Explainable Semantic Space by Grounding Language to Vision with Cross-Modal Contrastive Learning

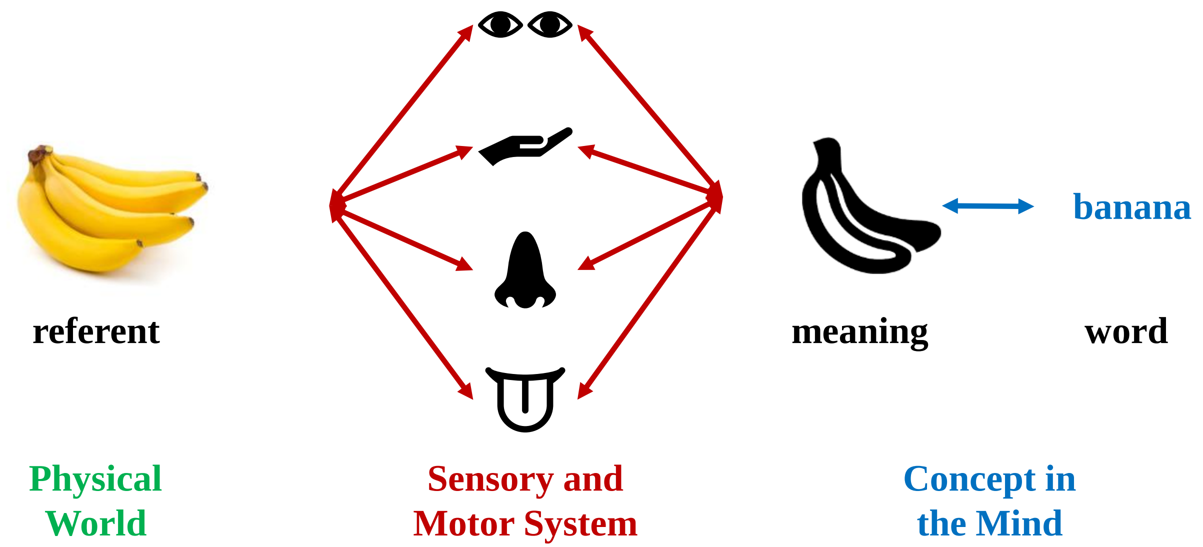
Yizhen Zhang[1,2], Minkyu Choi[1], Kuan Han[1], Zhongming Liu[1,3]

1 Department of Electrical Engineering and Computer Science, University of Michigan
2 Department of Neurological Surgery, University of California San Francisco
3 Department of Biomedical Engineering, University of Michigan

Motivation

How humans learn language?

- We learn the **meaning of a word** (e.g., banana) by associating it with the **sensory features** (e.g., shape, color, smell, taste etc.) of its **referent** (a real banana in the physical world).
- We learn **concepts** from **real world experiences**.



Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic bulletin & review*, 23(4), 979-990.

Motivation

Predominant language learning models learn word representations only from **textual** context instead of **multimodal** context.

Google AI Blog

The latest news from Google AI

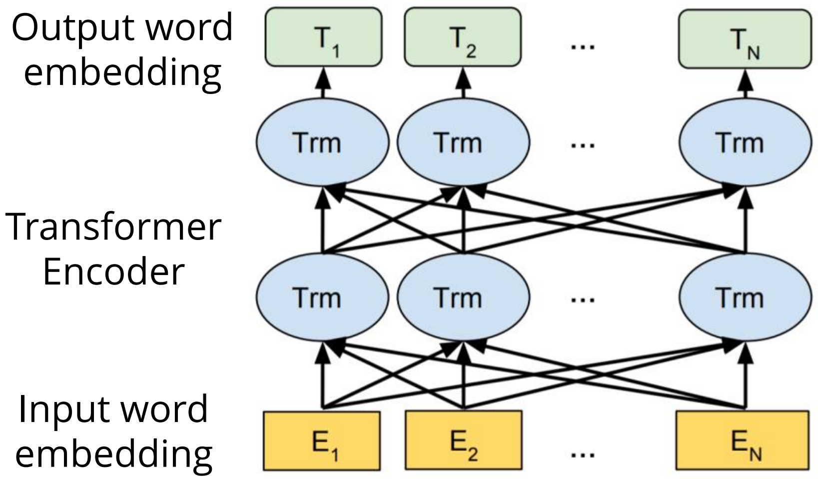
Bert: Pre-training of deep bidirectional transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org
We introduce a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, **BERT** is designed to pre-train deep bidirectional representations …
Cited by 20810 Related articles All 26 versions

Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing

Friday, November 2, 2018

Posted by Jacob Devlin and Ming-Wei Chang, Research Scientists, Google AI Language

Bert



Devlin et al. 2018

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

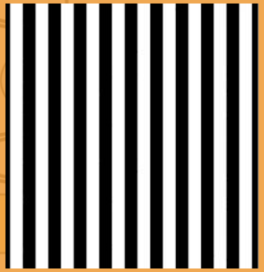
Motivation

Predominant language learning models learn word representations only from **textual** context instead of **multimodal** context.

The machine is doing a task like deciphering an ancient language – since ‘word symbols’ in these language models are not grounded in real world experience.

‘*ZEBRA*’ = HORSE + STRIPES


=

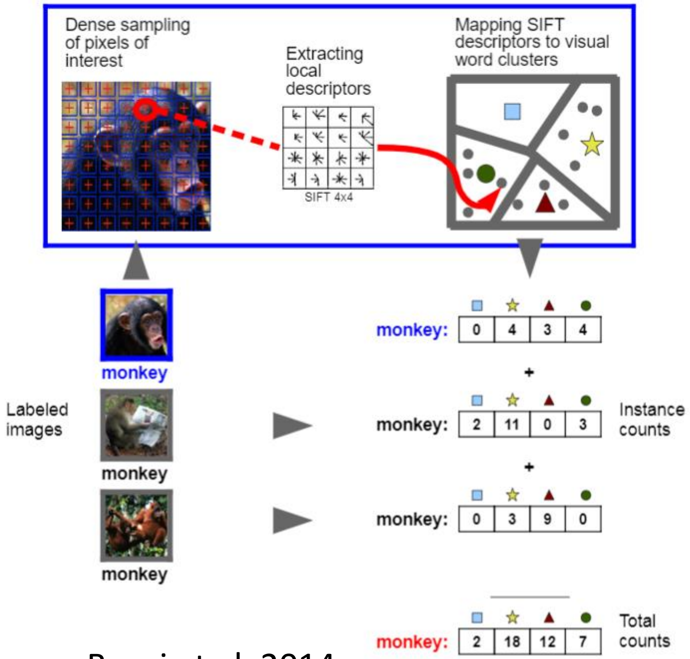
+


Figures adapted from online resources.
Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Background

Early works for grounding language in vision:

bag-of-visual-word



Bruni et al. 2014

capturing word similarity

Model	all	adjs	nouns	verbs	conc-q1	conc-q2	conc-q3	conc-q4	hard
Glove	40.8	**62.2**	42.8	19.6	**43.3**	41.6	42.3	40.2	27.2
Picturebook	37.3	11.7	48.2	17.3	14.4	27.5	46.2	**60.7**	28.8
Glove + Picturebook	**45.5**	46.2	**52.1**	**22.8**	36.7	**41.7**	**50.4**	57.3	**32.5**
Picturebook (Visual)	31.3	11.1	38.8	20.4	13.9	26.1	38.7	47.7	23.9
Picturebook (Semantic)	37.3	11.7	48.2	17.3	14.4	27.5	46.2	60.7	28.8
Picturebook (1)	24.5	2.6	33.5	12.1	4.7	17.8	32.8	47.8	13.6
Picturebook (2)	28.4	6.5	38.9	9.0	5.0	21.3	34.3	55.1	15.7
Picturebook (3)	30.3	11.9	41.9	3.1	2.6	24.3	37.5	58.3	18.4
Picturebook (5)	34.4	6.8	44.5	18.0	9.0	27.9	42.8	58.3	25.9
Picturebook (10)	37.3	11.7	48.2	17.3	14.4	27.5	46.2	60.7	28.8

Table 3: SimLex-999 results (Spearman's ρ). Best results overall are bolded. Best results per section are underlined. Bracketed numbers signify the number of images used. Some rows are copied across sections for ease of reading.

Kiros and Chan et al. 2018

capturing word analogy

Analogy	**Answer Candidates**	**GloVe**	**ViCo**
car:land::aeroplane:?	ocean, sky, road, railway	ocean	**sky**
clock:circle::tv:?	triangle, square, octagon, round	triangle	**square**
park:bench::church:?	door, sofa, cabinet, pew	door	**pew**
sheep:fur::person:?	hair, horn, coat, tail	coat	**hair**
monkey:zoo::cat:?	park, house, church, forest	park	**house**
leg:trouser::wrist:?	watch, shoe, tie, bandana	bandana	**watch**
yellow:banana::red:?	strawberry, lemon, mango, orange	mango	**strawberry**
rice:white::spinach:?	blue, green, red, yellow	blue	**green**
train:railway::car:?	land, desert, ocean, sky	land	**land**
can:metallic::bottle:?	wood, glass, cloth, paper	glass	**glass**
man:king::woman:?	queen, girl, female, adult	**queen**	girl
can:metallic::bottle:?	wood, plastic, cloth, paper	**plastic**	wood
train:railway::car:?	road, desert, ocean, sky	**road**	ocean

Table 6. **Answering Analogy Questions.** Out of 30 analogy pairings tested, we found both GloVe and ViCo to be correct 19 times, only ViCo was correct 8 times, and only Glove was correct 3 times. Correct answers are **highlighted**.

