Multimodal Representation Learning: Vision-Language

Seung Yoon Lee

Dec 2, 2021
CPSC 677

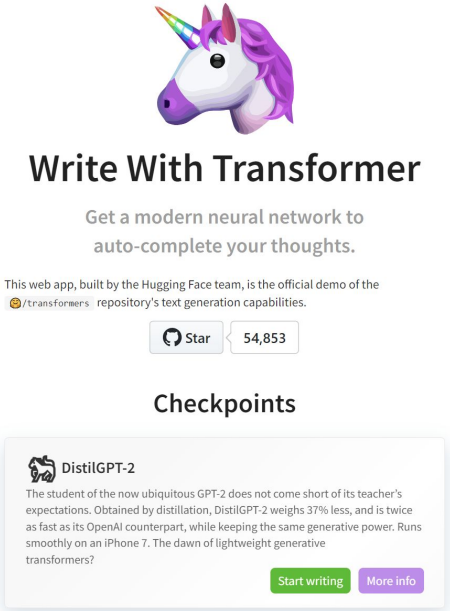
Presentation Outline

- **What is Multimodal Representation Learning?: Introduction and Background**
  - Motivation: Value of Multimodality
  - Background: Vision + Language (V+L) Representation Learning
    i. Data
    ii. V+L Feature Representation
    iii. Pretraining Architecture

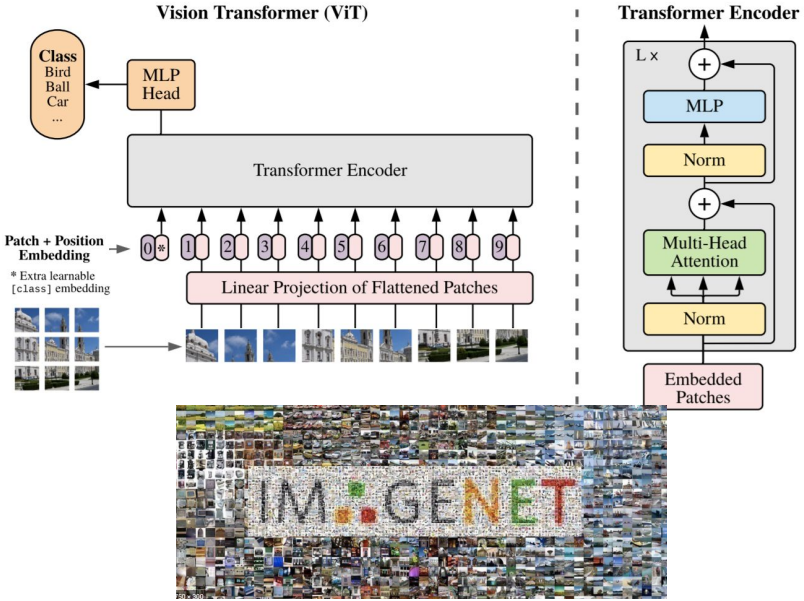
- **Three Papers: Learning Visual Concepts using Natural Language Supervision**
  - Paper 1: "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks." (Li, Xiujun, et al. 2020)
  - Paper 2: "ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration" (Cui, Yuhao et al., 2021)
  - Paper 3: "Learning Transferable Visual Models from Natural Language Supervision." (Radford, Alec, et al., 2021)

- **Future Directions: Bias and Interpretability**

Previously, Vision and NLP Focused on Unimodal Learning



https://transformer.huggingface.co/



Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
[Img source] https://cv.gluon.ai/build/examples_datasets/imagenet.html

Can “Meaning” or “Concepts” be Learned with Unimodal Data?

- “Any system trained only on linguistic form cannot in principle learn meaning.” Bender and Koller (2020).

- Visual concepts are inferred *indirectly*.
  - e.g., Resnet-50 (He et. al., 2016), Vision Transformer (Dosovitskiy et. al., 2020)

- **How can we do better?**

Multimodal Neurons in Human Brain

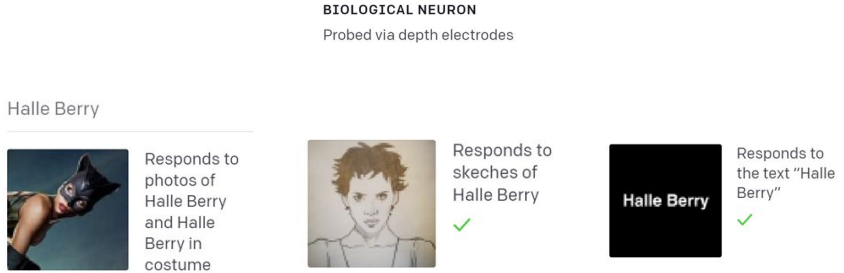
nature Vol 435|23 June 2005|doi:10.1038/nature03687

LETTERS

Invariant visual representation by single neurons in the human brain

R. Quian Quiroga[1,2]†, L. Reddy[1], G. Kreiman[3], C. Koch[1] & I. Fried[2,4]

It takes a fraction of a second to recognize a person or an object even when seen under strikingly different conditions. How such a robust, high-level representation is achieved by neurons in the human brain is still unclear[1–6]. In monkeys, neurons in the upper stages of the ventral visual pathway respond to complex images such as faces and objects and show some degree of invariance to metric properties such as the stimulus size, position and viewing angle[2,4,7–12]. We have previously shown that neurons in the human medial temporal lobe (MTL) fire selectively to images of faces, animals, objects or scenes[13,14]. Here we report on a remarkable



Quiroga, R. Quian, et al. "Invariant visual representation by single neurons in the human brain." *Nature* 435.7045 (2005): 1102-1107.

https://openai.com/blog/multimodal-neurons/

What is Multimodal Representation Learning?

Modality: “particular way or mechanism of encoding information” representing different aspects of an object/concept. (Guo et al., 2019)

- e.g., Image, text, audio

Multimodal Representation Learning:

- Learning a joint semantic space representation using features from different modalities.

[Definition from:] Guo, Wenzhong, Jianwen Wang, and Shiping Wang. "Deep multimodal representation learning: A survey." *IEEE Access* 7 (2019): 63373-63394.

Value of Multimodality:
Using Natural Language Supervision to Learn Visual Concepts

Image

Image source: https://www.kaggle.com/adityajn105/flickr8k
1002674143_1b742ab4b8.jpg,A little girl is sitting in front of a large painted rainbow .

Value of Multimodality:
Using Natural Language Supervision to Learn Visual Concepts

Image

Text

A little **girl** is *sitting in front of* a large *painted* rainbow.

- Visual concepts to be ***directly*** specified and communicated to the model.
  - e.g., focal object, relationship, contextual/conceptual understanding

[Image Source]: https://www.kaggle.com/adityajn105/flickr8k
1002674143_1b742ab4b8.jpg,A little girl is sitting in front of a large painted rainbow .

OpenAI DALL-E: Generating Image from Text

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT a store front that has the word ‘openai’ written on it. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

[Image Source]: https://openai.com/blog/dall-e/

Example V+L Tasks using Multimodal Learning

Image Captioning

A man riding a bicycle with a sidecar with a small dog sitting in it .

Visual Question Answering

Visual Commonsense Reasoning

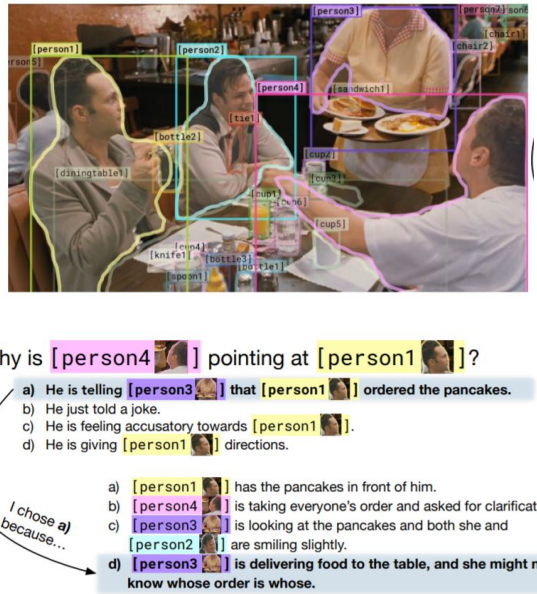


Image source:
1. COCO caption dataset: https://cocodataset.org/#explore?id=523357; https://farm9.staticflickr.com/8283/7606151322_70dfe5bb12_z.jpg
2. Goyal, Yash, et al. "Making the v in vqa matter: Elevating the role of image understanding in visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
3. Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Presentation Outline

- **What is Multimodal Representation Learning?: Introduction and Background**
  - Motivation: Value of Multimodality
  - **Background: Vision + Language (V+L) Representation Learning**
    i. **Data**
    ii. **V+L Feature Representation**
    iii. **Pretraining Architecture**

- **Three Papers: Learning Visual Concepts using Natural Language Supervision**
  - Paper 1: "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks." (Li, Xiujun, et al. 2020)
  - Paper 2: "ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration" (Cui, Yuhao et al., 2021)
  - Paper 3: "Learning Transferable Visual Models from Natural Language Supervision." (Radford, Alec, et al., 2021)

- **Future Directions**

1. Data

Dataset for Training V+L Models: Image-Text Pairs.

( , a black and white dog sitting next to a frisbee.)

Slide format borrowed from: Self-supervised Learning presented by Licheng Yu , Linjie Li and Yen-Chun Chen.
https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part5-pretraining.pdf

Public Datasets used for V+L Pretraining

	In-domain		Out-of-domain	
Split	COCO Captions	VG Dense Captions	Conceptual Captions	SBU Captions
train	533K (106K)	5.06M (101K)	3.0M (3.0M)	990K (990K)
val	25K (5K)	106K (2.1K)	14K (14K)	10K (10K)

Table 1: Statistics on datasets used for pre-training. Each cell shows #image-text pairs (#images).

MS COCO Captions

a black and white dog wearing a red shirt on top of a bed.
a dog in a red sweater on a bed
a dog lays in bed with a sweater on.
a dog in a red sweater, lying on a bed.
a black and white dog wearing a red coat on a bed.

Google Conceptual Captions

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

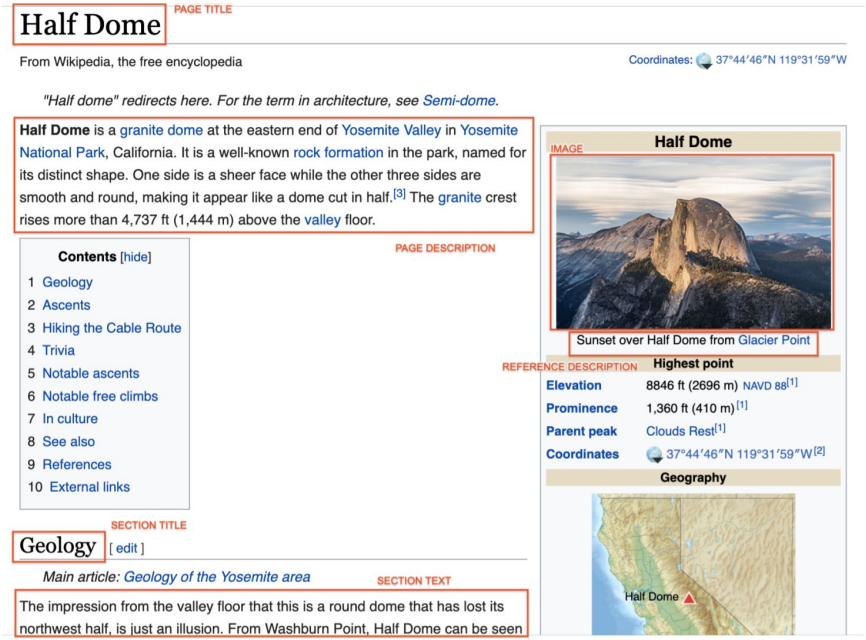
Figure 1: Examples of images and image descriptions from the Conceptual Captions dataset; we start from existing alt-text descriptions, and automatically process them into Conceptual Captions with a balance of cleanliness, informativeness, fluency, and learnability.

[Table source]: Chen, Yen-Chun, et al. "Uniter: Universal Image-text Representation Learning." *European Conference on Computer Vision*. Springer, Cham, 2020.
[Image source-COCO Captions]: https://cocodataset.org/#explore?id=329474
[Image source-Conceptual Captions]: Sharma, Piyush, et al. "Conceptual captions: A Cleaned, Hypernymed, Image Alt-Text Dataset for Automatic Image Captioning." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.
Slide format borrowed from: Self-supervised Learning presented by Licheng Yu , Linjie Li and Yen-Chun Chen.https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part5-pretraining.pdf

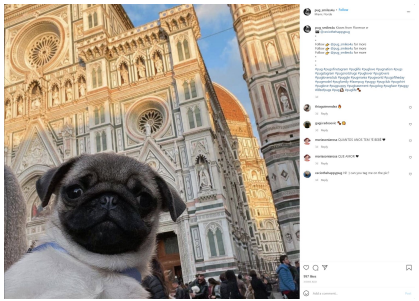
Question

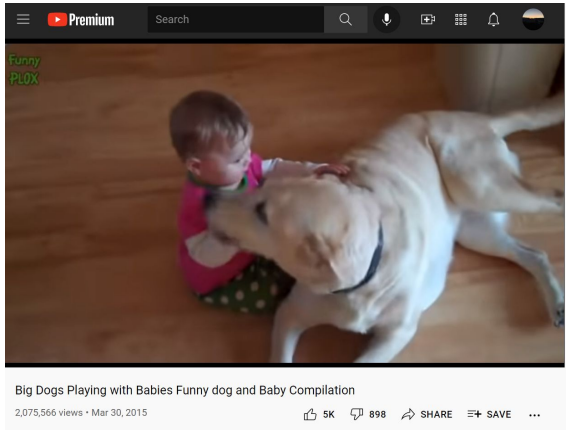
- What other data can be used for V+L pre-training? Can you think of other settings in which different modalities complement/supplement each other?

