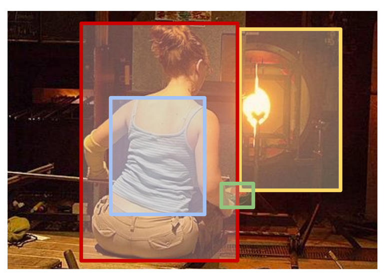
World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models

Ziqiao Ma*, Jiayi Pan*, Joyce Chai.
Computer Science and Engineering Division, University of Michigan.

Motivation

The ability to connect language units to their referents in the physical world, referred to as **grounding**, plays an important role in acquiring and understanding the meanings of words. Grounding has enabled humans to bootstrap new word learning with only minimal information, known as **fast mapping**[1].



`A lady wearing a navy blue stripe tank top is getting ready to burn glass in front of an incinerator.`

Despite the exciting performance of pre-trained vision-language models (VLMs) on downstream tasks, it remains unclear whether these models can truly understand or produce words with their grounded meanings in the perceived world, and how grounding may further bootstrap new word learning.

Grounded Open Vocabulary Acquisition

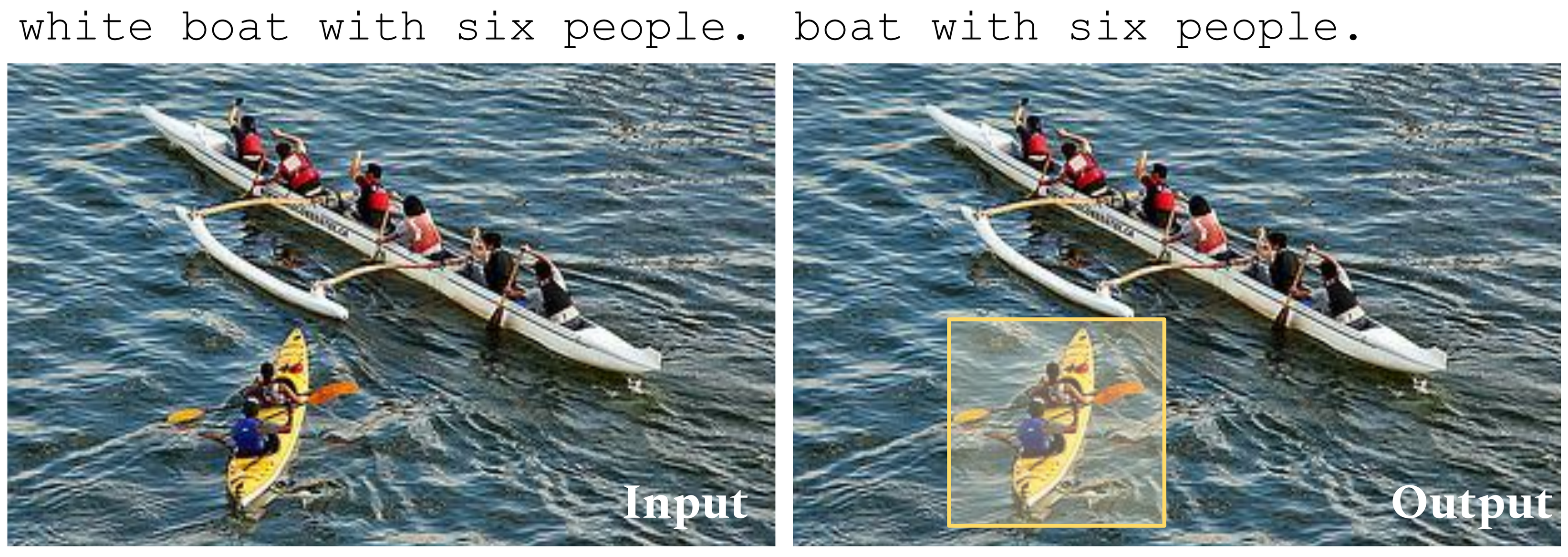
We evaluate **grounded language acquisition** through both *language modeling* and *object localization* tasks.