Gupta et al. 2019

better word clusters

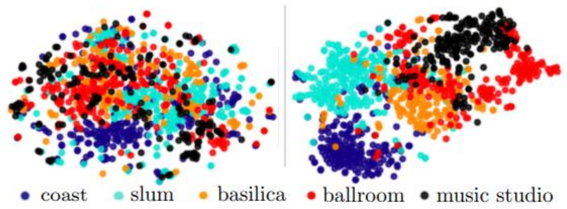
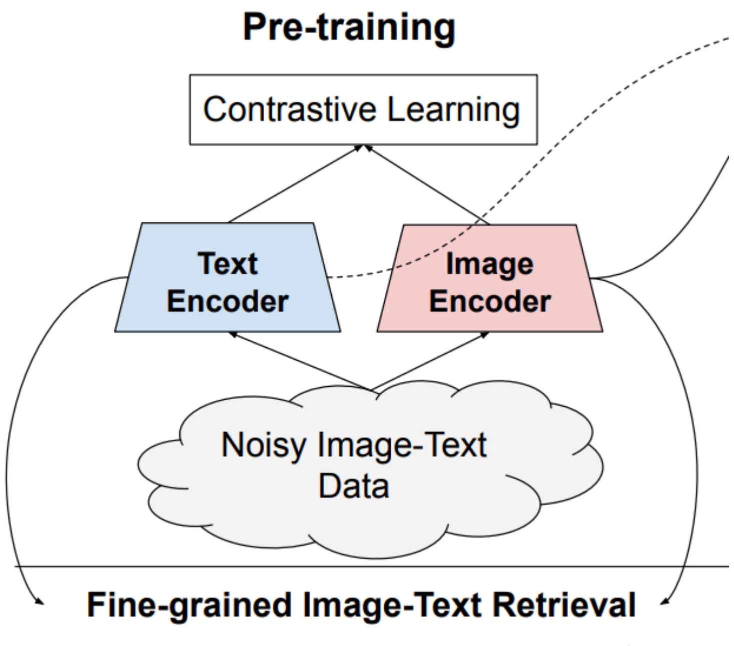


Figure 3: t-SNE visualization on CMPlaces sentences for a set of randomly sampled visual scenes. Left: textual model **T**. Right: grounded model $\mathbf{C}_g + \mathbf{P}_g$.

