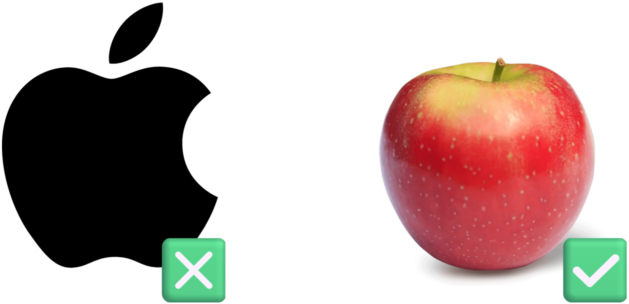
My favorite fruit is **apple**.

Meanings in the Real World

My favorite fruit is **apple**.

Meanings in the Real World

My favorite fruit is **apple**.



This is not purely ⟦apple⟧.

Meanings are grounded in the world.

Experience Grounds Language (Bisk et al., 2020)

*We posit that the present success of representation learning approaches trained on large, text-only corpora requires the parallel tradition of research on the broader **physical and social context** of language to address the deeper questions of communication.*

[Bisk, Y. et al., 2020. *Experience Grounds Language*. In *EMNLP*.]

What is Grounding?

Given the primary data source $\mathcal{X}$ and the ground $\mathcal{Y}$, **grounding** is the process of establishing a meaningful relationship between them.

It is implied that the mutual information $I(\mathcal{X};\mathcal{Y}) > 0$.

[Shi, H. F. 2024. *Learning language structures through grounding*. Ph.D. Thesis. Toyota Technological Institute at Chicago.]

What is Grounding?

Given the primary data source $\mathcal{X}$ and the ground $\mathcal{Y}$, **grounding** is the process of establishing a meaningful relationship between them.

It is implied that the mutual information $I(\mathcal{X};\mathcal{Y}) > 0$.

Usually, we also have $H(\mathcal{Y} \mid \mathcal{X}) > 0$—the ground is not specifically determined by the data source.

[Shi, H. F. 2024. *Learning language structures through grounding*. Ph.D. Thesis. Toyota Technological Institute at Chicago.]

What is Grounding?

Given the primary data source $\mathcal{X}$ and the ground $\mathcal{Y}$, **grounding** is the process of establishing a meaningful relationship between them.

It is implied that the mutual information $I(\mathcal{X};\mathcal{Y}) > 0$.

Usually, we also have $H(\mathcal{Y} \mid \mathcal{X}) > 0$—the ground is not specifically determined by the data source.

Examples

- $\mathcal{X}$: text, $\mathcal{Y}$: image (represent meaning of text with image)

[Chai, J. Y. et al. 2018. *Language to action: Towards interactive task learning with physical agents*. In *IJCAI*.]

[Shi, H. F. 2024. *Learning language structures through grounding*. Ph.D. Thesis. Toyota Technological Institute at Chicago.]

What is Grounding?

Given the primary data source $\mathcal{X}$ and the ground $\mathcal{Y}$, **grounding** is the process of establishing a meaningful relationship between them.

It is implied that the mutual information $I(\mathcal{X};\mathcal{Y}) > 0$.

Usually, we also have $H(\mathcal{Y} \mid \mathcal{X}) > 0$—the ground is not specifically determined by the data source.

Examples

- $\mathcal{X}$: text, $\mathcal{Y}$: image (represent meaning of text with image)
- $\mathcal{X}$: utterance from person A, $\mathcal{Y}$: mental state of person B (understanding the communication intention)

[Chai, J. Y. et al. 2018. *Language to action: Towards interactive task learning with physical agents*. In *IJCAI*.]

[Shi, H. F. 2024. *Learning language structures through grounding*. Ph.D. Thesis. Toyota Technological Institute at Chicago.]

What is Grounding?

Given the primary data source $\mathcal{X}$ and the ground $\mathcal{Y}$, **grounding** is the process of establishing a meaningful relationship between them.

It is implied that the mutual information $I(\mathcal{X};\mathcal{Y}) > 0$.