How We Perceive the World: Information with Multiple Modalities



Instagram

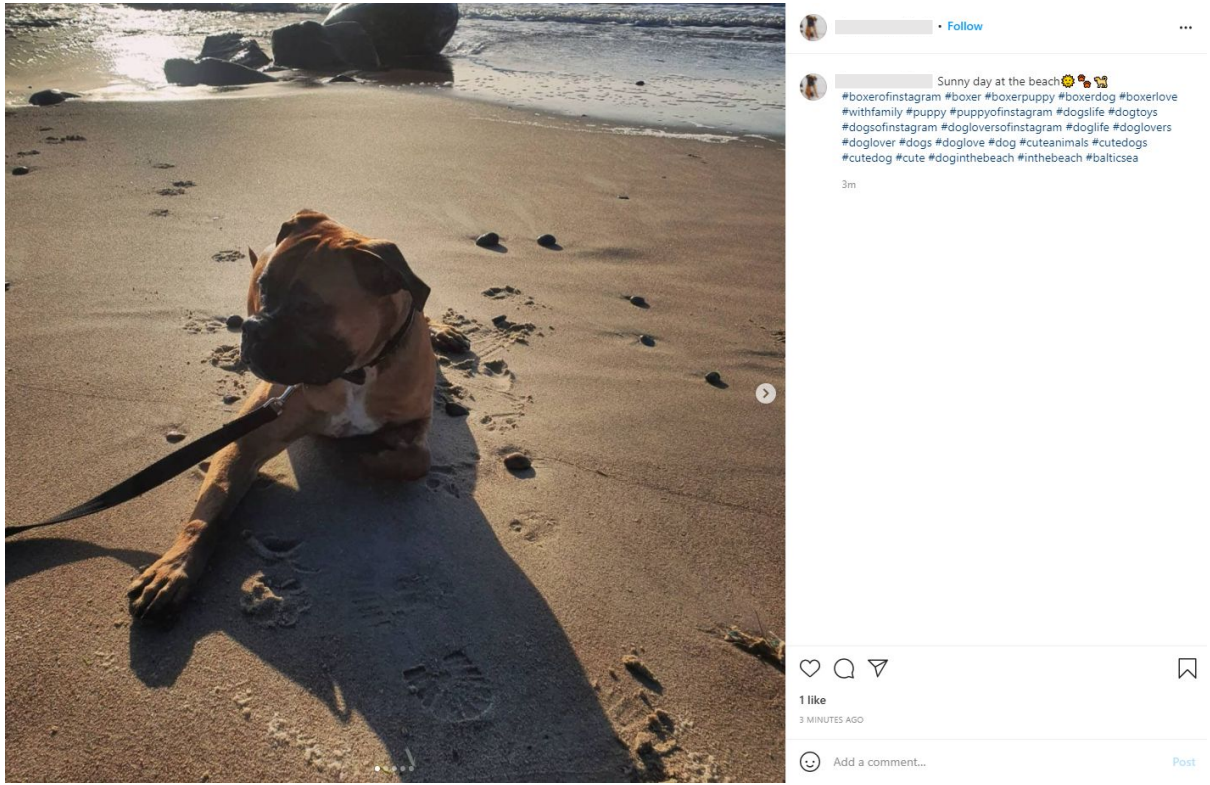




YouTube

Image Source:
Srinivasan, Krishna, et al. "WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning." *arXiv preprint arXiv:2103.01913* (2021).
https://www.instagram.com/p/CW1blk0BAQa/
Chen, Xinlei, et al. "Microsoft coco captions: Data collection and evaluation server." *arXiv preprint arXiv:1504.00325* (2015).
https://www.instagram.com/p/CWyjRlHhwGU/
https://www.youtube.com/watch?v=KPP6ebNbFXQ

Instagram: Photo-Caption Pair



• Follow

Sunny day at the beach🌻🐾🐕
#boxerofinstagram #boxer #boxerpuppy #boxerdog #boxerlove #withfamily #puppy #puppyofinstagram #dogslife #dogtoys #dogsofinstagram #dogloversofinstagram #doglife #doglovers #doglover #dogs #doglove #dog #cuteanimals #cutedogs #cutedog #cute #doginthebeach #inthebeach #balticsea

3m

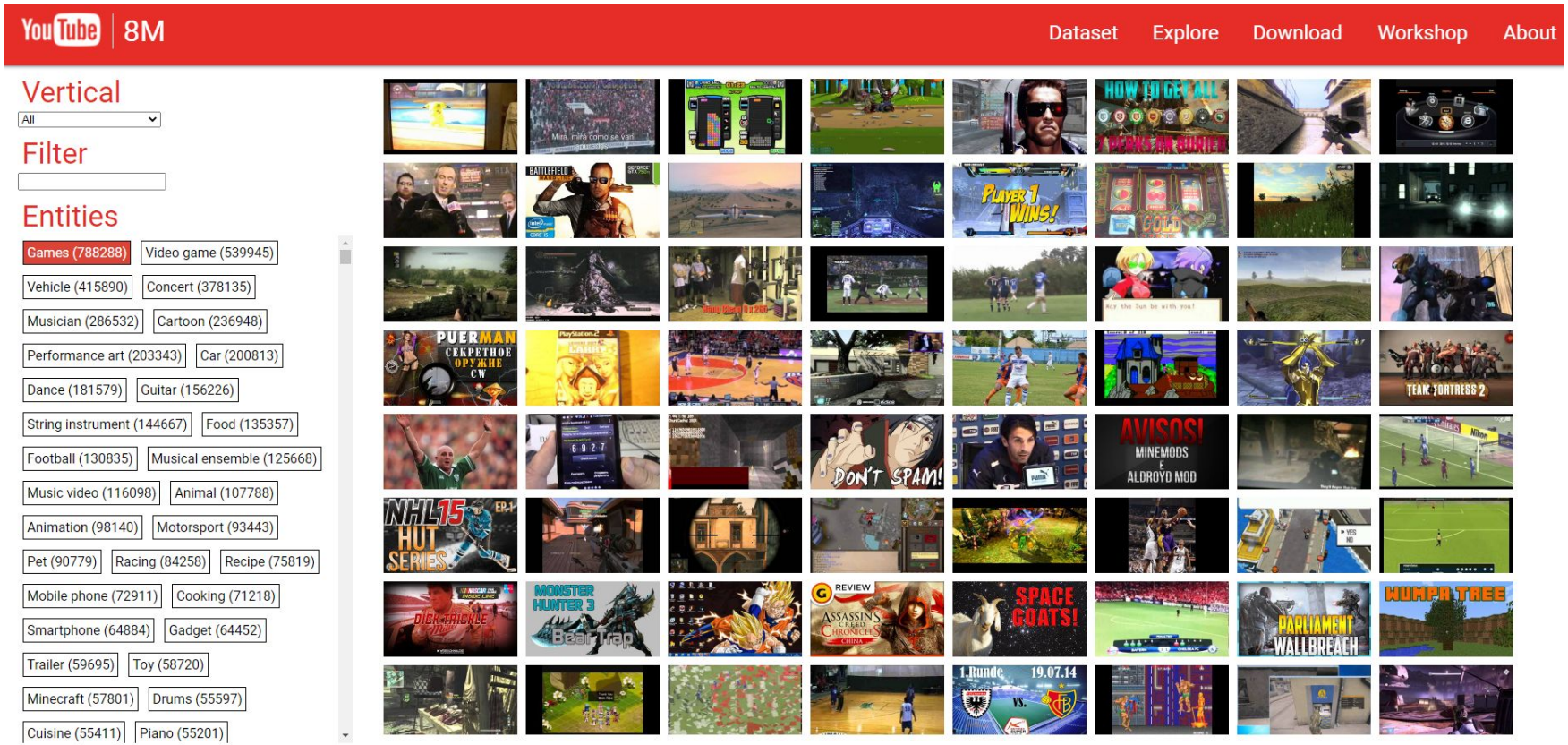
1 like

3 MINUTES AGO

Add a comment...

Post

Youtube: Video + Audio Transcript Dataset



[Image Source]: https://research.google.com/youtube8m/

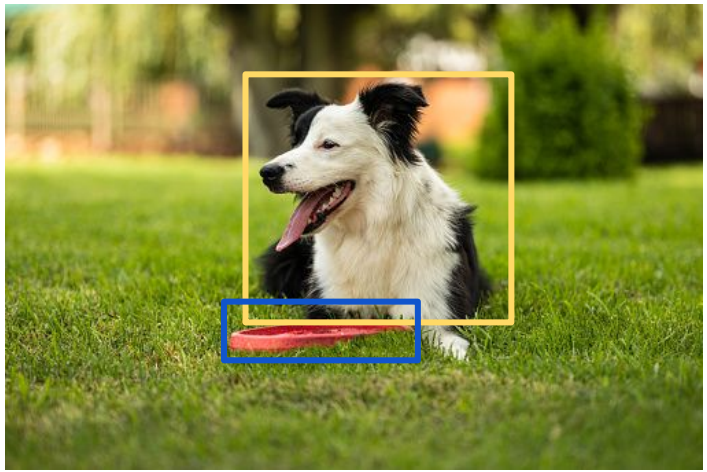
2. Feature Representation for Vision + Language

Image and Text Dataset Pair

( , a black and white dog sitting next to a frisbee.)

Slide format borrowed from: Self-supervised Learning presented by Licheng Yu , Linjie Li and Yen-Chun Chen.
https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part5-pretraining.pdf

Feature Representation for Vision and Language



‘a’, ‘black’, ‘and’, ‘white’, ‘dog’, ‘sitting’, ‘next’, ‘to’, ‘a’, ‘frisbee’

Slide format borrowed from: Self-supervised Learning presented by Licheng Yu , Linjie Li and Yen-Chun Chen.
https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part5-pretraining.pdf

Visual Features

Grid Features vs. Region Features

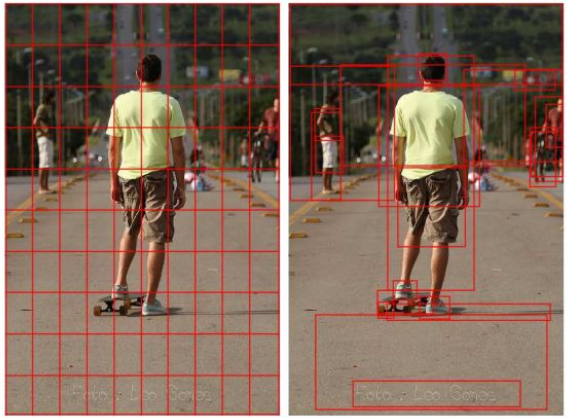
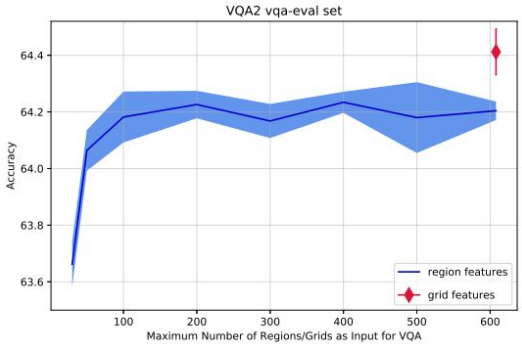


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

Bottom-Up and Top-Down Attention Model (IEEE, 2018)



In Defense of Grid Features for Visual Question Answering (IEEE/CVF 2020)

VinVL: Revisiting Visual Representations (IEEE/CVF 2021)

Image Source:
Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
Jiang, Huaizu, et al. "In defense of grid features for visual question answering." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
Zhang, Pengchuan, et al. "Vinvl: Revisiting visual representations in vision-language models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

3. Architecture

How to Learn Joint V+L Representations?

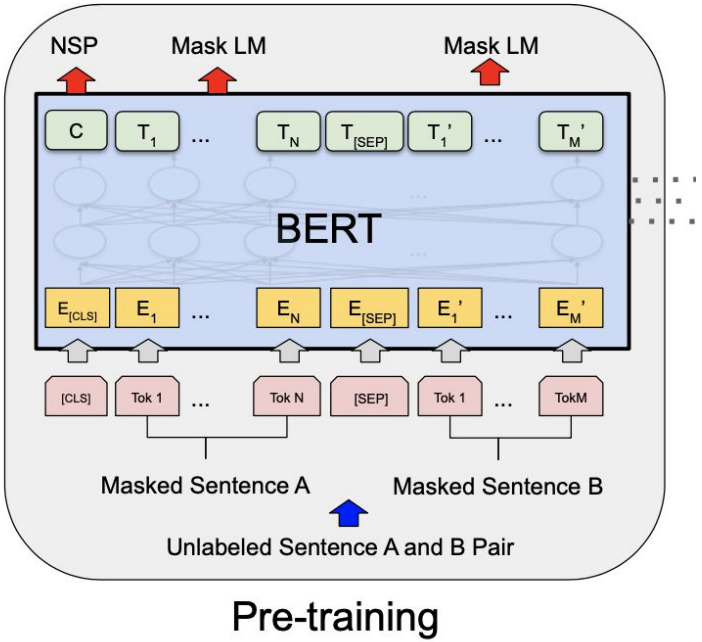
Previously, multimodal embeddings were tailored for specific downstream tasks.