Bordes and Zablocki et al. 2020

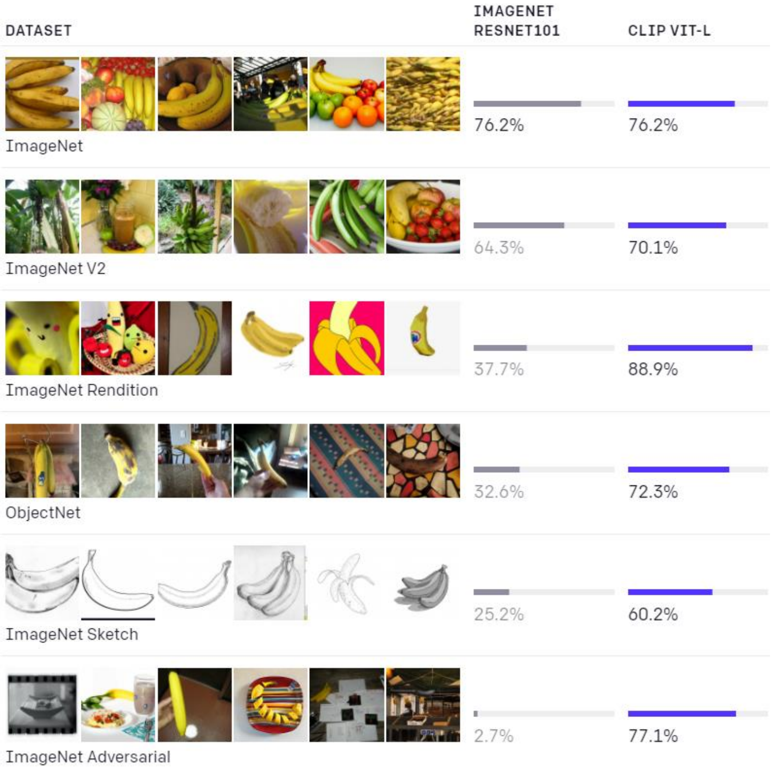
Background

Visual-language cross-modal learning with contrastive loss



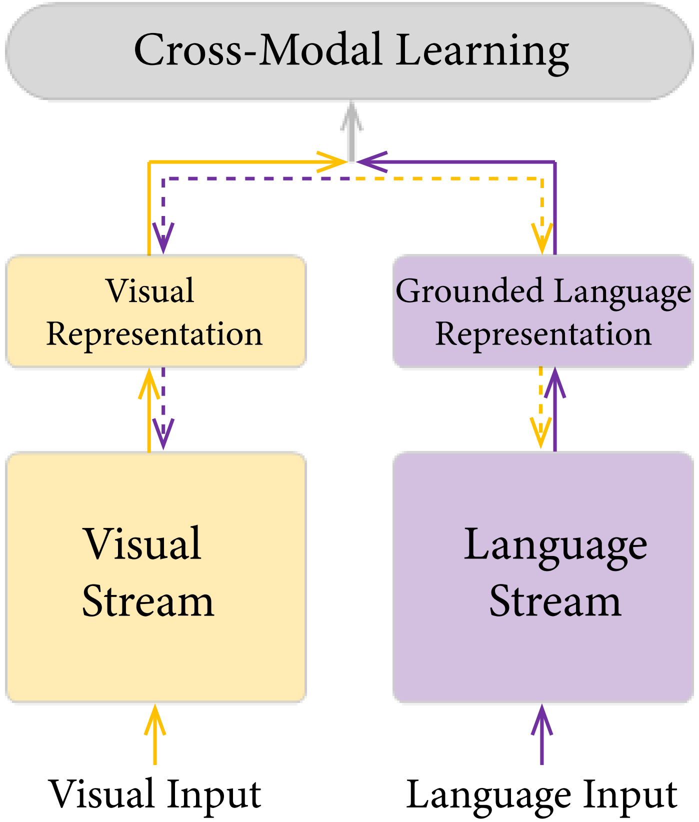
zero-shot image classification

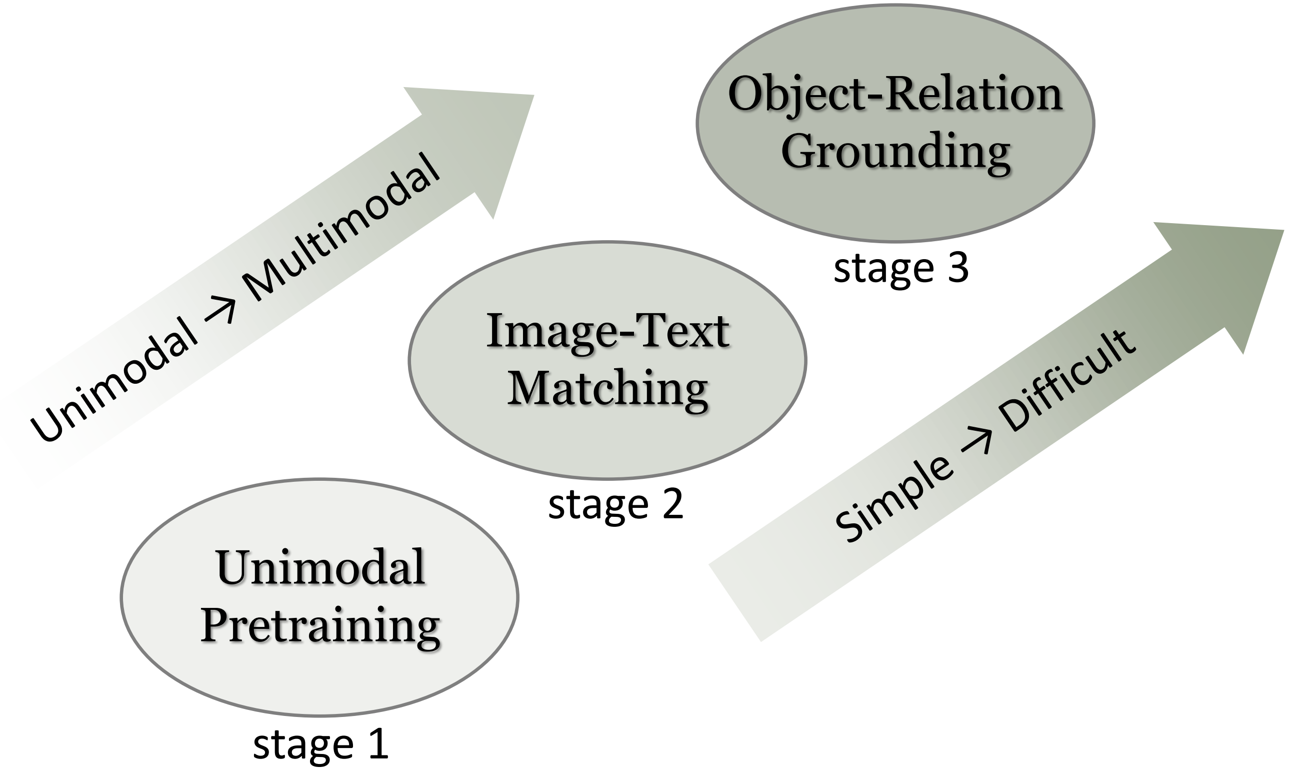
ALIGN: Jia et al. 2021



CLIP: Radford and Kim et al. 2021

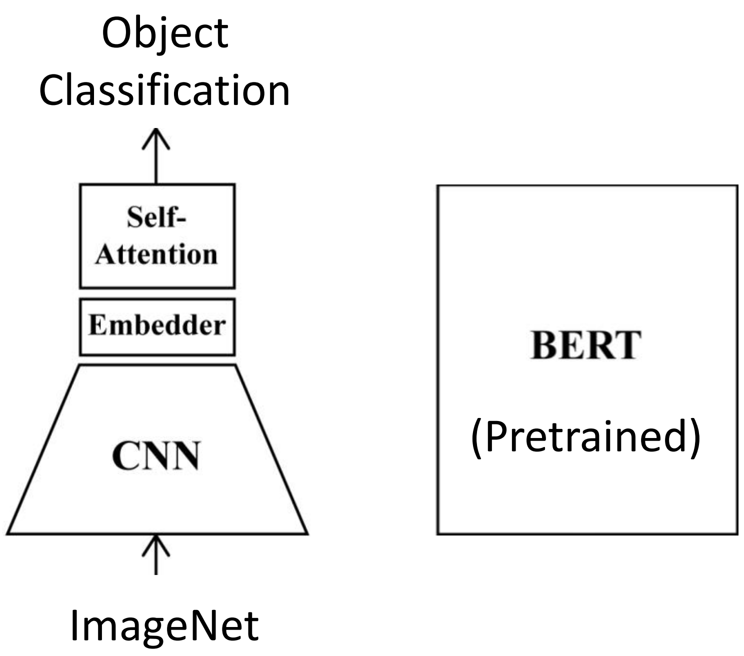
Method


Cross-Modal Learning
Visual Representation
Grounded Language Representation
Visual Stream
Language Stream
Visual Input
Language Input


Unimodal → Multimodal
Object-Relation Grounding
stage 3
Image-Text Matching
stage 2
Simple → Difficult
Unimodal Pretraining
stage 1

Model Design

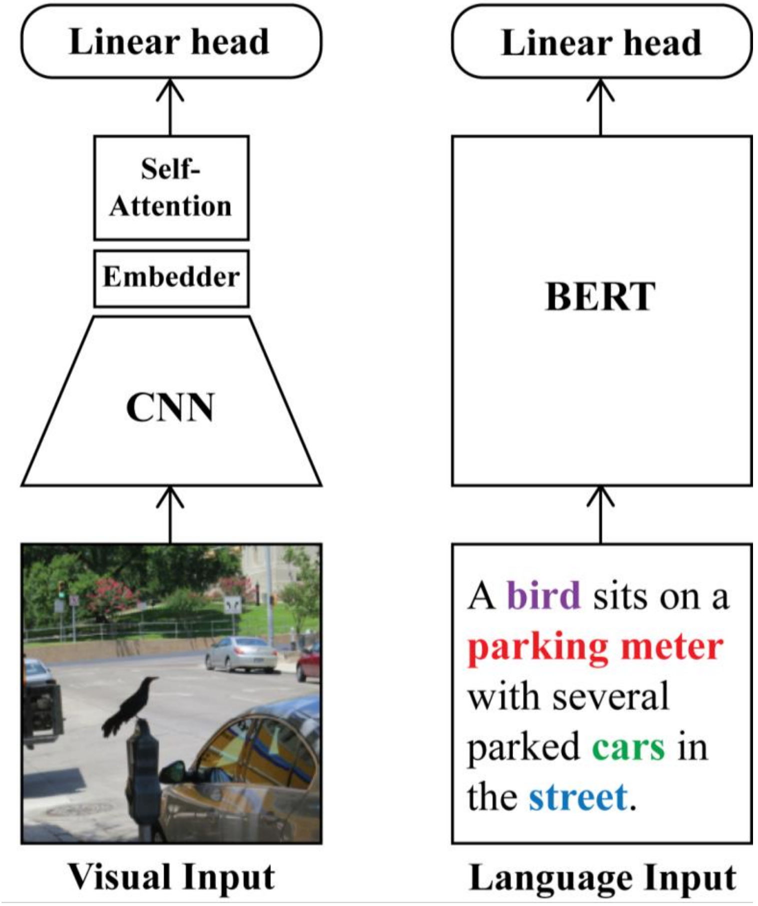
1. Unimodal Pretraining



Model Design

1. Unimodal Pretraining

2. Visual Grounding of Natural Language

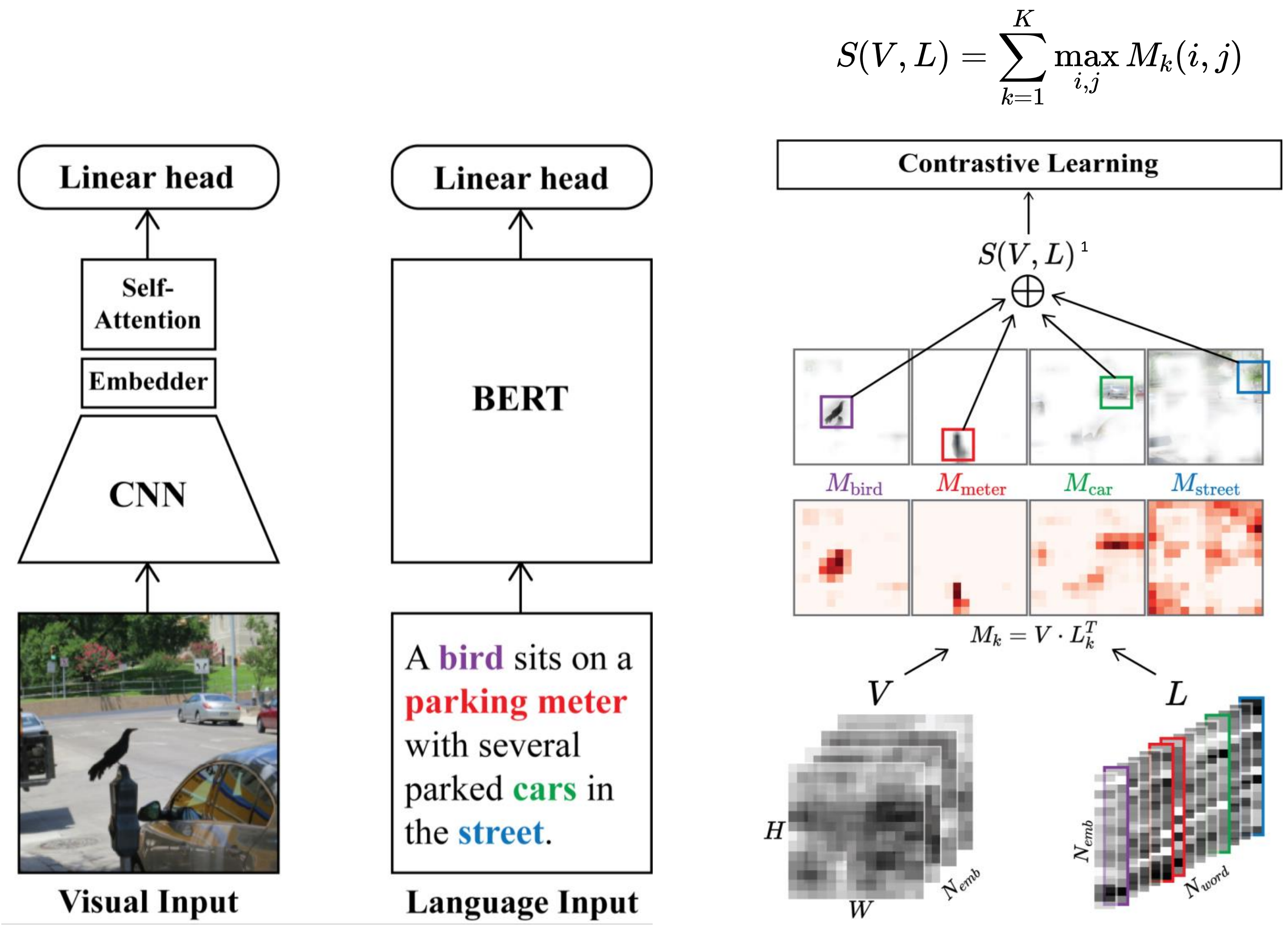


https://cocodataset.org/

Model Design

1. Unimodal Pretraining

2. Visual Grounding of Natural Language



1. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., & Glass, J. (2018). Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 649-665).