Usually, we also have $H(\mathcal{Y} \mid \mathcal{X}) > 0$—the ground is not specifically determined by the data source.

Examples

- $\mathcal{X}$: text, $\mathcal{Y}$: image (represent meaning of text with image)
- $\mathcal{X}$: utterance from person A, $\mathcal{Y}$: mental state of person B (understanding the communication intention)
- $\mathcal{X}$: text, $\mathcal{Y}$: audio (connecting text with corresponding audio)

[Chai, J. Y. et al. 2018. *Language to action: Towards interactive task learning with physical agents*. In *IJCAI*.]

[Shi, H. F. 2024. *Learning language structures through grounding*. Ph.D. Thesis. Toyota Technological Institute at Chicago.]

What is Grounding?

Given the primary data source $\mathcal{X}$ and the ground $\mathcal{Y}$, **grounding** is the process of establishing a meaningful relationship between them.

It is implied that the mutual information $I(\mathcal{X}; \mathcal{Y}) > 0$.

Usually, we also have $H(\mathcal{Y} \mid \mathcal{X}) > 0$—the ground is not specifically determined by the data source.

Examples

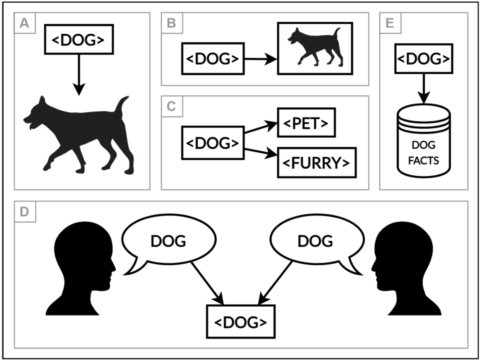
- $\mathcal{X}$: text, $\mathcal{Y}$: image (represent meaning of text with image)
- $\mathcal{X}$: utterance from person A, $\mathcal{Y}$: mental state of person B (understanding the communication intention)
- $\mathcal{X}$: text, $\mathcal{Y}$: audio (connecting text with corresponding audio)
- $\mathcal{X}$: image, $\mathcal{Y}$: text (image understanding with textual supervision)

[Chai, J. Y. et al. 2018. *Language to action: Towards interactive task learning with physical agents*. In *IJCAI*.]

[Shi, H. F. 2024. *Learning language structures through grounding*. Ph.D. Thesis. Toyota Technological Institute at Chicago.]

Grounding: A Comprehensive Taxonomy

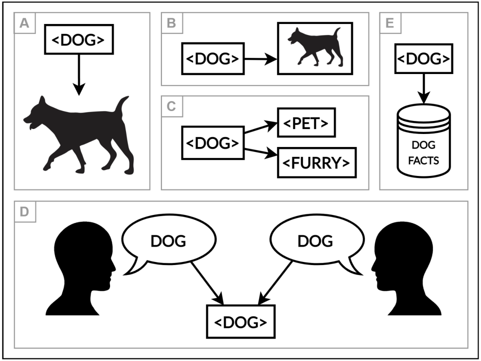
Grounding can be categorized into



[Figure credit: Mollo and Millière. The Vector Grounding Problem. https://arxiv.org/pdf/2304.01481]

Grounding: A Comprehensive Taxonomy

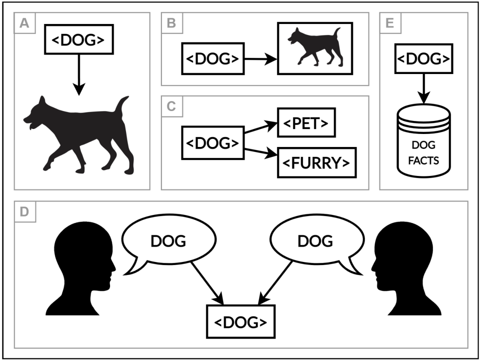
Grounding can be categorized into A. referential grounding,



[Figure credit: Mollo and Millière. The Vector Grounding Problem. https://arxiv.org/pdf/2304.01481]