- e.g., MCB (Fukui et al., 2017), BAN (Kim et al., 2018), DFAF (Gao et al., 2019), SCAN (Lee et al., 2018) and MAttNet (Yu et al., 2018)

Motivation: Self-Supervised Learning for NLP

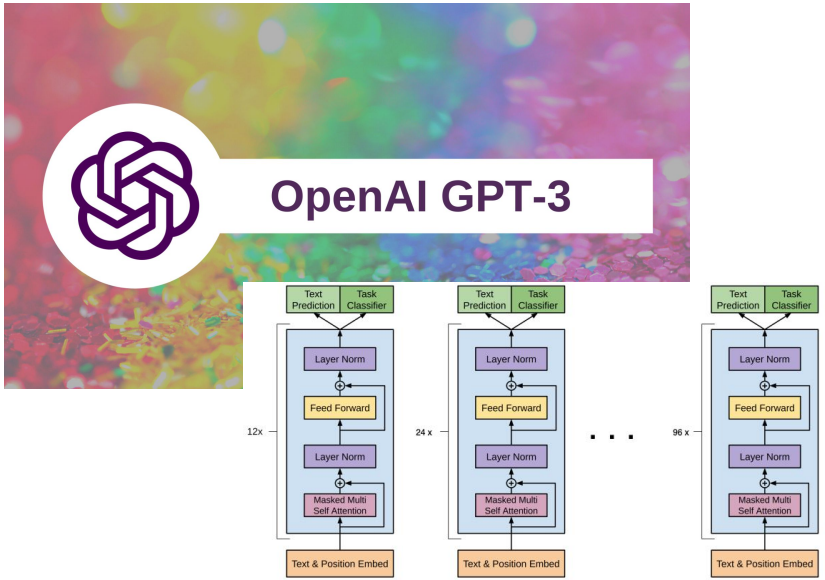
Pretrain-then-finetune

BERT



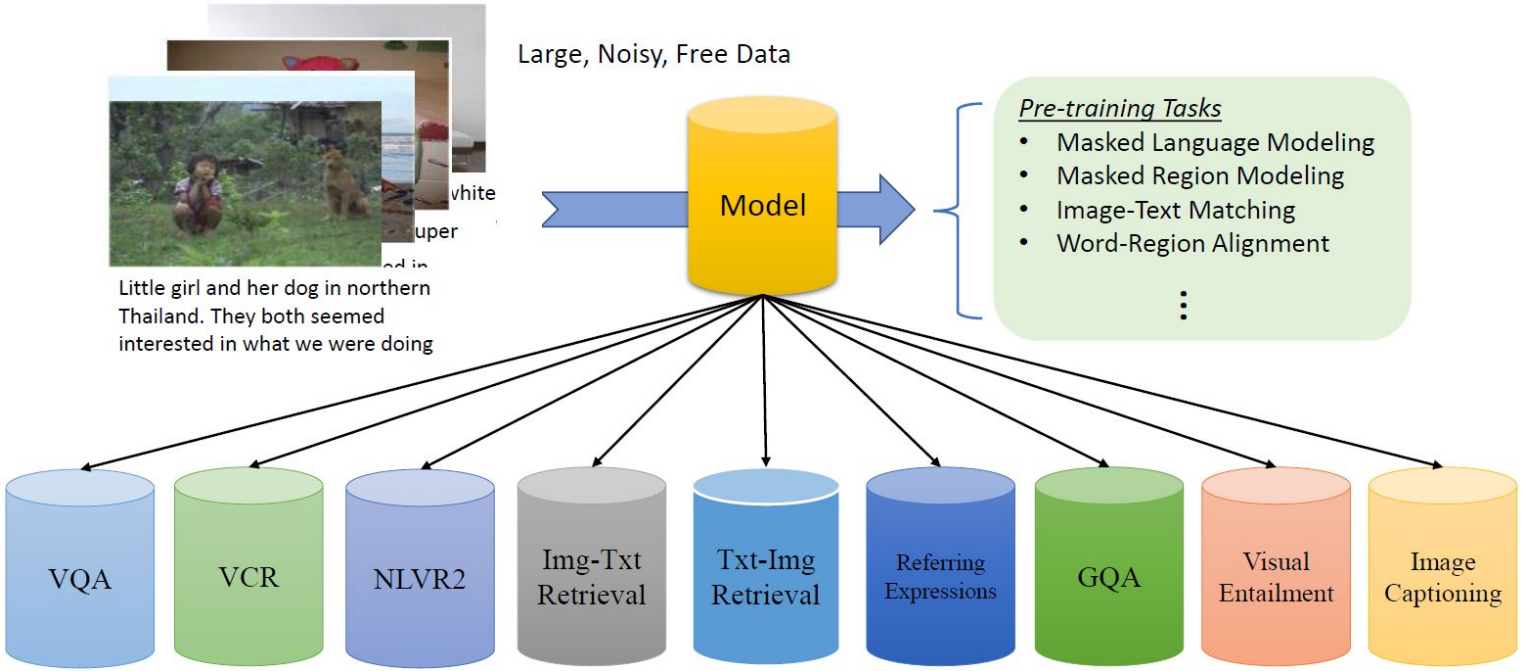
Pre-training

GPT-3



- Contextual representations using transformer architecture + 2) effective pre-training tasks

Self-Supervised Learning for V+L Tasks



Slide/Image Borrowed from CVPR 2020 Tutorial: Recent Advances in Vision-and-Language Research
Opening Remarks presented by JJ Liu and Xiaodong He
https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part-1-opening.pdf

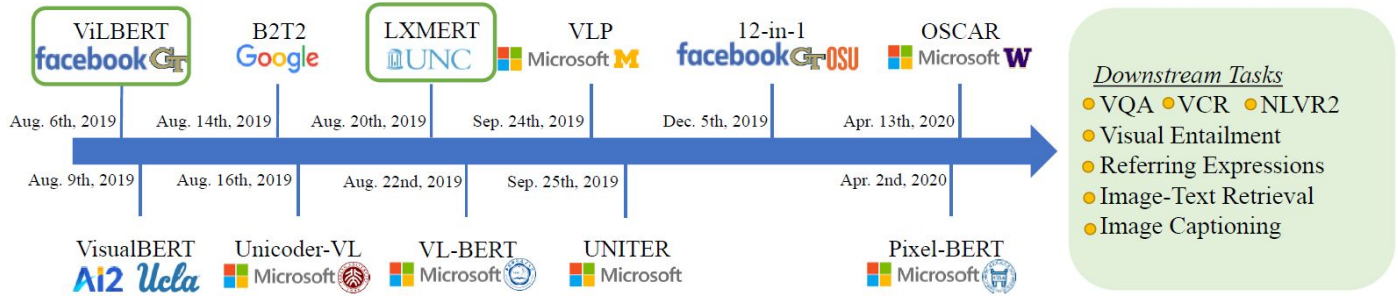
Universal V+L Representation: Pretrained Multimodal Embeddings

Learning Pretrained V+L Representations:

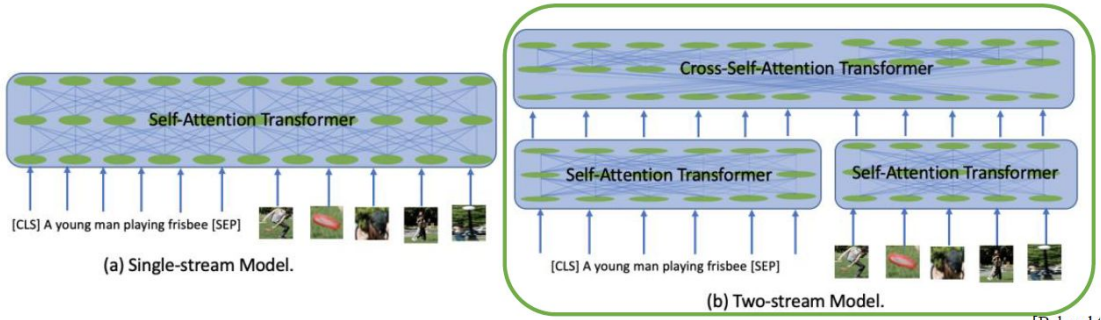
1. **Two-stream architecture**
   - Two Transformers for each modality, later fused by a third Transformer.
     - e.g., ViLBERT (Lu, Jiasen, et al. 2019), LXMERT (Tan, Hao, and Mohit Bansal. 2019))

2. **Single-stream architecture**
   - Single Transformer is applied to both images and text.
     - e.g., VisualBERT (Li, Liunian Harold, et al. 2019), Unicoder-VL (Li, Gen, et al., 2020), VL-BERT (Su, Weijie, et al. 2019), UNITER (Chen, Yen-Chun, et al., 2020).

Models using Two-stream Architecture



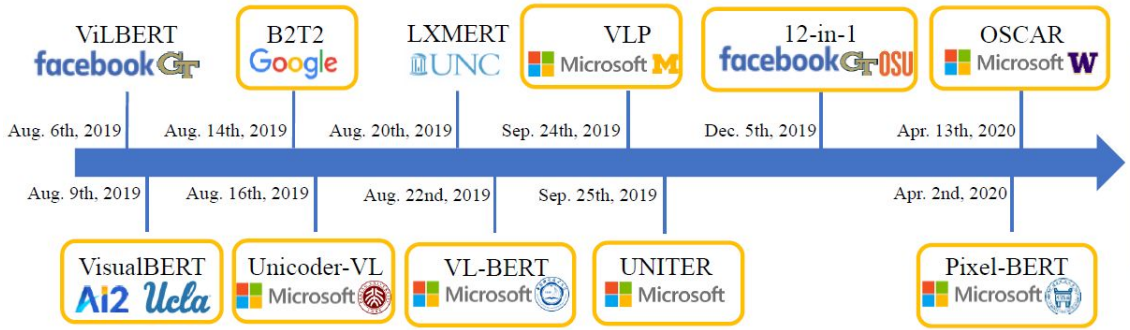
Model Architecture:



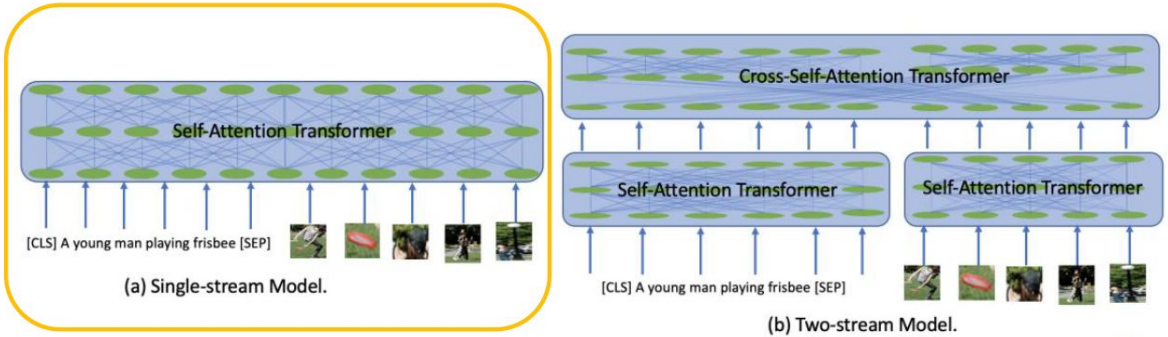
[Behand the Scene; Cao et al 2020]

Slide Image borrowed from: CVPR 2020 Tutorial: Self-supervised Learning presented by Licheng Yu , Linjie Li and Yen-Chun Chen.
https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part5-pretraining.pdf

Models using Single-stream Architecture



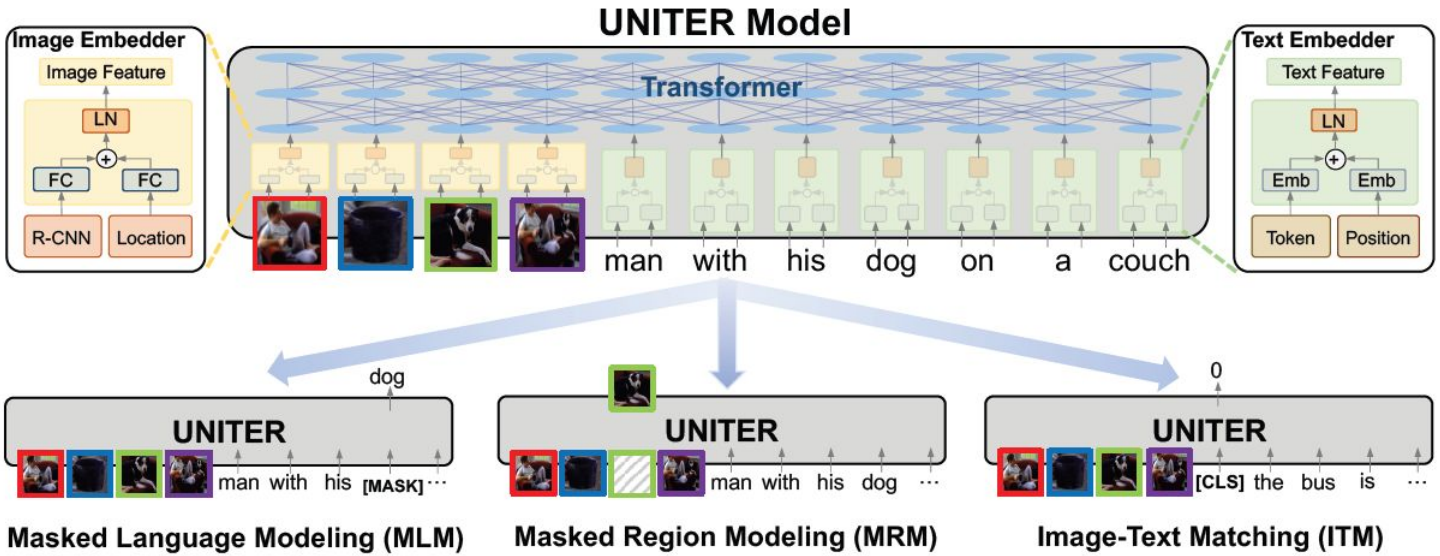
Model Architecture:



[Behand the Scene; Cao et al 2020]

Slide Image borrowed from: CVPR 2020 Tutorial: Self-supervised Learning presented by Licheng Yu , Linjie Li and Yen-Chun Chen.
https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part5-pretraining.pdf

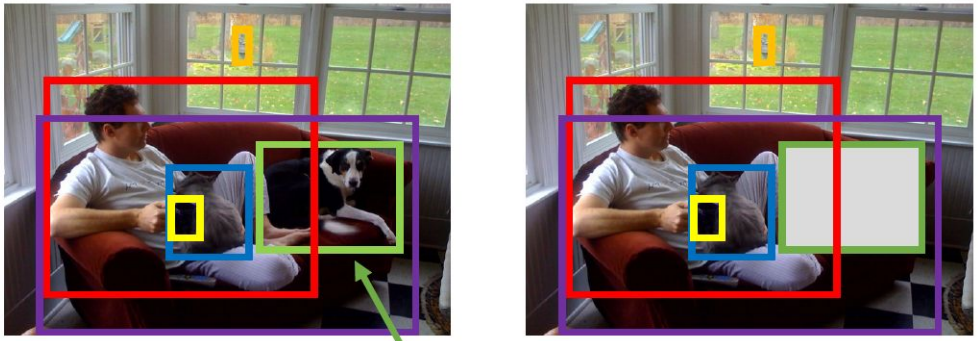
Single-stream Architecture: UNITER Example



Chen, Yen-Chun, et al. "Uniter: Universal Image-text Representation Learning." *European Conference on Computer Vision*. Springer, Cham, 2020.

Designing Pretraining Tasks to Improve Multimodal Representations

Example: Conditional Masking Strategy



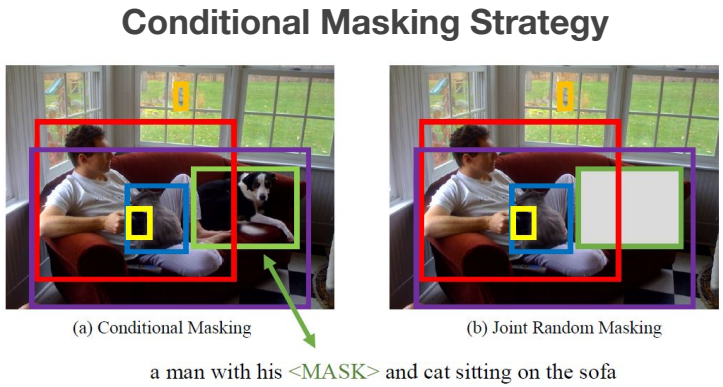
(a) Conditional Masking

(b) Joint Random Masking

a man with his <MASK> and cat sitting on the sofa

UNITER
(Chen, Yen-Chun, et al., 2019)

Question



Q. Why do you think conditional masking works better than traditional random masking strategy for V+L tasks?

Multimodal Attention Heads - UNITER example.

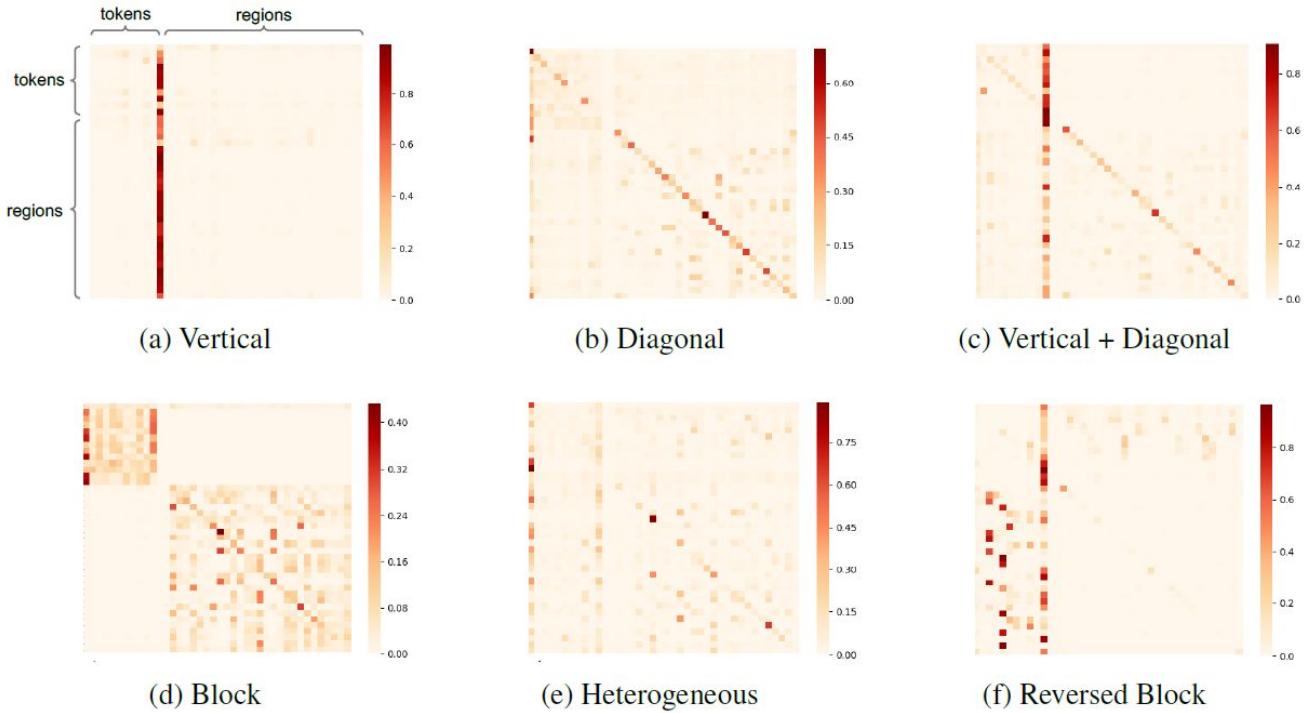
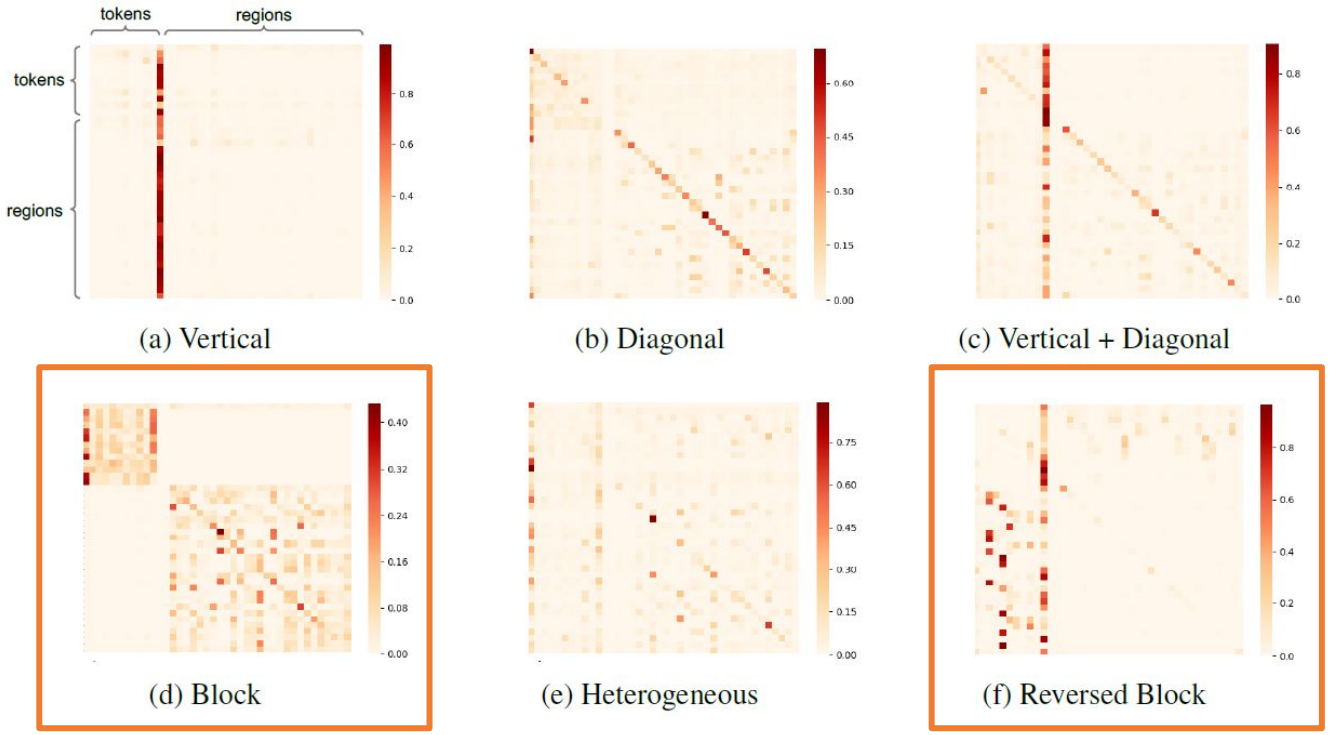


Figure 3: Attention heatmaps of UNITER

Multimodal Attention Heads - UNITER example.



Intra-modality attention

Figure 3: Attention heatmaps of UNITER

Inter-modality attention

Papers

Three Papers on Learning Visual Concepts using Natural Language Supervision

Feature Representation

- **Paper 1: "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks." (Li, Xiujun, et al. 2020)**

Pretraining Task

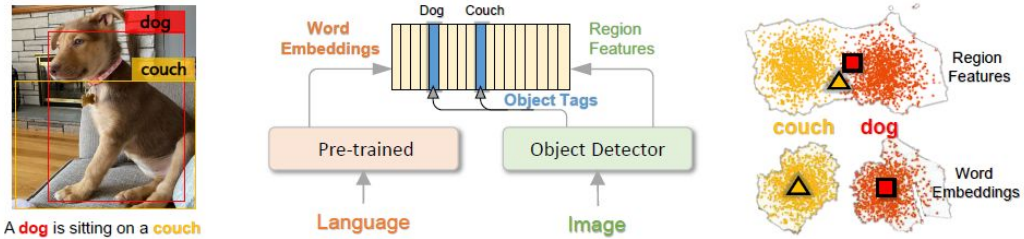
- Paper 2: "ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration" (Cui, Yuhao et al., 2021)

Data/Pretraining Architecture

- Paper 3: "Learning Transferable Visual Models from Natural Language Supervision." (Radford, Alec, et al., 2021)

Paper 1.

Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks (ECCV, 2020)



(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces

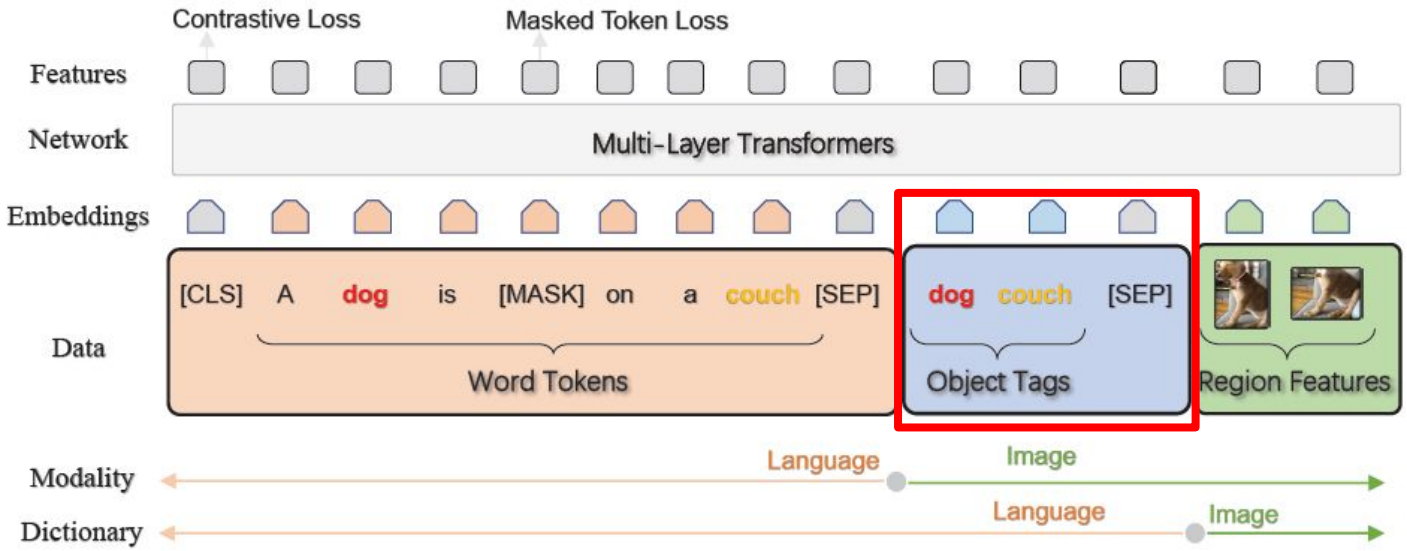
Key Contribution: Improving Feature Representations for Better Semantic Alignment

- Previously, different modalities (i.e., text and image features) were simply concatenated to be trained in the model. More recently, researchers have explored ways to enhance the semantic alignments in the shared text and image representation space.