Training

NT-Xent[1] loss: Normalized Temperature-scaled Cross Entropy Loss

positive key: correct caption

query: Image

$$\mathrm{Loss}_l = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B}\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_j)/\tau)}$$

$$\mathrm{Loss}_v = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B}\exp(S(\boldsymbol{V}_j, \boldsymbol{L}_i)/\tau)}$$

batch size

temperature hyperparameter

query: Image

negative keys: wrong captions

*Contrasting **key** representations from the **language** stream*

1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

Training

NT-Xent[1] loss: Normalized Temperature-scaled Cross Entropy Loss

query: Image

positive key: correct caption

$$\text{Loss}_l = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B}\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_j)/\tau)}$$

batch size

query: Image

negative keys: wrong captions

temperature hyperparameter

$$\text{Loss}_v = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B}\exp(S(\boldsymbol{V}_j, \boldsymbol{L}_i)/\tau)}$$

*Contrasting **key** representations from the **language** stream*

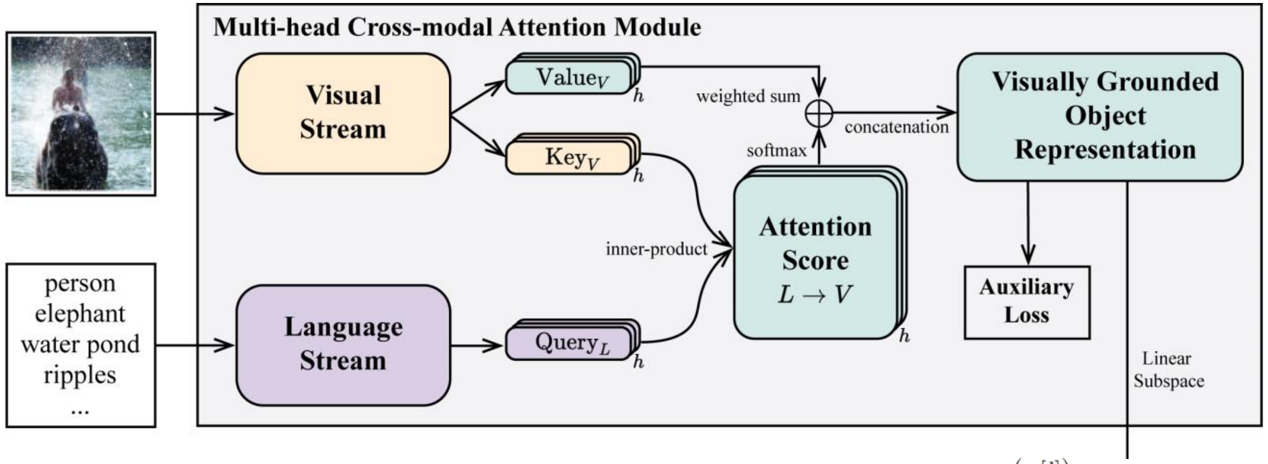
1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

Model Design

1. Unimodal Pretraining

2. Visual Grounding of Natural Language

3. Visual Grounding of object relations



$$\text{auxiliary loss}(\boldsymbol{x}, l) = -\log \frac{\exp\left(\boldsymbol{x}[l]\right)}{\sum_{i=1}^{N} \exp\left(\boldsymbol{x}[i]\right)}$$

VISUALGENOME

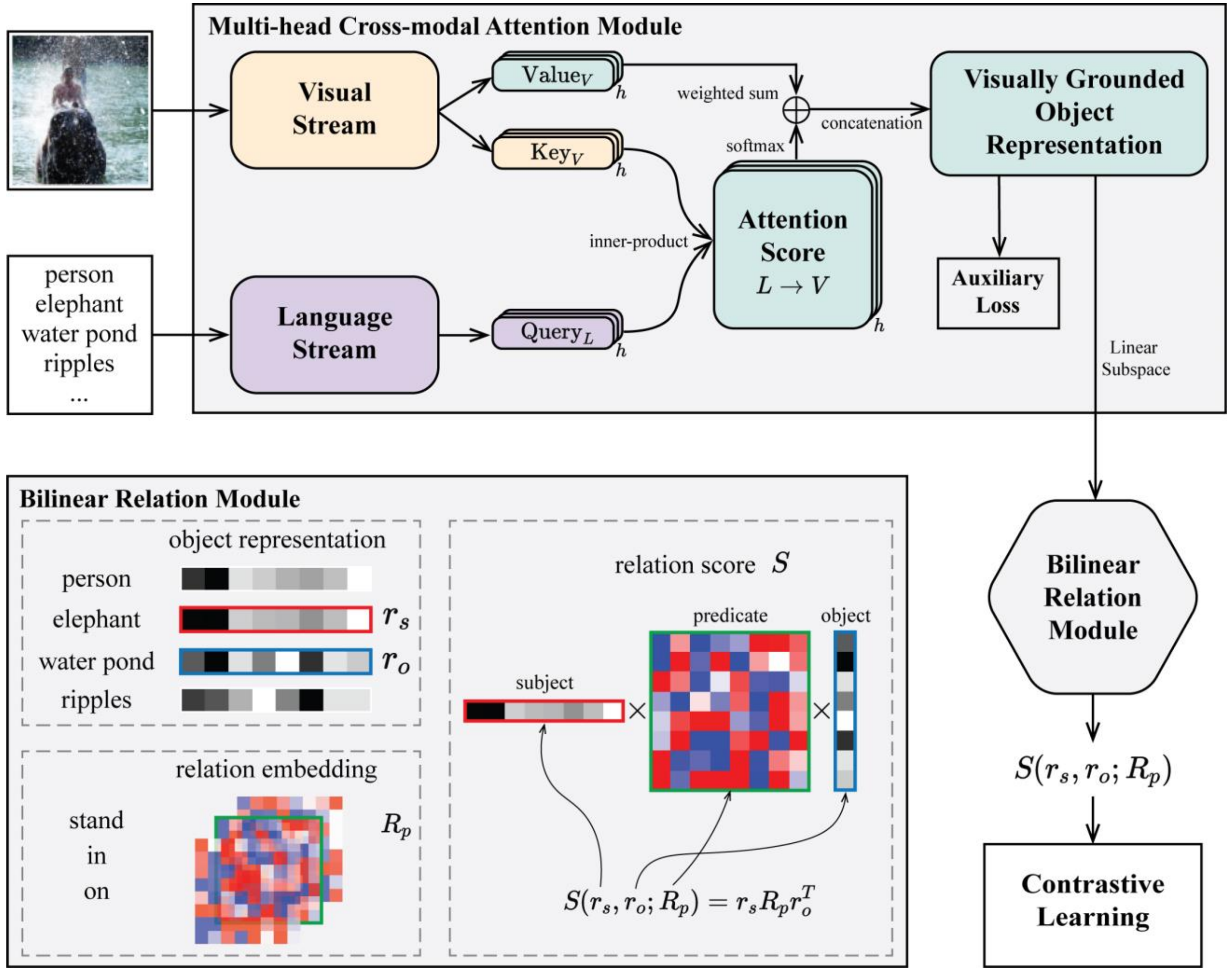
https://visualgenome.org/

Model Design

1. Unimodal Pretraining

2. Visual Grounding of Natural Language

3. Visual Grounding of object relations

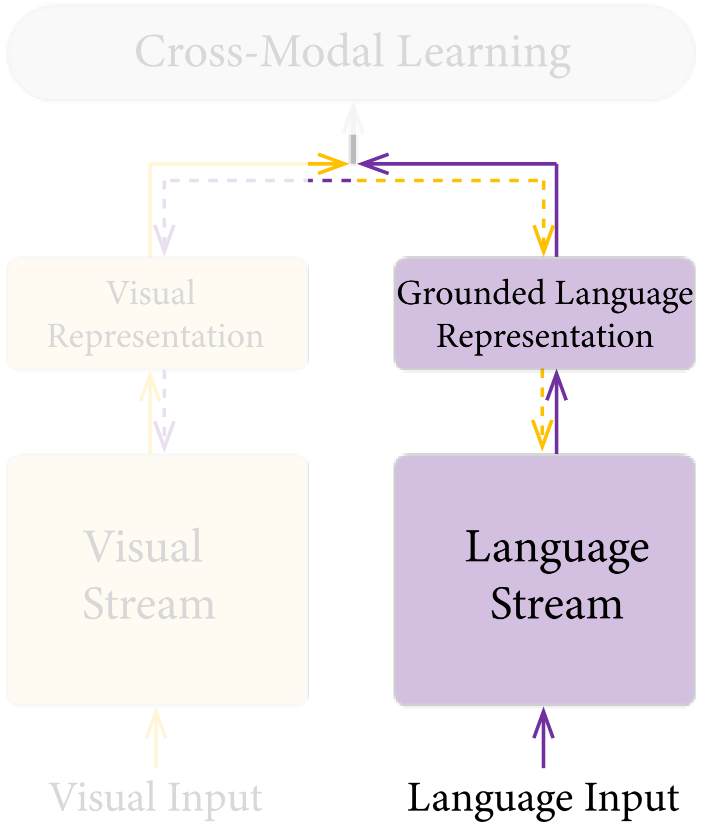


$$\text{Loss}_{\text{rel}} = -\frac{1}{|\mathcal{B}|} \sum_{(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p) \in \mathcal{B}} \log \frac{\exp(S(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p)/\tau)}{\sum_{k \in \mathcal{K}_{\text{rel}}} \exp\big(S(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p^k)/\tau\big)}$$

$$\text{Loss}_{\text{obj}} = -\frac{1}{|\mathcal{B}|} \sum_{(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p) \in \mathcal{B}} \log \frac{\exp(S(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p)/\tau)}{\sum_{k \in \mathcal{K}_{\text{obj}}} \exp(S(\boldsymbol{r}_s^k, \boldsymbol{r}_o^k; \boldsymbol{R}_p)/\tau)}$$