Grounding: A Comprehensive Taxonomy

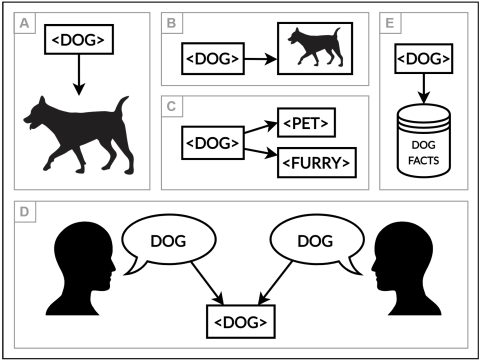
Grounding can be categorized into A. referential grounding, B. sensorimotor grounding,



[Figure credit: Mollo and Millière. The Vector Grounding Problem. https://arxiv.org/pdf/2304.01481]

Grounding: A Comprehensive Taxonomy

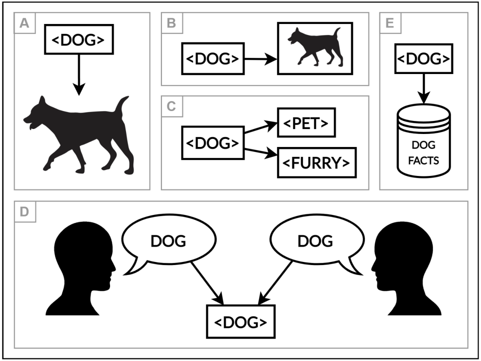
Grounding can be categorized into A. referential grounding, B. sensorimotor grounding, C. relational grounding,



[Figure credit: Mollo and Millière. The Vector Grounding Problem. https://arxiv.org/pdf/2304.01481]

Grounding: A Comprehensive Taxonomy

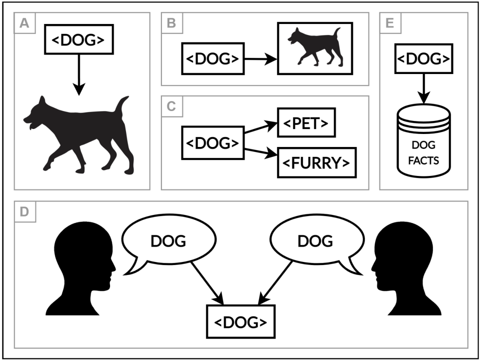
Grounding can be categorized into A. referential grounding, B. sensorimotor grounding, C. relational grounding, D. communicative grounding,



[Figure credit: Mollo and Millière. The Vector Grounding Problem. https://arxiv.org/pdf/2304.01481]

Grounding: A Comprehensive Taxonomy

Grounding can be categorized into A. referential grounding, B. sensorimotor grounding, C. relational grounding, D. communicative grounding, and E. epistemic grounding.



[Figure credit: Mollo and Millière. The Vector Grounding Problem. https://arxiv.org/pdf/2304.01481]

Grounding: This Lecture

Grounding is a broad topic that goes beyond semantics—communicative grounding is the key of pragmatics.

Grounding: This Lecture

Grounding is a broad topic that goes beyond semantics—communicative grounding is the key of pragmatics.

However, in this lecture, we focus on **semantic grounding** (or more specifically, sensorimotor grounding): representing meanings of text with data from other modalities (e.g., images).

Recap: (Ungrounded) Pure-Text Language Models

Two popular types of (ungrounded) pure-text language models:

- Autoregressive models (e.g., GPT):

$$P_{\Theta}(w_i \mid w_1, \ldots, w_{i-1})$$

[Radford, A. et al. 2018. *Improving language understanding by generative pretraining*.]

Recap: (Ungrounded) Pure-Text Language Models

Two popular types of (ungrounded) pure-text language models:

- Autoregressive models (e.g., GPT):

$$P_\Theta(w_i \mid w_1, \ldots, w_{i-1})$$

- Masked language models (e.g., BERT):

$$P_\Theta(w_i \mid w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n)$$

[Radford, A. et al. 2018. *Improving language understanding by generative pretraining*.]

[Devlin, J. et al. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *NAACL*.]