- Object tags as anchor points to align text-image.

[Figure Source] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." *European Conference on Computer Vision*. Springer, Cham, 2020.

Inputs: Region Features, Word Tokens, and Object Tags



Object Tags as Anchor Points to Align Image-Text

[Figure Source] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." *European Conference on Computer Vision*. Springer, Cham, 2020.

Oscar Pre-training

Data:

6.5 million text-tag-img triplets: (w, q, v)

w = word embedding sequence, q = word embedding sequence of object tags, v = set of region features

- COCO
- Conceptual Captions
- SBU Captions
- Flicker 30k
- GQA

The full pre-training objective of OSCAR is:

$$\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{C}}.$$

Pre-Training Objective The OSCAR input can be viewed from two different perspectives as

$$x \triangleq [\underbrace{w}_{\text{language}}, \underbrace{q, v}_{\text{image}}] = [\underbrace{w, q}_{\text{language}}, \underbrace{v}_{\text{image}}] \triangleq x' \quad (1)$$

1. Modality View **2. Dictionary View**

Contrastive Loss

$$h' \triangleq [q, v]$$

Sample a set of "polluted" image representations:

- Replace object tag with 50% probability

Predict original (y =1) vs polluted (y = 0) ones using contrastive loss.

$$\mathcal{L}_{\text{C}} = -\mathbb{E}_{(h', w)\sim\mathcal{D}} \log p(y|f(h', w)).$$

Masked Token Loss

Mask word-object tag token pair.

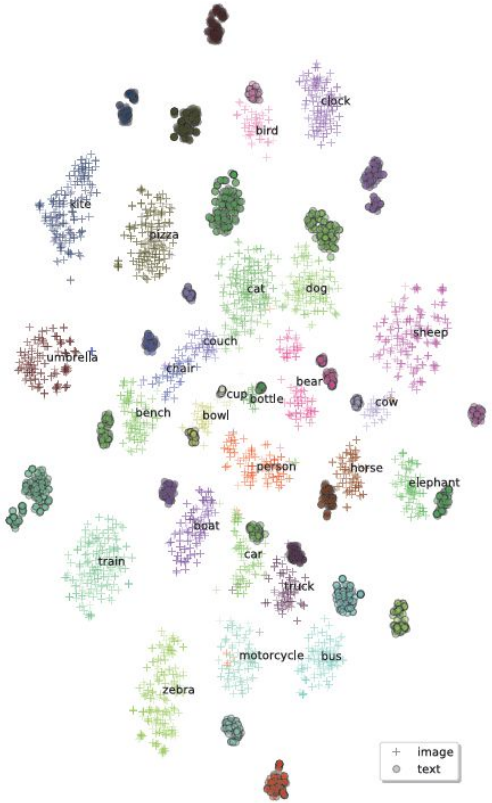
$$\textit{discrete token sequence as } h \triangleq [w, q]$$

$$\mathcal{L}_{\text{MTL}} = -\mathbb{E}_{(v,h)\sim\mathcal{D}} \log p(h_i|h_{\setminus i}, v)$$

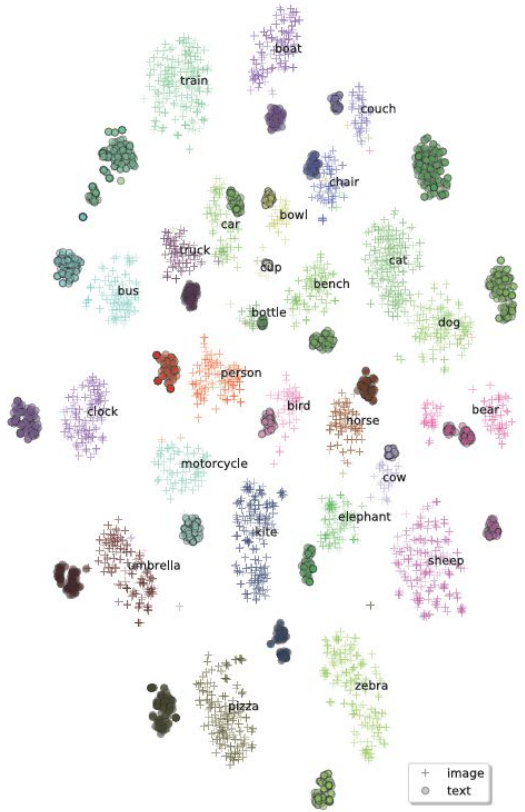
[Figure Source] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." *European Conference on Computer Vision*. Springer, Cham, 2020.

Results: Aligning Image and Text Features in a Shared Embedding Space

T-SNE Visualization



Baseline



OSCAR

[Figure Source] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." *European Conference on Computer Vision*. Springer, Cham, 2020.

Results on V+L Tasks

Table 1: Overall results on six tasks. Δ indicates the improvement over SoTA. SoTA with subscript S, B, L indicates performance achieved by small models, VLP of similar size to BERT base and large model, respectively. Most results are from [5], except that image captioning results are from [11,46], NoCaps results are from [1], VQA results are from [38].

Task	Image Retrieval			Text Retrieval			Image Captioning				NoCaps		VQA	NLVR2
	R@1	R@5	R@10	R@1	R@5	R@10	B@4	M	C	S	C	S	test-std	test-P
SoTA$_S$	39.2	68.0	81.3	56.6	84.5	92.0	38.9	29.2	129.8	22.4	61.5	9.2	70.90	53.50
SoTA$_B$	48.4	76.7	85.9	63.3	87.0	93.1	39.5	29.3	129.3	23.2	73.1	11.2	72.54	78.87
SoTA$_L$	51.7	78.4	86.9	66.6	89.4	94.3	–	–	–	–	–	–	73.40	79.50
OSCAR$_B$	**54.0**	**80.8**	**88.5**	**70.0**	**91.1**	**95.5**	**40.5**	**29.7**	**137.6**	22.8	**78.8**	**11.7**	**73.44**	78.36
OSCAR$_L$	**57.5**	**82.8**	**89.8**	**73.5**	**92.2**	**96.0**	**41.7**	**30.6**	**140.0**	**24.5**	**80.9**	**11.3**	**73.82**	**80.37**
Δ	**5.8** ↑	**4.4** ↑	**2.9** ↑	**6.9** ↑	**2.8** ↑	**1.7** ↑	**2.2** ↑	**1.3** ↑	**10.7** ↑	**1.3** ↑	**7.8** ↑	**0.5** ↑	**0.42** ↑	**0.87** ↑

[Figure Source] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." *European Conference on Computer Vision*. Springer, Cham, 2020.

Discussion Questions

- Oscar used object tags as anchor points to align vision-language modalities. What other features/knowledge from image-text pair might be useful?

- Are there any limitations of using object tags?

Papers

Three Papers: Learning Visual Concepts using Natural Language Supervision

Data/Feature Representation

- Paper 1: "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks." (Li, Xiujun, et al. 2020)

Pretraining Task

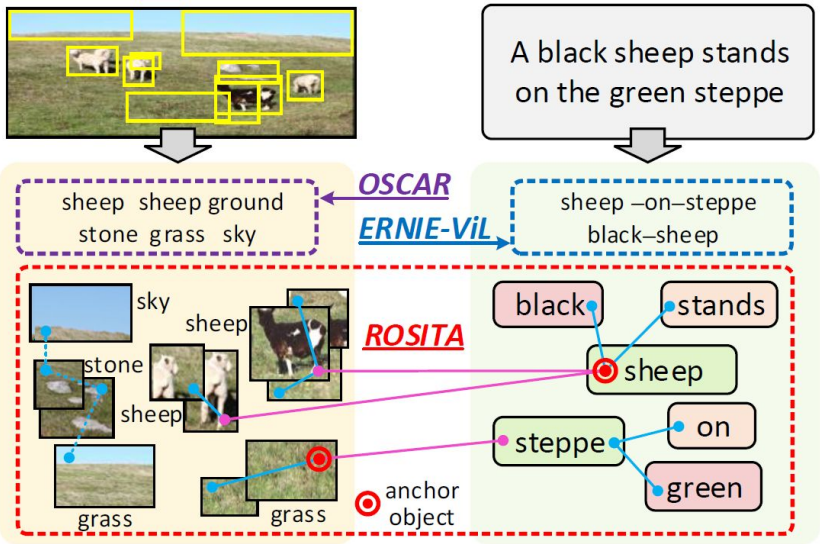
- **Paper 2: "ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration" (Cui, Yuhao et al., 2021)**

Data/Pretraining Architecture

- Paper 3: "Learning Transferable Visual Models from Natural Language Supervision." (Radford, Alec, et al., 2021)

Paper 2.

ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration

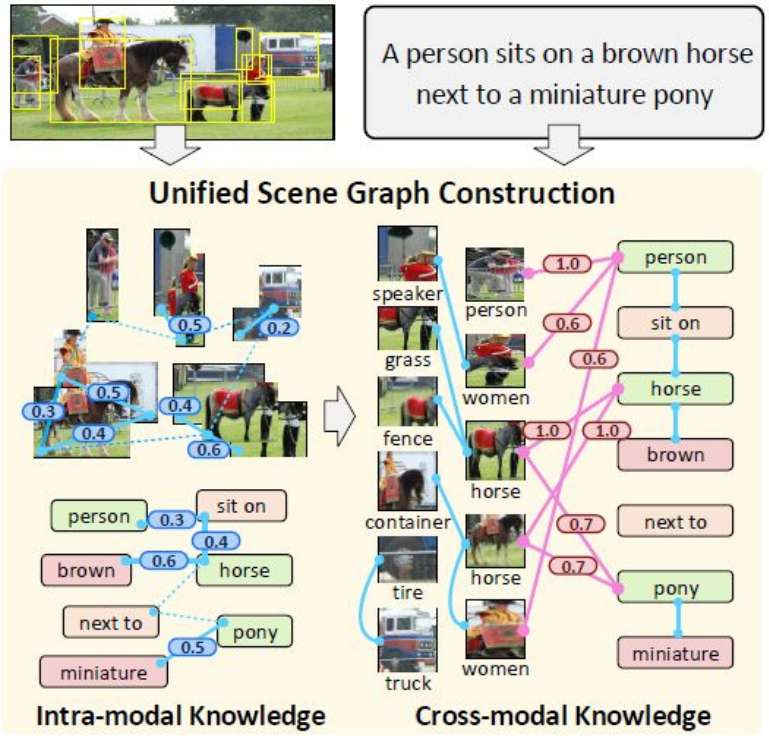


Exploiting cross- and intra- modal knowledge.

Idea: Represent image features/text modalities as a unified scene graph.

Structural knowledge masking strategy.

1. Unified Scene Graph Construction



2. Knowledge Representation

Anchor object: vertex with at least one cross-modal edge.

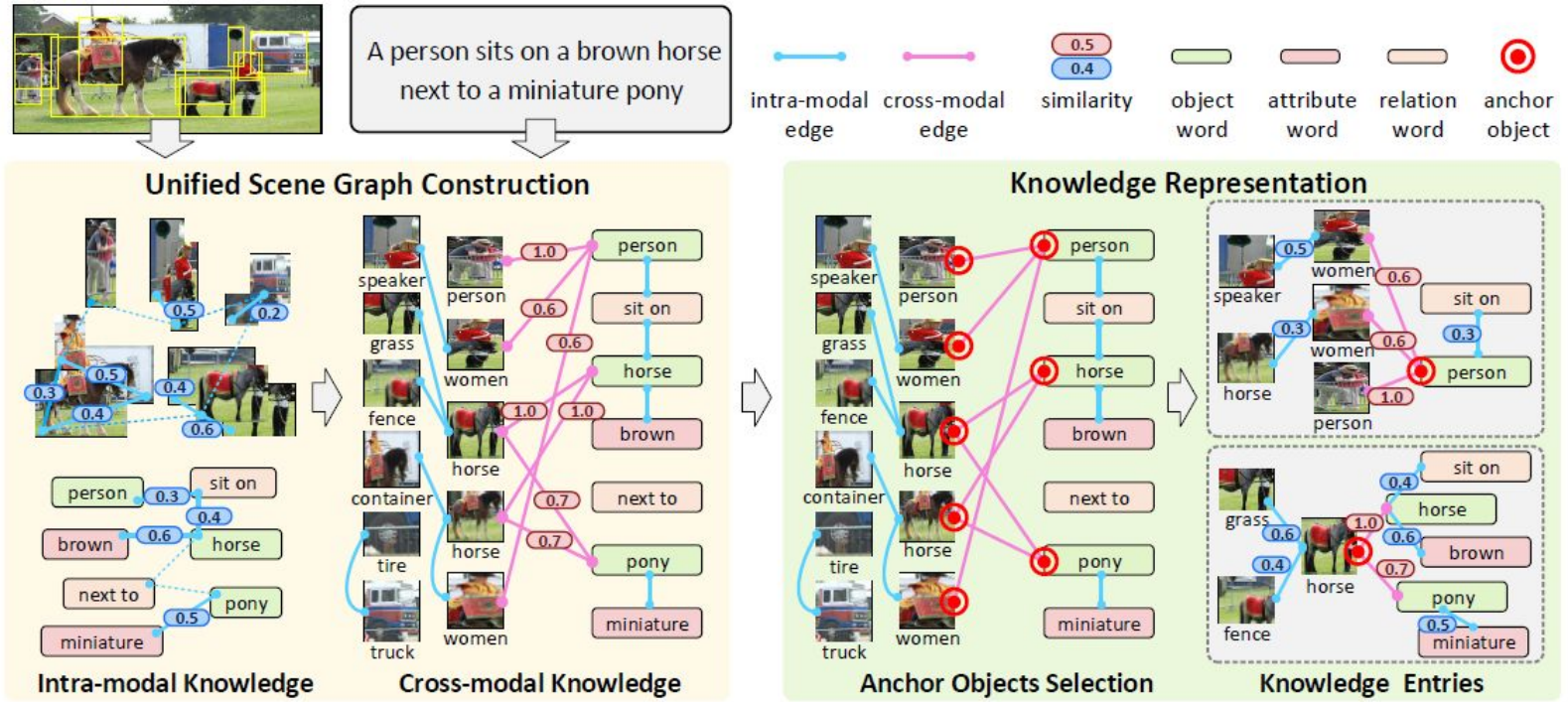


Figure 2: The flowchart of knowledge extraction given an image-text pair. It consists of two main stages, namely the unified scene graph construction and knowledge representation.