Interpreting the grounded semantic space

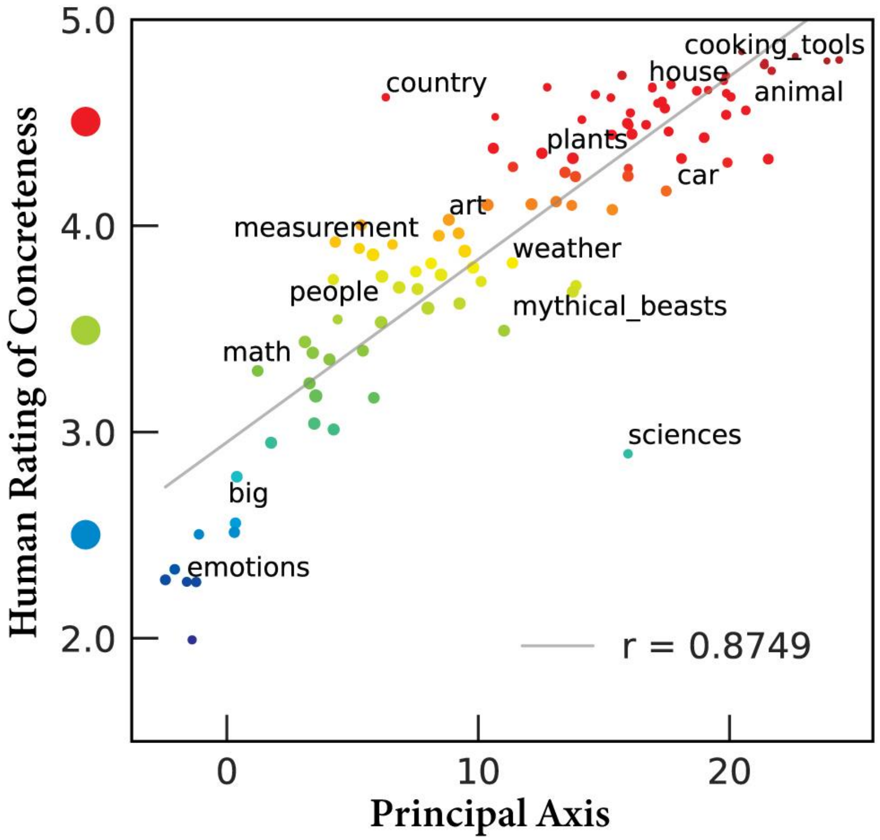
How does visual grounding reshape the semantic space in the language stream?



After visual grounding, we

- detach the language stream
- extract the grounded word embeddings
- apply intrinsic evaluations
    - Principal Component Analysis
    - Concreteness gradient
    - Clustering of Word Categories
    - Concept composition
    - Cross-modal image search
- interpret the semantic space by human intuition and neurobiological knowledge

Results - PCA



Examples:

word	category	principal axis	human rating
oven	cooking tool	30.86	4.97
zebra	animal	24.74	4.86
car	car	22.06	4.89
furniture	house	20.83	4.89
defense	country	4.65	4.19
chemistry	sciences	18.67	3.64
wood	plants	16.22	4.85
cartoon	art	11.84	4.33
humid	weather	11.58	3.48
angel	mythical beasts	7.22	3.82
lover	people	6.01	3.68
thousand	math	3.65	3.07
huge	big	0.17	3.54
cheerful	emotions	- 2.54	2.34

Human-rated Word Concreteness: Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911.

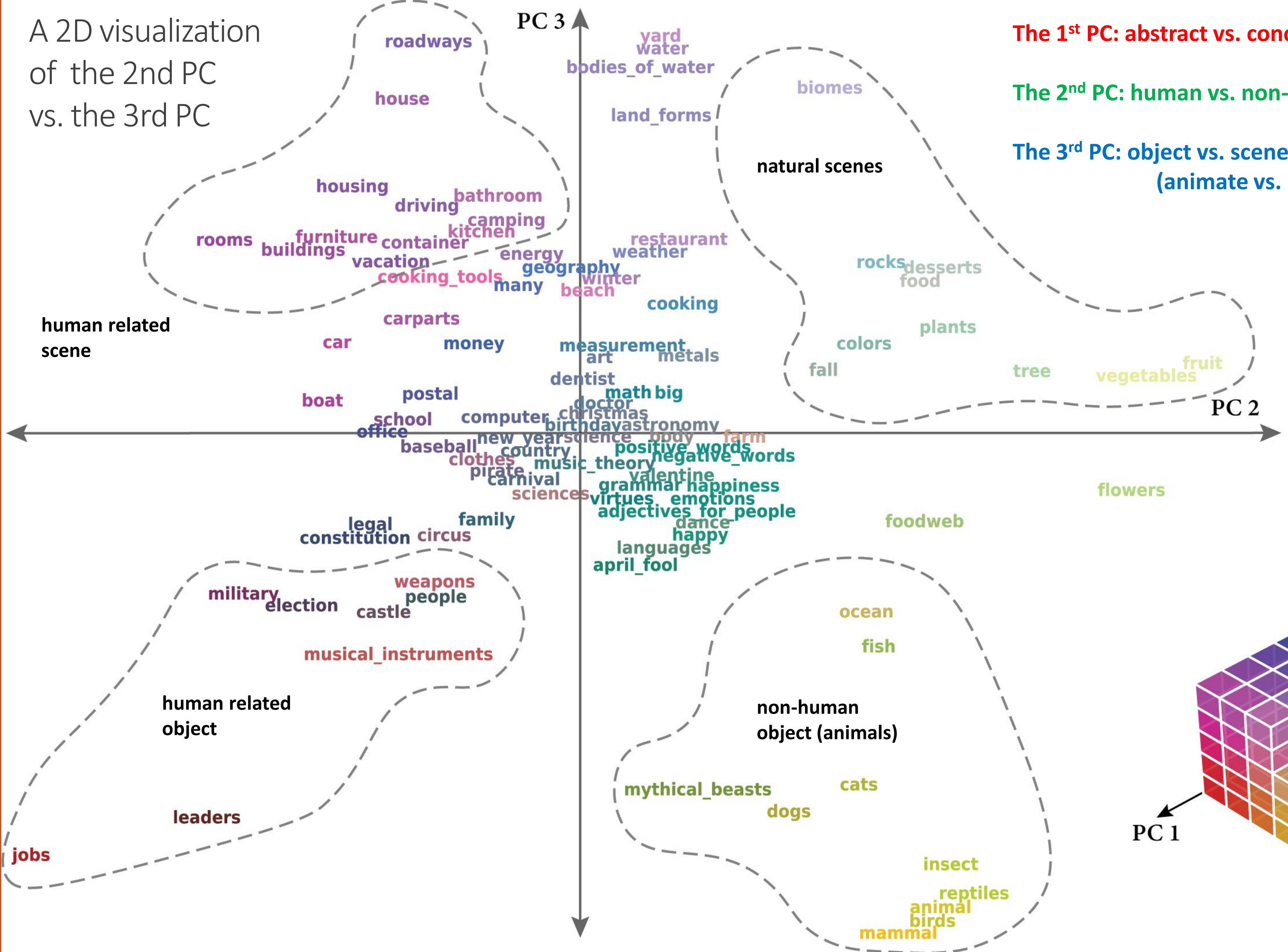
Semcat dataset: Şenel, L. K., Utlu, I., Yücesoy, V., Koc, A., & Cukur, T. (2018). Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1769-1779.

Results - PCA

	Correlation (Pearson's r)		
Group	Bert	Grounded	Relational Grounded
word-level	0.1040	0.6615	**0.6948**
category-level	0.3538	**0.8749**	0.8001

Semcat dataset: Şenel, L. K., Utlu, I., Yücesoy, V., Koc, A., & Cukur, T. (2018). Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1769-1779.

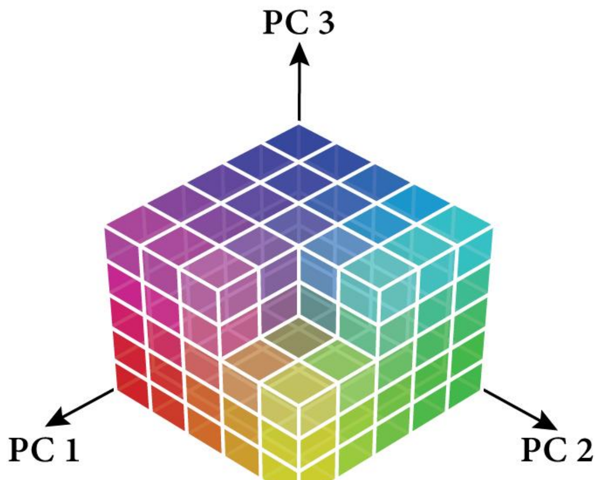
A 2D visualization
of the 2nd PC
vs. the 3rd PC