Recap: (Ungrounded) Pure-Text Language Models

Two popular types of (ungrounded) pure-text language models:

- Autoregressive models (e.g., GPT):

$$P_\Theta(w_i \mid w_1, \ldots, w_{i-1})$$

- Masked language models (e.g., BERT):

$$P_\Theta(w_i \mid w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n)$$

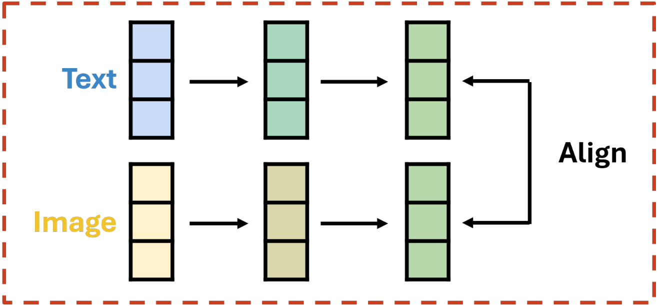
Whether these pure-text language models encode meaning, and to what extent, is still under debate.

[Radford, A. et al. 2018. *Improving language understanding by generative pretraining*.]

[Devlin, J. et al. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *NAACL*.]

Joint Visual-Semantic Embedding Space

Encode visual and textual information into a **shared space**.



Learning Joint Visual Semantic Space

Training data: pairs of images and text descriptions.

[Kiros, R. et al. 2014. *Unifying visual-semantic embeddings with multimodal neural language models*.]

Learning Joint Visual Semantic Space

Training data: pairs of images and text descriptions.

Core idea: encode images and text into a **joint embedding space** by minimizing the hinge-based triplet loss.

$$\Theta^* = \arg\min_{\Theta} \sum_{(I^+, T^+, T^-)} \max\left(0, \alpha - \text{sim}(I_{\Theta}^+, T_{\Theta}^+) + \text{sim}(I_{\Theta}^+, T_{\Theta}^-)\right) + \sum_{(T^+, I^+, I^-)} \max\left(0, \alpha - \text{sim}(T_{\Theta}^+, I_{\Theta}^+) + \text{sim}(T_{\Theta}^+, I_{\Theta}^-)\right)$$

[Kiros, R. et al. 2014. *Unifying visual-semantic embeddings with multimodal neural language models*.]

Learning Joint Visual Semantic Space

Training data: pairs of images and text descriptions.

Core idea: encode images and text into a **joint embedding space** by minimizing the hinge-based triplet loss.

$$\Theta^* = \arg\min_{\Theta} \sum_{(I^+, T^+, T^-)} \max\left(0, \alpha - \mathrm{sim}(I^+_\Theta, T^+_\Theta) + \mathrm{sim}(I^+_\Theta, T^-_\Theta)\right) + \sum_{(T^+, I^+, I^-)} \max\left(0, \alpha - \mathrm{sim}(T^+_\Theta, I^+_\Theta) + \mathrm{sim}(T^+_\Theta, I^-_\Theta)\right)$$

I^+: 

I^-: 

T^+: *There is a cat standing on the lawn.*

T^-: *There is an apple on the table.*

[Kiros, R. et al. 2014. *Unifying visual-semantic embeddings with multimodal neural language models*.]

Properties of the Joint Space

Images and text descriptions are close in the joint space if they are semantically related.

Properties of the Joint Space

Images and text descriptions are close in the joint space if they are semantically related.

Example Applications:

- Bidirectional image-caption retrieval: encode the query (image or text), and the "database" into the joint space and retrieve the closest neighbors.

Properties of the Joint Space

Images and text descriptions are close in the joint space if they are semantically related.

Example Applications:

- Bidirectional image-caption retrieval: encode the query (image or text), and the "database" into the joint space and retrieve the closest neighbors.
- Image captioning: encode the image into the joint space, and train a decoder to generate text conditioned on the image encoding.

Text in the training corpus can be at any level of granularity (e.g., word, phrase, sentence, paragraph).