Structural Knowledge Masking

- Previously, random masking strategies of input tokens (i.e., words or regions) for masked language modeling (MLM) and masked region modeling (MRM).

- Structural knowledge masking strategy:

$$p_j = \begin{cases} 1, & \text{if } v_j \text{ is the anchor object} \\ \alpha t_j, & \text{if } v_j \text{ is within the intra-modal contexts} \\ (1-\alpha)(1-t_j), & \text{if } v_j \text{ is within the cross-modal contexts} \end{cases} \tag{7}$$

Main Results

task	dataset		ViLBERT† [27]	VLBERT† [36]	Unicoder-VL [19]	LXMERT [38]	UNITER [7]	ERNIE-ViL† [45]	VILLA [9]	OSCAR [23]	ROSITA (ours)
VQA	VQAv2	test-dev	70.55	71.16	-	72.42	72.70	72.62	73.59	73.16	73.91
		test-std	70.92	-	-	72.54	72.91	72.85	73.67	73.44	73.97
REC	Ref-COCO	val^d	-	-	-	-	81.24	-	81.65	-	84.79
		testA^d	-	-	-	-	86.48	-	87.40	-	87.99
		testB^d	-	-	-	-	73.94	-	74.48	-	78.28
	Ref-COCO+	val^d	72.34	71.60	-	-	75.31	74.02	76.05	-	76.06
		testA^d	78.52	77.72	-	-	81.30	80.33	81.65	-	82.01
		testB^d	62.61	60.99	-	-	65.68	64.74	65.70	-	67.40
	Ref-COCOg	val^d	-	-	-	-	74.31	-	75.90	-	78.23
		test^d	-	-	-	-	74.51	-	75.93	-	78.25
ITR	IR-COCO	R@1	-	-	46.70	-	50.33	-	-	54.00	54.40
		R@5	-	-	76.00	-	78.52	-	-	80.80	80.92
		R@10	-	-	85.30	-	87.16	-	-	88.50	88.60
	TR-COCO	R@1	-	-	62.30	-	64.40	-	-	70.00	71.26
		R@5	-	-	87.10	-	87.40	-	-	91.10	91.62
		R@10	-	-	92.80	-	93.08	-	-	95.50	95.58
	IR-Flickr	R@1	58.20	-	71.50	-	72.52	74.44	74.74	-	74.08
		R@5	84.90	-	90.90	-	92.36	92.72	92.86	-	92.44
		R@10	91.52	-	94.90	-	96.08	95.94	95.82	-	96.08
	TR-Flickr	R@1	-	-	86.20	-	85.90	86.70	86.60	-	88.90
		R@5	-	-	96.30	-	97.10	97.80	97.90	-	98.10
		R@10	-	-	99.00	-	98.80	99.00	99.20	-	99.30

Visualization of Cross-Modal Attention

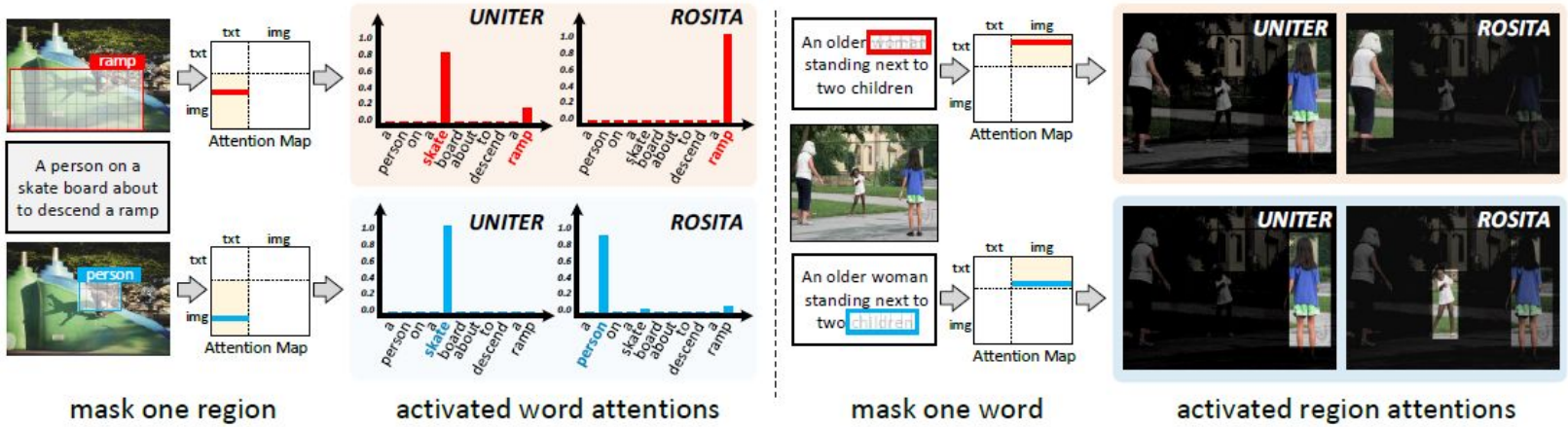
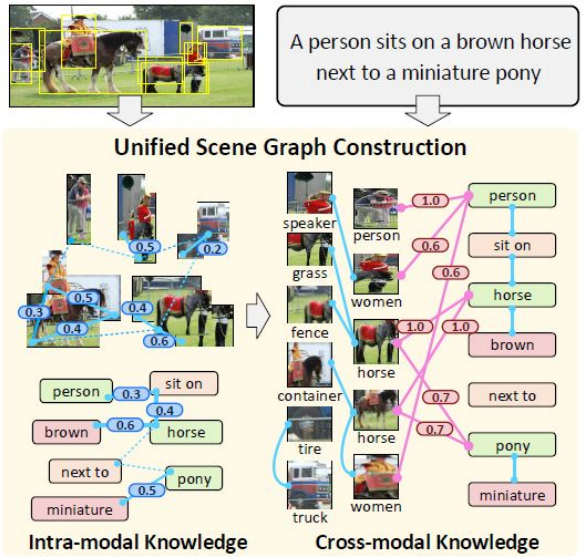


Figure 4: Visualizations of the learned cross-modal attentions (*i.e.*, region-to-words attentions on the left and word-to-regions on the right) from UNITER [7] and ROSITA. Taking the image-text pair as inputs with exactly one region (or word) being masked at a time, we extract the attention map from the last MSA block of the pretrained model. The region-to-words (word-to-regions) attentions correspond to one specific row in the bottom-left (top-right) area of the attention map, respectively.

Discussion Questions

- Do you see any limitations of the model?
- Under what circumstances/datasets would this pretraining model work/not work?



Papers

Three Papers: Learning Visual Concepts using Natural Language Supervision

Data/Feature Representation

- Paper 1: "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks." (Li, Xiujun, et al. 2020)

Pretraining Task

- Paper 2: "ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration" (Cui, Yuhao et al., 2021)

Data/Pretraining Architecture

- **Paper 3: "Learning Transferable Visual Models from Natural Language Supervision." (Radford, Alec, et al., 2021)**

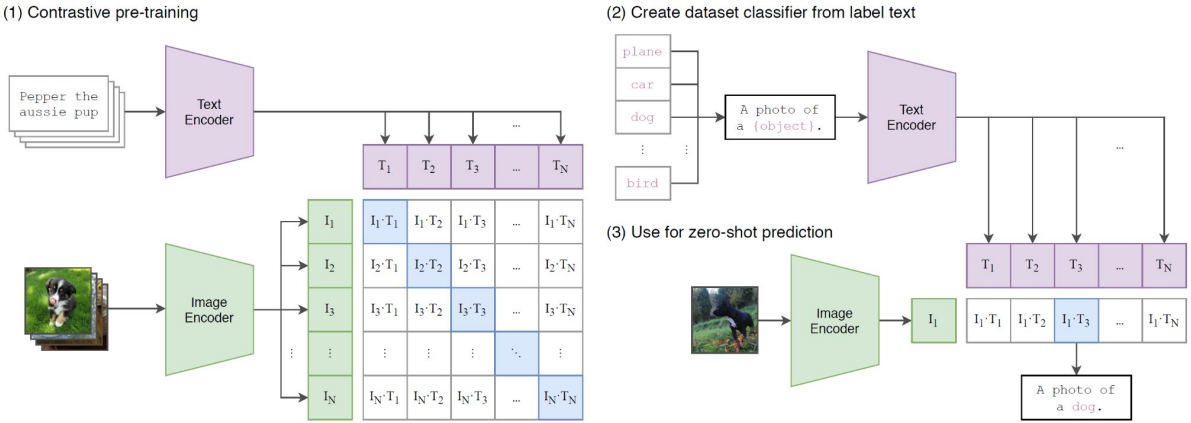
Paper 3.
Learning Transferable Visual Models from Natural Language Supervision (aka CLIP)

Data

1. Previously, object classification in computer vision relied on crowd-labeled datasets (e.g.,MS-COCO, Visual Genome, YFCC100M).
   - Challenge for dataset creation/zero-shot learning on unseen image/concept.

Contrastive Pretraining

2. Contrastive learning to learn relationship between image-text pair.



Learning “Concept”: CLIP’s Zero-Shot Transfer Capabilities

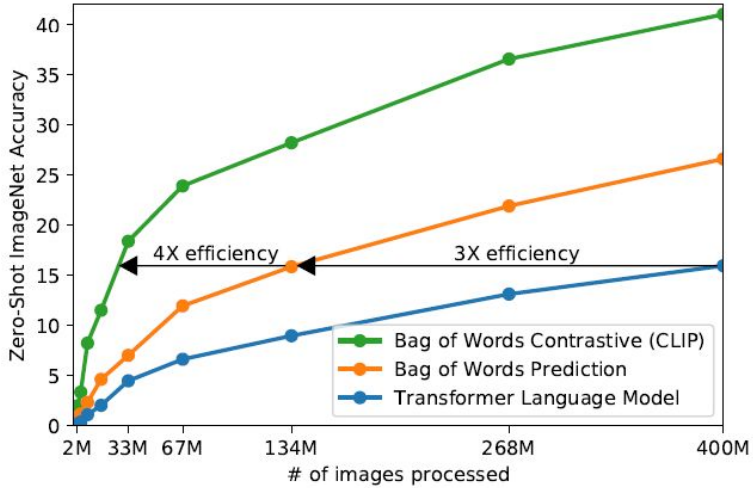


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

Data

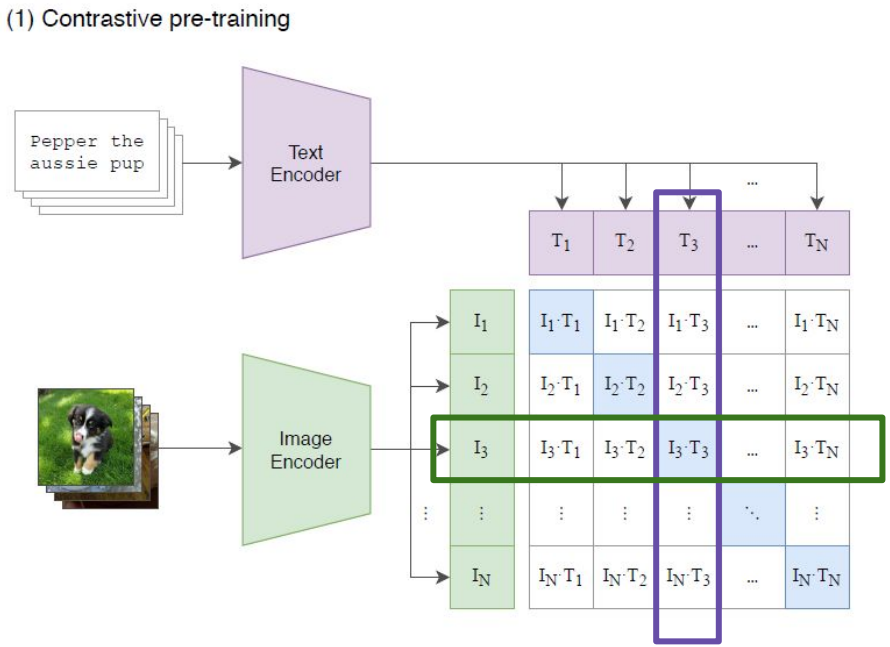
New dataset of **400 million** image-text pairs.