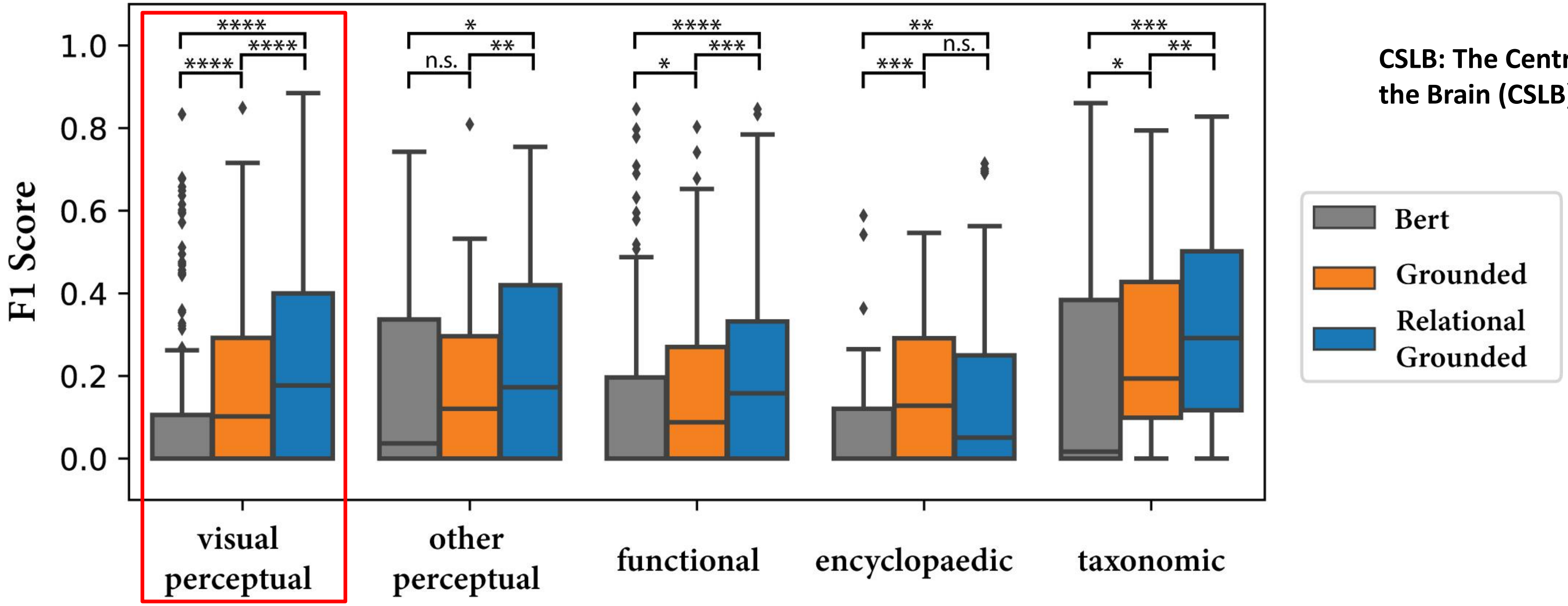
The 1st PC: abstract vs. concrete

The 2nd PC: human vs. non-human

The 3rd PC: object vs. scene
(animate vs. inanimate)



Results – Semantic Norm Prediction



e.g., has_wheels has_flavors does_cut is_dangerous is_clothing

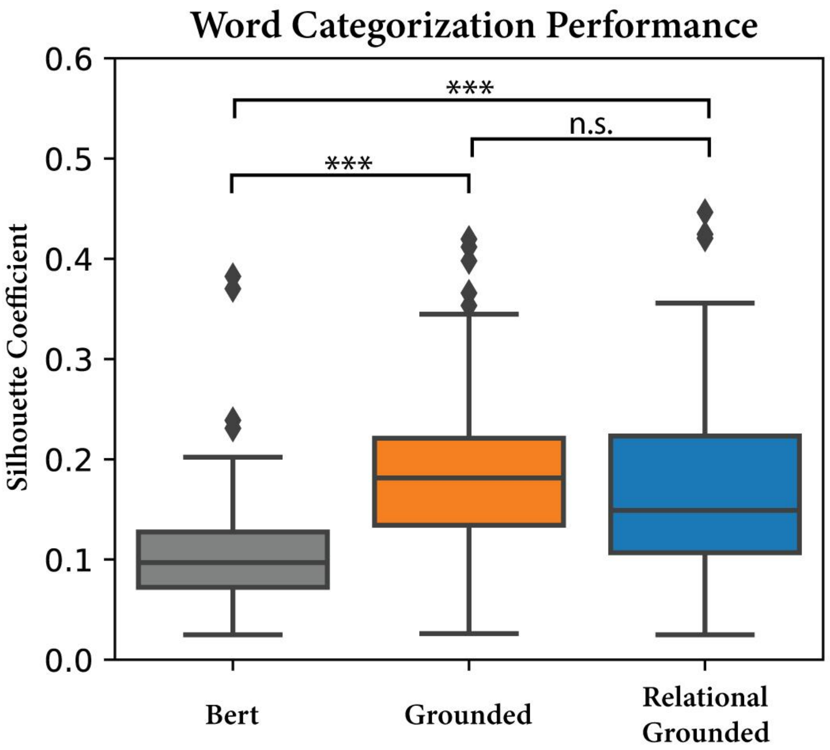
CSLB: The Centre for Speech, Language and the Brain (CSLB) concept property norms.

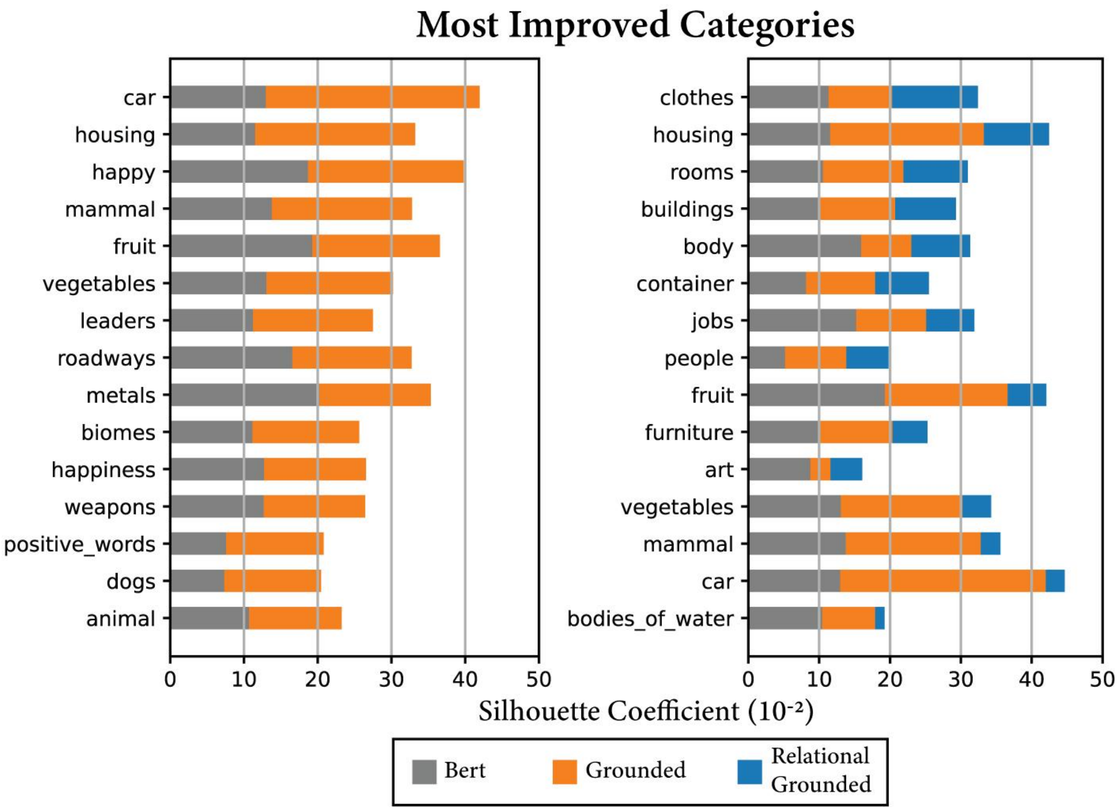
Wilcoxon Signed Rank Test
Significant level:

n.s.	not significant
*	p < 0:05
**	p < 0:01
***	p < 0:001
****	p < 0:0001

CSLB dataset: Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior research methods*, *46*(4), 1119-1127.

Results – word clustering



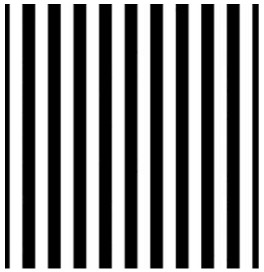


Semcat dataset: Şenel, L. K., Utlu, I., Yücesoy, V., Koc, A., & Cukur, T. (2018). Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1769-1779.

Results - Vision based compositional reasoning

Where is the phrase “striped horse” represented in the Semantic Space?

Is it a ZEBRA ?


+

=


What is a “Striped horse” in the grounded semantic space?


tomcat


seahorse


squirrel


antelope


swordfish

Representation Similarity of Visually Inferred Concepts
Striped Horse : Zebra

Similarity Ranking

Cosine	Word	Rank
0.61	tomcat	1
0.59	seahorse	2
0.58	squirrel	3
0.57	antelope	4
0.57	bobcat	5
0.56	wildcat	6
0.55	swordfish	7
0.55	orangutan	8
0.54	rhinoceros	9
0.53	lion	10
0.53	monkey	11
0.53	giraffe	12
0.49	leopard	30
0.47	donkey	38
0.44	mule	60
0.41	horse	84
0.40	camel	95
0.36	mare	141
0.34	unicorn	189
0.32	stallion	245
0.25	chestnut	682
0.12	zebra	2914

Before Visual Grounding

Rank	Word	Cosine
1	horse	0.80
2	mule	0.74
3	mare	0.73
4	stallion	0.72
5	seahorse	0.70
6	donkey	0.70
7	camel	0.69
8	antelope	0.66
9	unicorn	0.63
10	leopard	0.63
11	chestnut	0.61
12	zebra	0.60
20	rhinoceros	0.57
28	wildcat	0.54
40	lion	0.51
41	orangutan	0.51
43	bobcat	0.51
56	tomcat	0.47
60	squirrel	0.47
62	monkey	0.46
179	giraffe	0.36
577	swordfish	0.25

After Visual Grounding

horse

mule

mare

stallion

zebra

Images for illustrative purpose only. Adapted from online resources.

What is a “Striped horse” in the grounded semantic space?