Variations of Training Objective: Hard Negative Mining

Original:

$$\Theta^* = \arg\min_{\Theta} \sum_{(I^+, T^+, T^-)} \left[\alpha - \text{sim}(I^+_\Theta, T^+_\Theta + \text{sim}(I^+_\Theta, T^-_\Theta))\right]_+$$

$$+ \sum_{(T^+, I^+, I^-)} \left[\alpha - \text{sim}(T^+_\Theta, I^+_\Theta) + \text{sim}(T^+_\Theta, I^-_\Theta)\right]_+$$

$$[\cdot]_+ = \max(0, \cdot)$$

Variations of Training Objective: Hard Negative Mining

Original:

$$\Theta^* = \arg\min_{\Theta} \sum_{(I^+, T^+, T^-)} \left[\alpha - \text{sim}(I^+_\Theta, T^+_\Theta + \text{sim}(I^+_\Theta, T^-_\Theta))\right]_+ + \sum_{(T^+, I^+, I^-)} \left[\alpha - \text{sim}(T^+_\Theta, I^+_\Theta) + \text{sim}(T^+_\Theta, I^-_\Theta)\right]_+$$

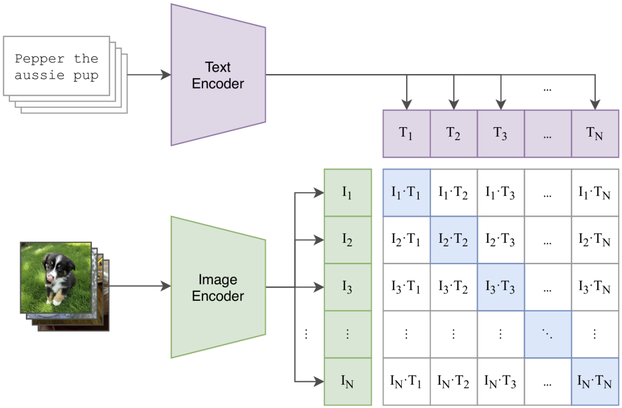
$$[\cdot]_+ = \max(0, \cdot)$$

Modified:

$$\Theta^* = \arg\min_{\Theta} \sum_{(I^+, T^+)} \max_{T^-} \left[\alpha - \text{sim}(I^+_\Theta, T^+_\Theta) + \text{sim}(I^+_\Theta, T^-_\Theta)\right]_+ + \sum_{(T^+, I^+)} \max_{I^-} \left[\alpha - \text{sim}(T^+_\Theta, I^+_\Theta) + \text{sim}(T^+_\Theta, I^-_\Theta)\right]_+$$

[Faghri, F. et al. 2017. *VSE++: Improving visual-semantic embeddings with hard negatives*. In *BMVC*.]

Variations of Training Objective: Contrastive Learning



[Radford, A. et al. 2021. *Learning transferable visual models from natural language supervision.*]

Variations of Training Objective: Contrastive Learning

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{[(I_1,T_1),\ldots(I_n,T_n)]} \left[\sum_i -\log P_\Theta(T_i \mid I_i; [T_{1\ldots n}]) - \log P_\Theta(I_i \mid T_i; [I_{1\ldots n}]) \right]$$

Variations of Training Objective: Contrastive Learning

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{[(I_1,T_1),\dots(I_n,T_n)]} \left[\sum_i -\log P_\Theta(T_i \mid I_i; [T_{1\dots n}]) - \log P_\Theta(I_i \mid T_i; [I_{1\dots n}]) \right]$$

$[(I_1, T_1), \dots, (I_n, T_n)]$: a batch of image-text pairs.

Variations of Training Objective: Contrastive Learning

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{[(I_1,T_1),\ldots(I_n,T_n)]} \left[\sum_i -\log P_\Theta(T_i \mid I_i; [T_{1\ldots n}]) - \log P_\Theta(I_i \mid T_i; [I_{1\ldots n}]) \right]$$

$[(I_1, T_1), \ldots, (I_n, T_n)]$: a batch of image-text pairs.

There exists a probabilistic interpretation of the training loss.