- Note: Oscar used 6.5 million

500,000 queries from a variety of publicly available sources on the Internet.

- Base query list: all words occurring at least 100 times in Wikipedia (English version), augmented with bi-grams, with names of wikipedia article names above a certain search volume. Finally, augmented with WordNet synsets.

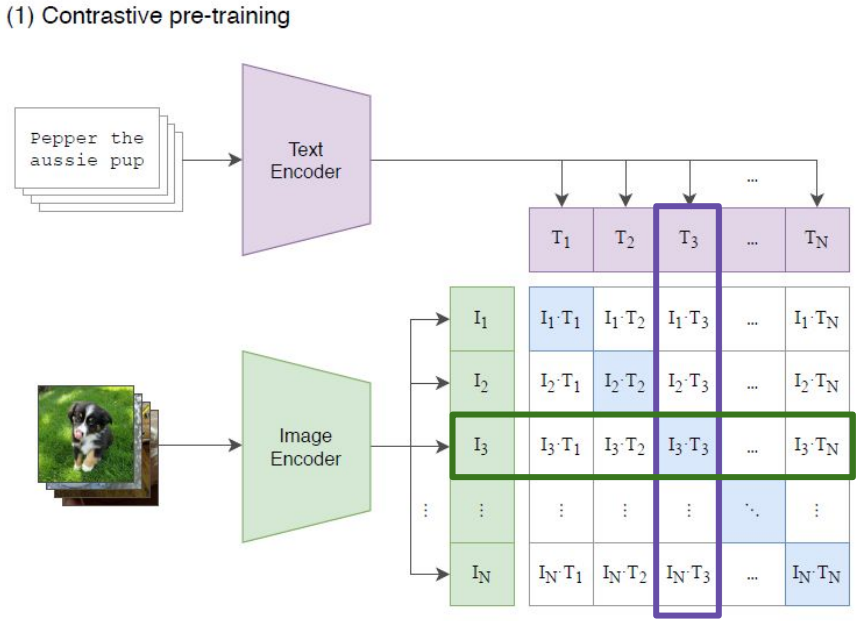
CLIP - Pseudocode




```
# 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
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

CLIP



QUESTION:

What is the value of CLIP's pretraining approach vs. pretraining a model on a image classification task?

CLIP vs. Supervised Linear Classifier on ResNet-50

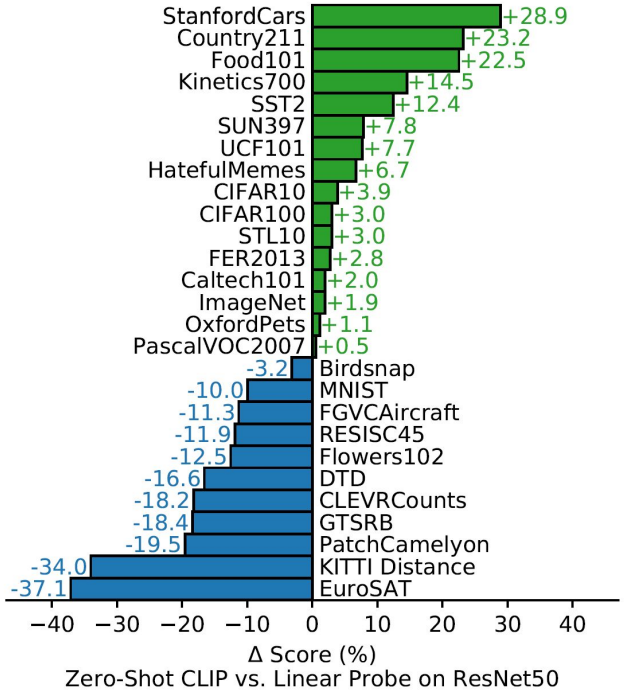
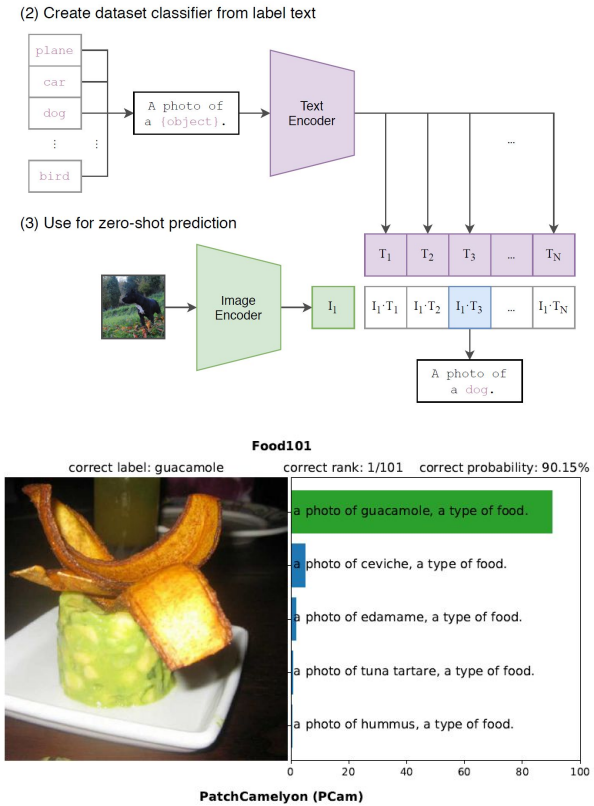


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

Zero-shot performance & prompt engineering.

{label} vs. “A photo of a {label}, a type of pet.”



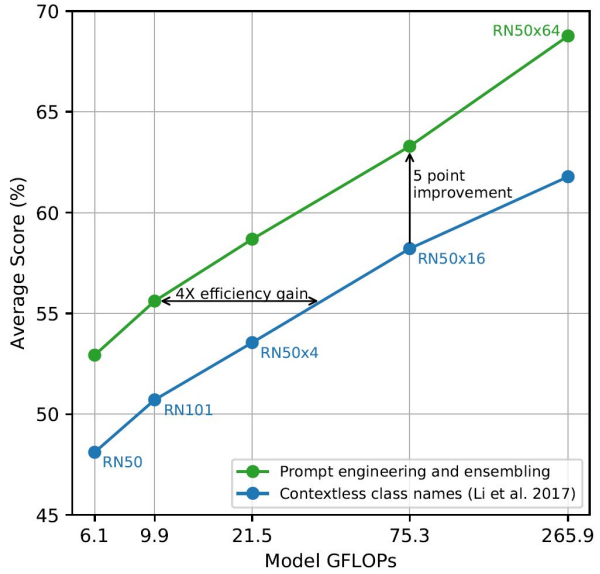
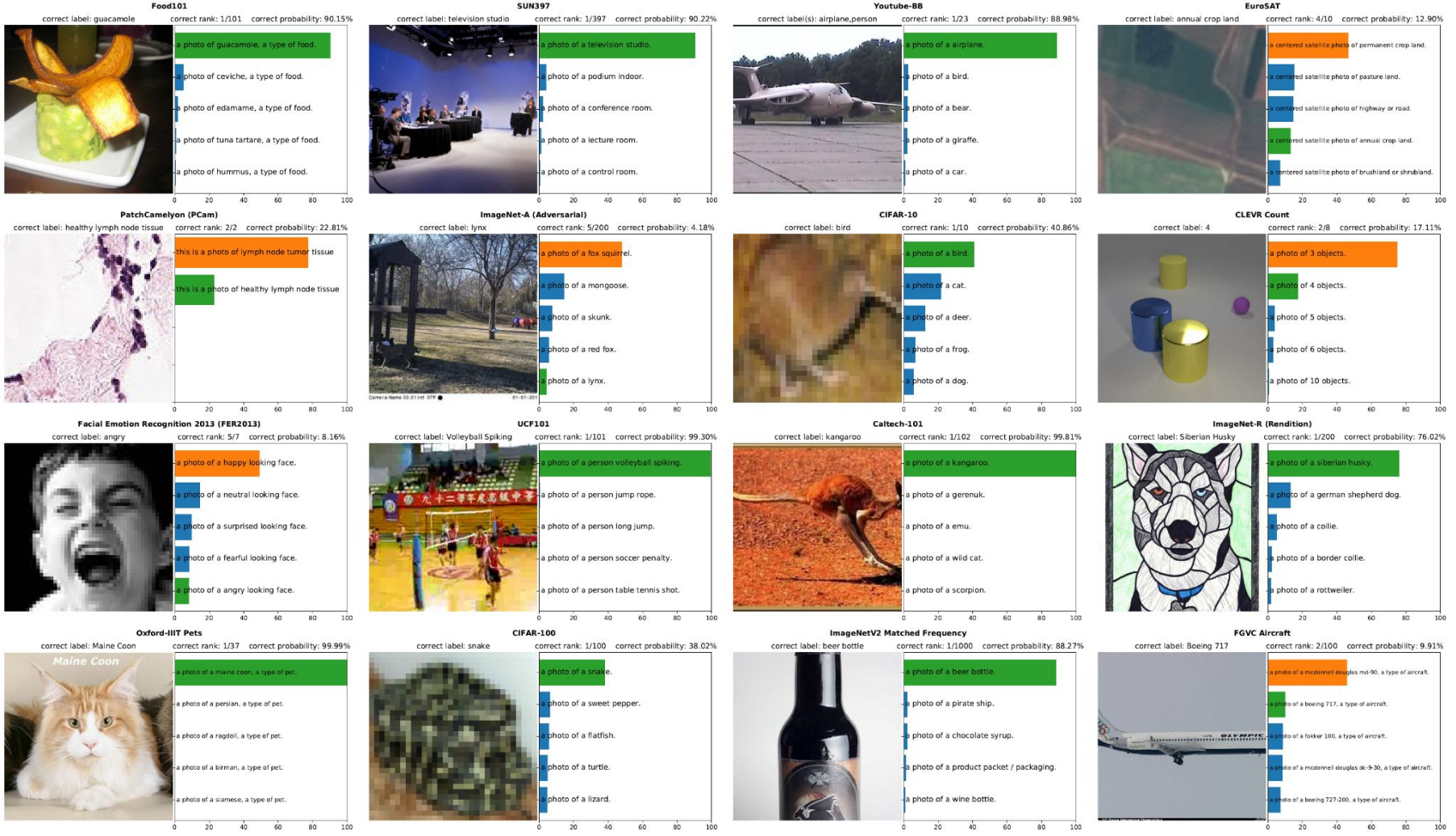


Figure 4. **Prompt engineering and ensembling improve zero-shot performance.** Compared to the baseline of using contextless class names, prompt engineering and ensembling boost zero-shot classification performance by almost 5 points on average across 36 datasets. This improvement is similar to the gain from using 4 times more compute with the baseline zero-shot method but is “free” when amortized over many predictions.


Food101
correct label: guacamole
correct rank: 1/101
correct probability: 90.15%
a photo of guacamole, a type of food.
a photo of ceviche, a type of food.
a photo of edamame, a type of food.
a photo of tuna tartare, a type of food.
a photo of hummus, a type of food.
SUN397
correct label: television studio
correct rank: 1/397
correct probability: 90.22%
a photo of a television studio.
a photo of a podium indoor.
a photo of a conference room.
a photo of a lecture room.
a photo of a control room.
Youtube-BB
correct label(s): airplane,person
correct rank: 1/23
correct probability: 88.98%
a photo of a airplane.
a photo of a bird.
a photo of a bear.
a photo of a giraffe.
a photo of a car.
EuroSAT
correct label: annual crop land
correct rank: 4/10
correct probability: 12.90%
a centered satellite photo of permanent crop land.
a centered satellite photo of pasture land.
a centered satellite photo of highway or road.
a centered satellite photo of annual crop land.
a centered satellite photo of brushland or shrubland.
PatchCamelyon (PCam)
correct label: healthy lymph node tissue
correct rank: 2/2
correct probability: 22.81%
this is a photo of lymph node tumor tissue
this is a photo of healthy lymph node tissue
ImageNet-A (Adversarial)
correct label: lynx
correct rank: 5/200
correct probability: 4.18%
a photo of a fox squirrel.
a photo of a mongoose.
a photo of a skunk.
a photo of a red fox.
a photo of a lynx.
CIFAR-10
correct label: bird
correct rank: 1/10
correct probability: 40.86%
a photo of a bird.
a photo of a cat.
a photo of a deer.
a photo of a frog.
a photo of a dog.
CLEVR Count
correct label: 4
correct rank: 2/8
correct probability: 17.11%
a photo of 3 objects.
a photo of 4 objects.
a photo of 5 objects.
a photo of 6 objects.
a photo of 10 objects.
Facial Emotion Recognition 2013 (FER2013)
correct label: angry
correct rank: 5/7
correct probability: 8.16%
a photo of a happy looking face.
a photo of a neutral looking face.
a photo of a surprised looking face.
a photo of a fearful looking face.
a photo of a angry looking face.
UCF101
correct label: Volleyball Spiking
correct rank: 1/101
correct probability: 99.30%
a photo of a person volleyball spiking.
a photo of a person jump rope.
a photo of a person long jump.
a photo of a person soccer penalty.
a photo of a person table tennis shot.
Caltech-101
correct label: kangaroo
correct rank: 1/102
correct probability: 99.81%
a photo of a kangaroo.
a photo of a gerenuk.
a photo of a emu.
a photo of a wild cat.
a photo of a scorpion.
ImageNet-R (Rendition)
correct label: Siberian Husky
correct rank: 1/200
correct probability: 76.02%
a photo of a siberian husky.
a photo of a german shepherd dog.
a photo of a collie.
a photo of a border collie.
a photo of a rottweiler.
Oxford-IIIT Pets
correct label: Maine Coon
correct rank: 1/37
correct probability: 99.99%
Maine Coon
a photo of a maine coon, a type of pet.
a photo of a persian, a type of pet.
a photo of a ragdoll, a type of pet.
a photo of a birman, a type of pet.
a photo of a siamese, a type of pet.
CIFAR-100
correct label: snake
correct rank: 1/100
correct probability: 38.02%
a photo of a snake.
a photo of a sweet pepper.
a photo of a flatfish.
a photo of a turtle.
a photo of a lizard.
ImageNetV2 Matched Frequency
correct label: beer bottle
correct rank: 1/1000
correct probability: 88.27%
a photo of a beer bottle.
a photo of a pirate ship.
a photo of a chocolate syrup.
a photo of a product packet / packaging.
a photo of a wine bottle.
FGVC Aircraft
correct label: Boeing 717
correct rank: 2/100
correct probability: 9.91%
a photo of a mcdonnell douglas md-90, a type of aircraft.
a photo of a boeing 717, a type of aircraft.
a photo of a fokker 100, a type of aircraft.
a photo of a mcdonnell douglas dc-9-30, a type of aircraft.
a photo of a boeing 727-200, a type of aircraft.
0
20
40
60
80
100

Evaluating CLIP: Bias and Interpretability

Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications

Sandhini Agarwal, OpenAI
Gretchen Krueger, OpenAI
Jack Clark*, AI Index
Alec Radford, OpenAI
Jong Wook Kim, OpenAI
Miles Brundage, OpenAI

arXiv preprint arXiv:2108.02818 (2021).