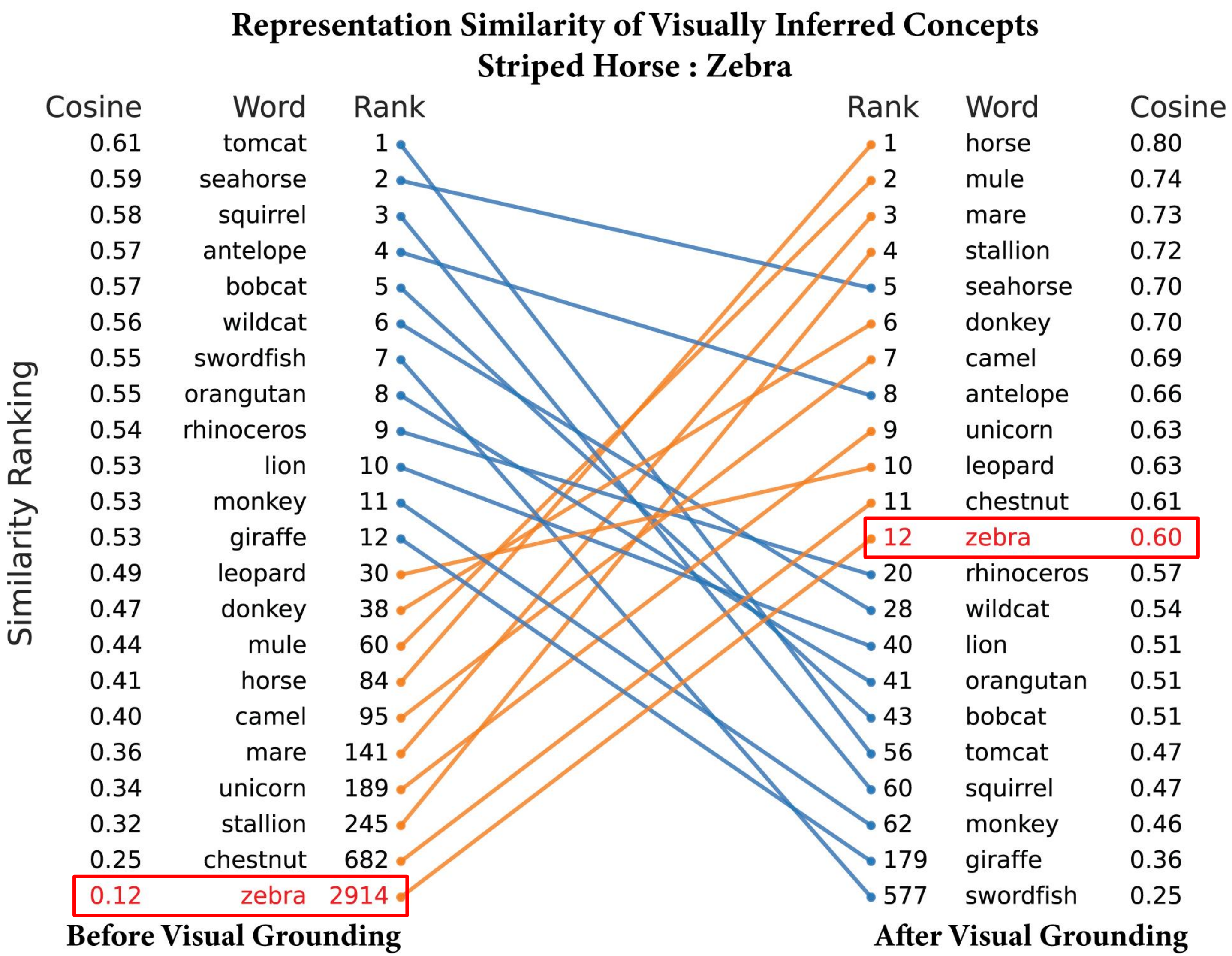

tomcat


seahorse


squirrel


antelope


swordfish




horse


mule


mare


stallion


zebra

Images for illustrative purpose only. Adapted from online resources.

What is a “Red Fruit” in the grounded semantic space?

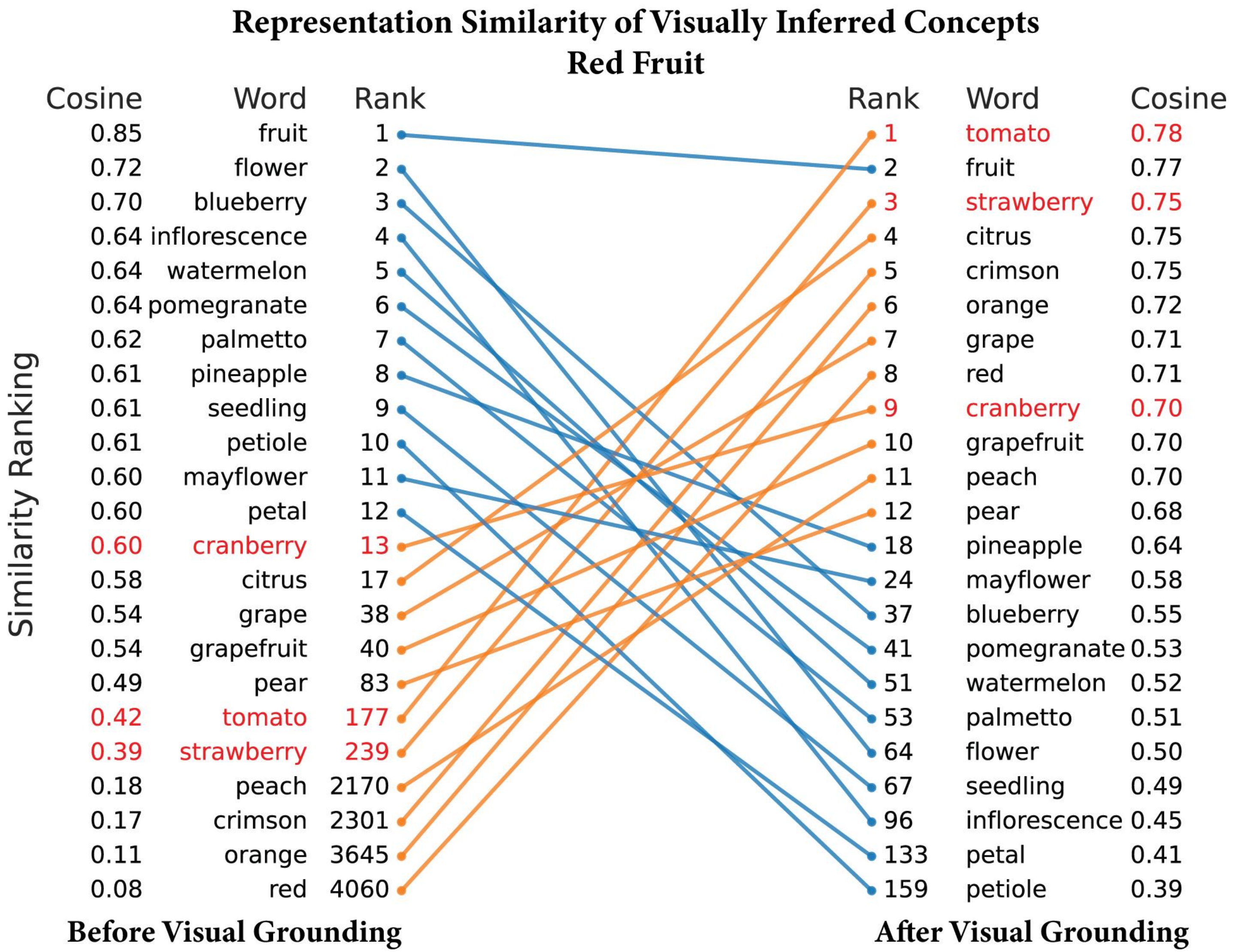

blueberry


watermelon


pomegranate


palmetto


pineapple




tomato


strawberry


citrus


grape

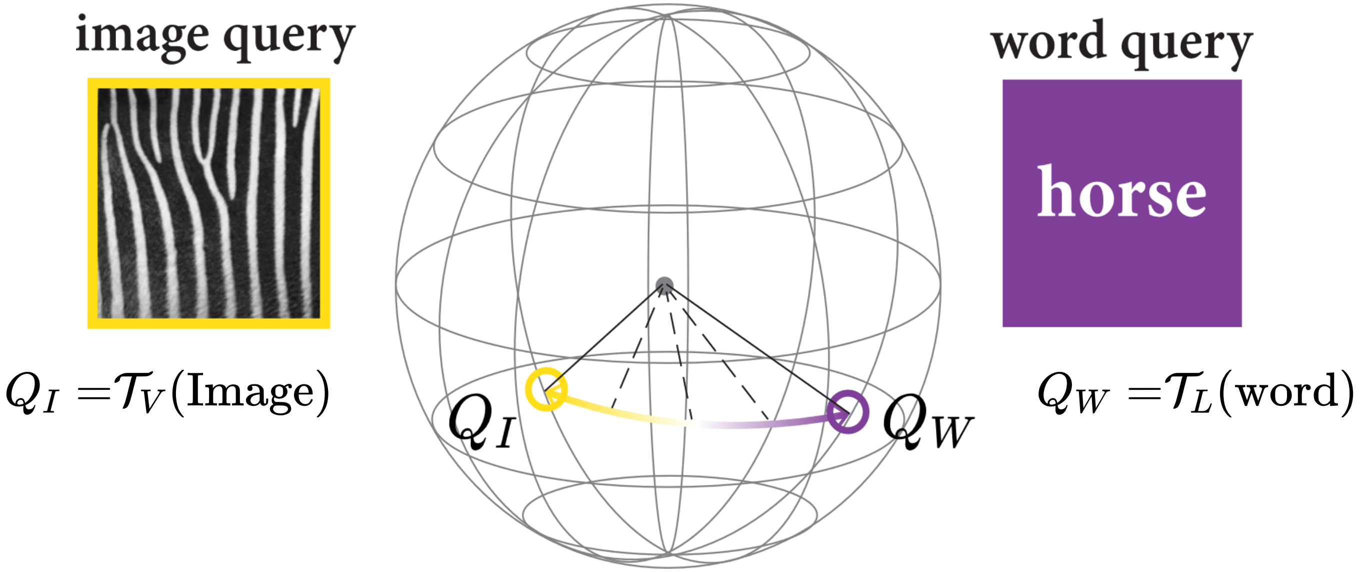

cranberry


peach

Results - Vision based compositional reasoning

Query Phrase	Target Word	Similarity (cosine | rank)					
		Bert		Grounded		Relational	
striped horse	zebra	0.12	2914	0.60	12	**0.63**	**8**
black and white bear	panda	0.13	2478	0.69	**2**	**0.81**	**2**
flying car	plane	0.36	167	**0.66**	**4**	0.61	11
round container	bowl	0.25	489	0.56	8	**0.67**	**2**
red fruit	strawberry	0.39	239	0.75	**3**	**0.85**	**3**
young dog	puppy	0.40	94	0.92	**2**	**0.93**	**2**
iced mountain	glacier	0.44	20	**0.86**	**1**	0.73	5
clear sky	sunny	0.27	631	0.31	184	**0.34**	**61**
hot weather	summer	0.27	903	0.52	14	**0.53**	**6**