Variations of Training Objective: Contrastive Learning

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{[(I_1,T_1),\ldots(I_n,T_n)]} \left[\sum_i -\log P_\Theta(T_i \mid I_i; [T_{1\ldots n}]) - \log P_\Theta(I_i \mid T_i; [I_{1\ldots n}])\right]$$

$[(I_1, T_1), \ldots, (I_n, T_n)]$: a batch of image-text pairs.

There exists a probabilistic interpretation of the training loss.

- **Query**: image I_i.
- **Database**: text descriptions $T_1, \ldots, T_n$.
- **Ground truth**: T_i.

$P_\Theta(T_i \mid I_i; [T_{1\ldots n}])$: the probability of T_i being the correct retrieval result in the above settings.

Variations of Training Objective: Contrastive Learning

The softmax function converts a list of real values (e.g., $\mathbf{x} \in \mathbb{R}^n$) to a probability distribution.

$$\mathrm{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

Variations of Training Objective: Contrastive Learning

The softmax function converts a list of real values (e.g., $\mathbf{x} \in \mathbb{R}^n$) to a probability distribution.

$$\mathrm{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

$$P_\Theta(T_i \mid I_i; [T_{1\ldots n}]) = \frac{\exp(\langle I_i, T_i \rangle)}{\sum_{j=1}^{n} \exp(\langle I_i, T_j \rangle)}$$

Variations of Training Objective: Contrastive Learning

The softmax function converts a list of real values (e.g., $\mathbf{x} \in \mathbb{R}^n$) to a probability distribution.

$$\mathrm{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

$$P_\Theta(T_i \mid I_i; [T_{1\ldots n}]) = \frac{\exp(\langle I_i, T_i \rangle)}{\sum_{j=1}^{n} \exp(\langle I_i, T_j \rangle)} = [\mathrm{softmax}(I_i^\intercal [T_{1\ldots n}])]_i$$

Visually Grounded Grammar Induction

Input: Captioned images.

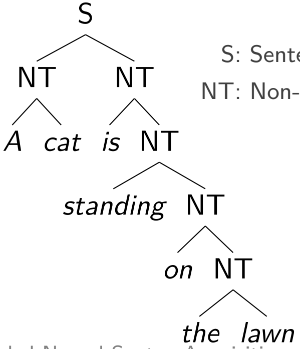
A cat is standing on the lawn.

[Source: Shi et al. 2019. Visually Grounded Neural Syntax Acquisition. In ACL.]

Visually Grounded Grammar Induction

Input: Captioned images.
Output: Linguistically plausible structure for captions.

A cat is standing on the lawn.

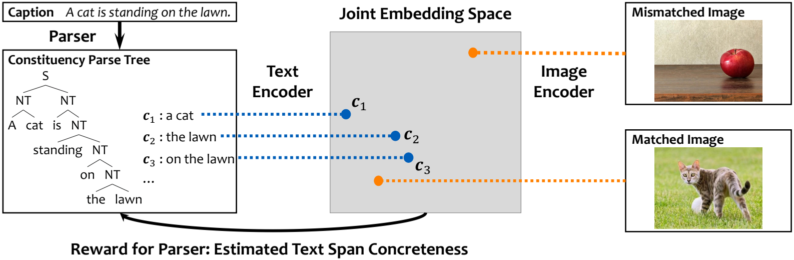


S: Sentence
NT: Non-Terminal

[Source: Shi et al. 2019. Visually Grounded Neural Syntax Acquisition. In ACL.]

Grounded Signals for Syntax Acquisition

Hypothesis: more visually concrete word spans are more likely to be constituents.