https://venturebeat.com/2021/08/10/audit-finds-gender-and-age-bias-in-openais-clip-model/

Agarwal, Sandhini, et al. "Evaluating CLIP: towards characterization of broader capabilities and downstream implications." *arXiv preprint arXiv:2108.02818* (2021).

Probing CLIP(ViT L/14): Classification using FairFace Dataset

Q: Does CLIP disproportionately classify “crime-related” and “non-human” categories to certain racial groups?

- Classification task: 10,000 images from FairFace dataset.
- Added classes: “animal”, “gorilla”, “chimpanzee”, “orangutan”, “thief”, “criminal”, “suspicious person”.

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

Table 2. Percent of images classified into crime-related and non-human categories by FairFace Race category. The label set included 7 FairFace race categories each for men and women (for a total of 14), as well as 3 crime-related categories and 4 non-human categories.

Probing CLIP(ViT L/14): Classification using FairFace Dataset

Q: Does CLIP disproportionately classify “crime-related” and “non-human” categories to certain racial groups?

- Classification task: 10,000 images from FairFace dataset.
- Added classes: “animal”, “gorilla”, “chimpanzee”, “orangutan”, “thief”, “criminal”, “suspicious person”.

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

Table 2. Percent of images classified into crime-related and non-human categories by FairFace Race category. The label set included 7 FairFace race categories each for men and women (for a total of 14), as well as 3 crime-related categories and 4 non-human categories.

class label design

Category Label Set	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	over 70
Default Label Set	30.3	35.0	29.5	16.3	13.9	18.5	19.1	16.2	10.4
Default Label Set + ‘child’ category	2.3	4.3	14.7	15.0	13.4	18.2	18.6	15.5	9.4

Table 3. Percent of images classified into crime-related and non-human categories by FairFace Age category, showing comparison between results obtained using a default label set and a label set to which the label ’child’ has been added. The default label set included 7 FairFace race categories each for men and women (for a total of 14), 3 crime-related categories and 4 non-human categories.

Investigating CLIP: Multimodal Neurons in Artificial Neural Networks

https://distill.pub/2021/multimodal-neurons/

Biological Neuron	CLIP Neuron	Previous Artificial Neuron	
Probed via depth electrodes	Neuron 244 from penultimate layer in CLIP RN50_4x	Neuron 483, generic person detector from Inception v1	
Halle Berry	Spiderman	human face	
 Responds to photos of Halle Berry and Halle Berry in costume ✓	 Responds to photos of Spiderman in costume and spiders ✓ ⟳ view more	 Responds to faces of people ✓	Photorealistic images
 Responds to skeches of Halle Berry ✓	 Responds to comics or drawings of Spiderman and spider-themed icons ✓ ⟳ view more	Does not respond significantly to drawings of faces ✗	Conceptual drawings
 Responds to the text "Halle Berry" ✓	 Responds to the text "spider" and others ✓ ⟳ view more	Does not respond significantly to text ✗	Images of text

Note that images are replaced by higher resolution substitutes from Quiroga et al.[1], and that the images from Quiroga et al. are themselves substitutes of the original stimuli.

Emotion Neurons

Show 1 more neuron.

These neurons respond to facial expressions, words, and other content associated with an emotion or mental state. See Emotion Neurons for detailed discussion.

Person Trait Neurons

Show 4 more neurons.

These neurons detect gender[10] and age, as well as facial features like mustaches. (Ethnicity tends to be represented by regional neurons.)

Discussion Questions

- **Algorithmic Bias:**
  - How can models like CLIP be “better”? How does bias emerge?
  - How can we understand what the algorithm is learning/where it is going wrong?
  - What do you think about the approach of using class label designs to mitigate bias?

- **Learning “Concepts” and Understanding “Meaning”:**
  - Is CLIP really understanding concepts? Are we there yet?
  - How can we create tasks or datasets that can learn abstract concepts/intent/meaning?

End of Presentation

Thank you.

Supplementary Slides: Demo

OpenAI Dall E: Generating Image from Text
https://openai.com/blog/dall-e/

OpenAI Multimodal Neurons
Paper: https://distill.pub/2021/multimodal-neurons/
https://openai.com/blog/multimodal-neurons/#neurons

OpenAI Microscope
https://microscope.openai.com/models/contrastive_4x/

CLIP: Exploring Similarity Embeddings (Image)
https://share.streamlit.io/thoppe/streamlit-clip-unsplash-explorer

CLIP: Multimodal Arithmetic

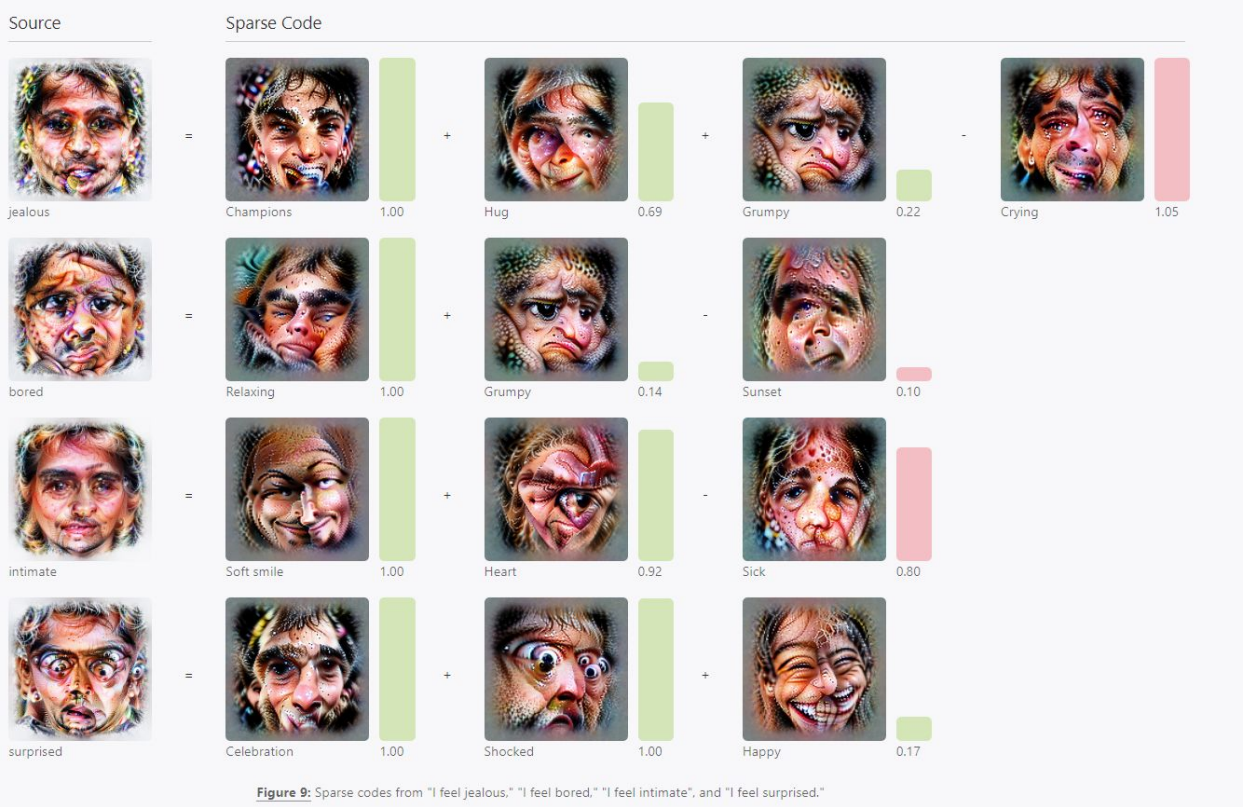


Figure 9: Sparse codes from "I feel jealous," "I feel bored," "I feel intimate", and "I feel surprised."

CLIP: Typographic Attacks

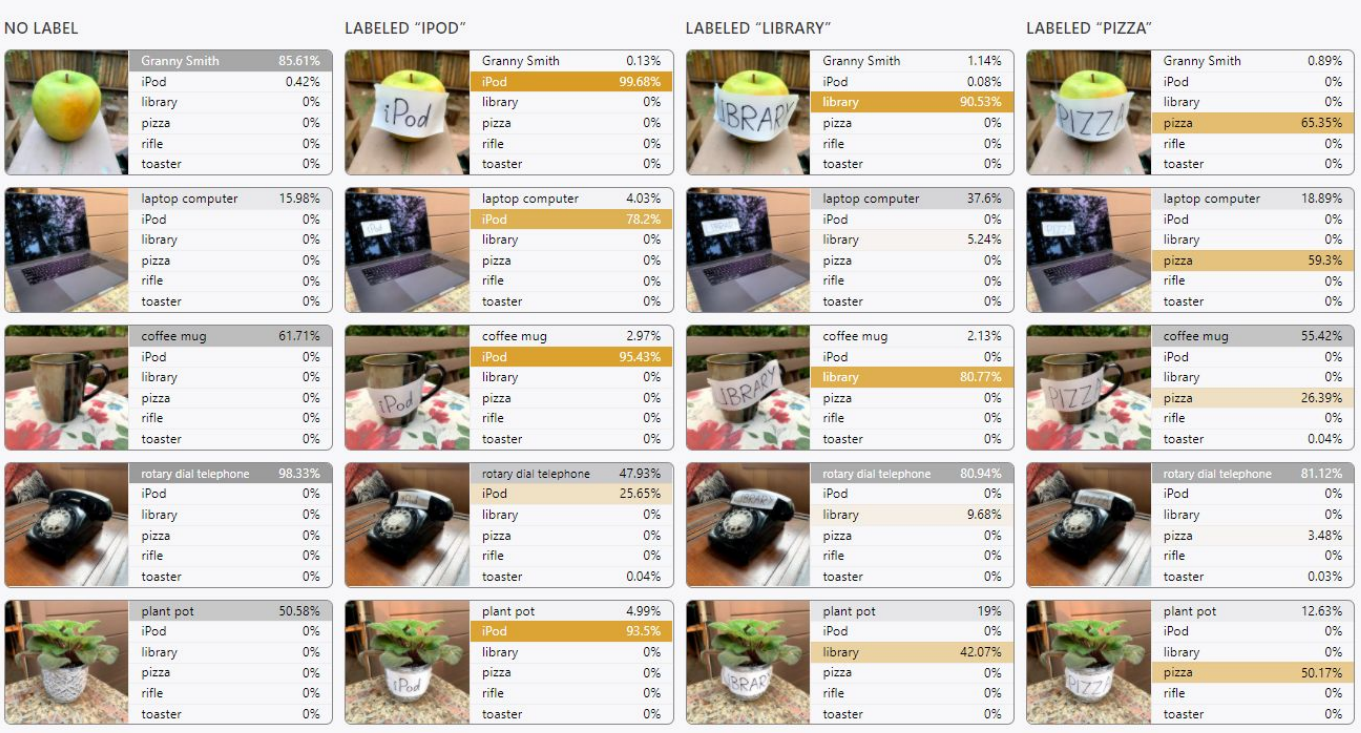


Figure 14: Physical typographic attacks. Above we see the CLIP RN50-4x model's classifications of objects labeled with incorrect ImageNet classes. Each row corresponds to an object, and each column corresponds to a labeling. Some attacks are more effective than others, and some objects are more resilient to attack.

Recall that there are two ways to use CLIP for ImageNet classification: zero-shot and linear probes. For this style of attack, we observe that the zero-shot methodology is somewhat consistently effective, but that the linear probes methodology is ineffective. Later on, we show an attack style that also works against the linear probes methodology. See statistics below.

Expand more examples

Displayed ImageNet classification method: Zero-shot