Results - Continuous Image Search from Text & Image



$\mathcal{T}_V, \mathcal{T}_L$:

The nonlinear transformation from the image / word space to the L2-normed multimodal representational space.

Search query: $$Q_{\text{search}} = (1 - \alpha) \cdot Q_I + \alpha \cdot Q_W$$

Searching: $$\underset{K \in \text{Image Pool}}{\text{argmax}} \cos(Q_{\text{search}}, \mathcal{T}(K))$$

Results - Continuous Image Search from Text & Image

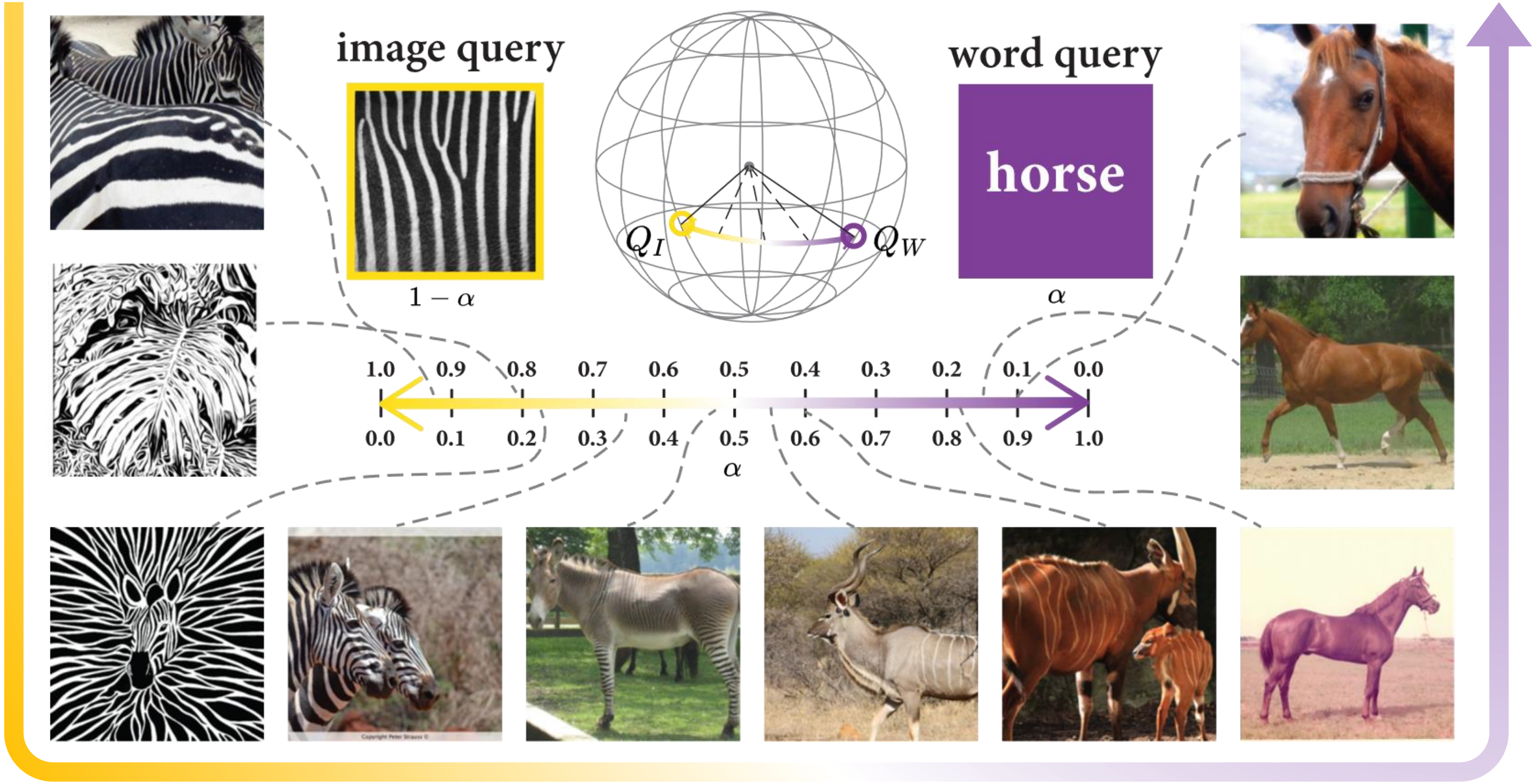


Image Pool: 41620 images from Open Images validation dataset

Results - Continuous Image Search from Text & Image

Image Query **Word Query**

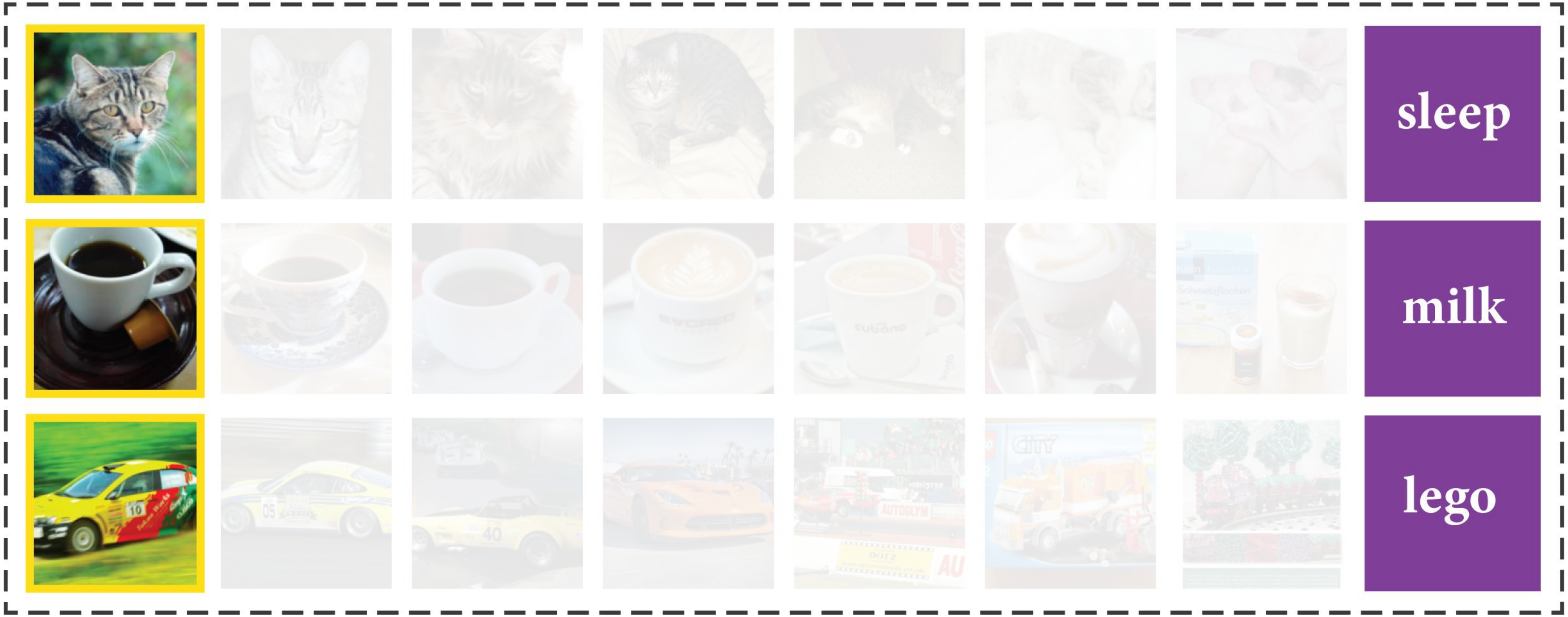


Image Pool: 41620 images from Open Images validation dataset

Results - Continuous Image Search from Text & Image

Image Query **Word Query**

Image Pool: 41620 images from Open Images validation dataset

Summary

We design a two-stream model for grounding language learning in vision:

- Progressive training
- cross-modal contrastive learning

After training, we analyze the language model as a stand-alone system. In this grounded word embedding space:

- The first principal axis = concrete vs. abstract gradient
- Principal axes are explainable by human intuition.
- Word representation captures human-defined semantic feature norms.
- Concepts are better clustered.

Besides, “zebra = striped horse” in both word embedding space and joint representational space.

Thank you!

QUESTIONS & COMMENTS?