Concreteness Estimation with the Joint Embedding Space

$$\ell(c; i, i') = \mathrm{sim}(i', c) - \mathrm{sim}(i, c)$$

Concreteness Estimation with the Joint Embedding Space

Image i

Candidate Constituent c

a cat

on the

$$\ell(c; i, i') = \mathrm{sim}(i', c) - \mathrm{sim}(i, c)$$

Concreteness Estimation with the Joint Embedding Space

Image i

Candidate Constituent c

a cat
on the

Another Image i'

$$\ell(c; i, i') = \mathrm{sim}(i', c) - \mathrm{sim}(i, c)$$

Value of ℓ

Concreteness Estimation with the Joint Embedding Space

Image i

Candidate Constituent c

a cat

on the

Another Image i'

$$\ell(c; i, i') = \mathrm{sim}(i', c) - \mathrm{sim}(i, c)$$

Value of ℓ

sim( , *a cat*) = 0.2 sim( , *a cat*) = 0.9 $\ell = -0.7$

Concreteness Estimation with the Joint Embedding Space

Image i

Candidate Constituent c

a cat

on the

Another Image i'

$$\ell(c; i, i') = \mathrm{sim}(i', c) - \mathrm{sim}(i, c)$$

Value of ℓ

sim( , *a cat*) = 0.2 sim( , *a cat*) = 0.9 $\ell = -0.7$

sim( , *on the*) = 0.4 sim( , *on the*) = 0.4 $\ell = 0$

Concreteness Estimation with the Joint Embedding Space

Image i

Candidate Constituent c

a cat

on the

Another Image i'

$$\ell(c; i, i') = \mathrm{sim}(i', c) - \mathrm{sim}(i, c)$$

Value of ℓ

sim( , *a cat*) = 0.2 sim( , *a cat*) = 0.9 $\ell = -0.7$

sim( , *on the*) = 0.4 sim( , *on the*) = 0.4 $\ell = 0$

Key Idea: Smaller $\ell(c) \leftrightsquigarrow c$ is more visually concrete.

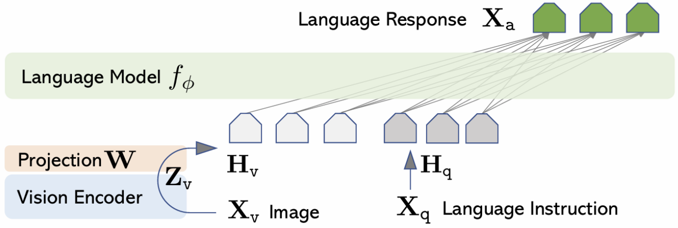
Quantify *visual concreteness* of word spans using loss values.

LLaVA: Visual Instruction Tuning

Use GPT-style language modeling objective.

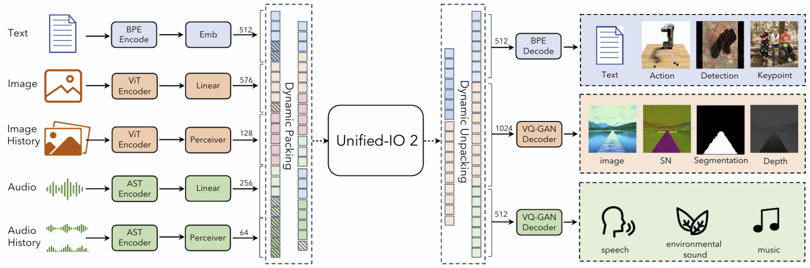
Encode images with different resolutions into "visual tokens."

Project the visual tokens into the textual (joint) space.



[Liu, H. et al. 2023. *Visual instruction tuning*. In *NeurIPS*.]

Towards Encoding Everything in the World

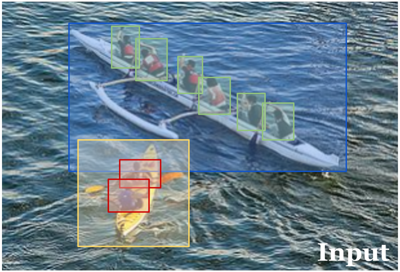


[Lu, J. et al. 2024. Unified-IO 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*]

Finer-Grained Vision-Language Tasks

Object retrieval

Note: the object bounding boxes are given in both training and testing.



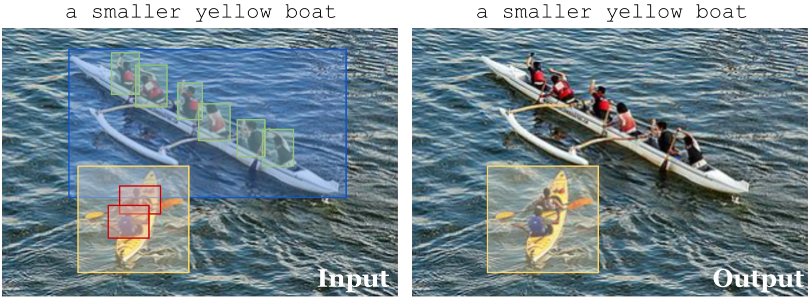
[Baillargeon, R. et al. 1985. *Object permanence in five-month-old infants*. In *Cognition*.]

Finer-Grained Vision-Language Tasks

Object retrieval

Note: the object bounding boxes are given in both training and testing.

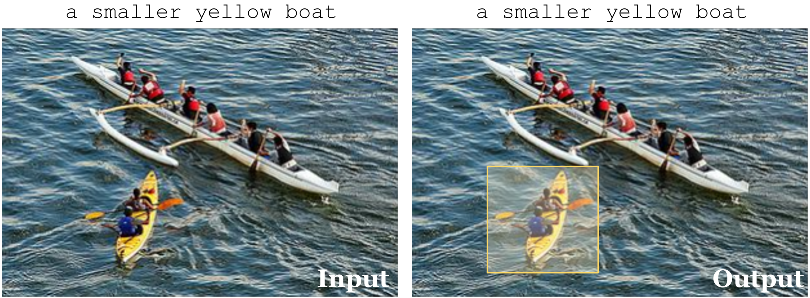
This is a reasonable assumption, as cognitive scientists have shown that 5-month infants recognize objects well.



[Baillargeon, R. et al. 1985. *Object permanence in five-month-old infants*. In *Cognition*.]

Finer-Grained Vision-Language Tasks

Multimodal coreference resolution

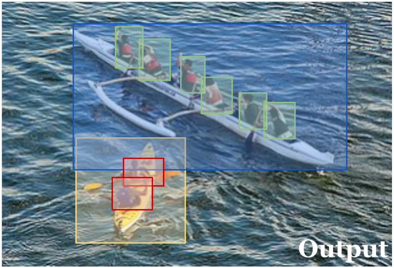


Finer-Grained Vision-Language Tasks

Phrase grounding

Two boats of people kayaking, a smaller yellow boat with two people and a larger white boat with six people.

Two boats of people kayaking, a smaller yellow boat with two people and a larger white boat with six people.



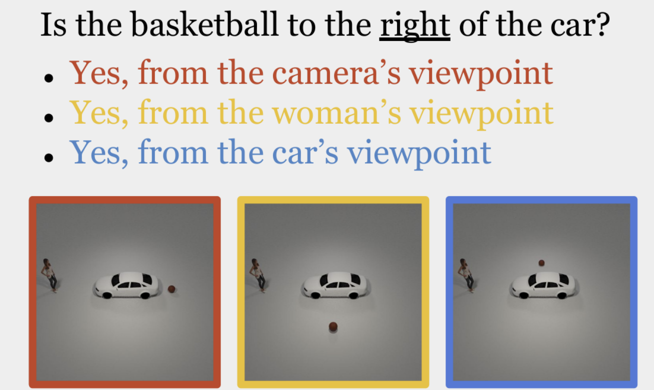
Limitation of Current Vision-Language Models

- Lack of full understanding of the physical world.

[Sarkar, A. et al. 2024. *Shadows don't lie and lines can't bend! Generative models don't know projective geometry...for now*. In *CVPR*.]

Limitation of Current Vision-Language Models

- Poor in recognizing spatial relations, especially poor adapting different spatial frames of reference.



[Zhang Z. et al. 2024. Do vision-language models represent space and how? Evaluating spatial frame of reference under ambiguities.]

Limitation of Current Vision-Language Models

- Highly biased towards cultures with more presence in the training data.

[Bhatia, M. et al. 2024. *From local concepts to universals: Evaluating the multicultural understanding of vision-language models.*]

Next

Pragmatics