- Use the log pseudo-perplexity to evaluate language modeling for each word w: $\log \mathrm{PPL}(w) = -\log P(w \mid x_{\mathrm{img}}, x_{\mathrm{cap}})$.
- Use the intersection-over-union (IoU) for object localization. With n ground truth boxes $B = \{b_i\}$ and m predicted boxes $\widetilde{B} = \{\widetilde{b_j}\}$: $\mathrm{IoU}_{\mathrm{any}} = \frac{1}{n}\sum_i \max_j \mathrm{IoU}(b_i, \widetilde{b_j})$ and $\mathrm{IoU}_{\mathrm{all}} = \mathrm{IoU}(\cup B, \cup \widetilde{B})$.
- Use **grounded perplexity** (G-PPL) for cross-modal evaluation:

$$\log \text{G-PPL}(w) = \begin{cases} \infty & \text{if IoU} = 0 \\ \log \mathrm{PPL}(w) - \log \mathrm{IoU} & \text{else} \end{cases}$$

`Two boats of people, a smaller yellow <mask> with two people and a larger white boat with six people.`

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

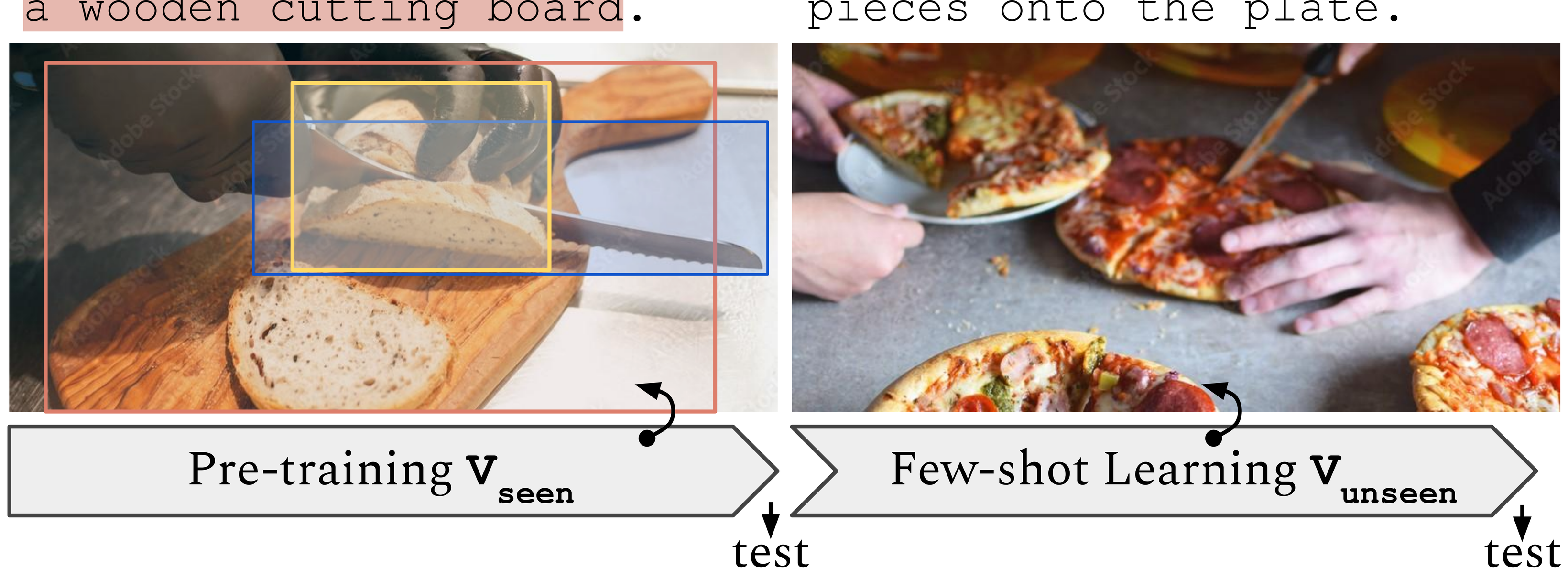


We introduce **few-shot new word learning**:

- Motivation: The costly grounding annotation can hardly cover the vocabulary during pre-training. Models should acquire grounded new words in a few shots without explicit mappings.
- Setup: the model first pre-trains on a grounding dataset with base words $\mathcal{V}_{\mathrm{seen}}$, and then acquires unseen words $\mathcal{V}_{\mathrm{unseen}}$ from a few shots of raw text-image pairs.

`Someone is slicing a loaf of bread using a knife on a wooden cutting board.`

`I am slicing the pizza with a knife and stacking the pieces onto the plate.`



We build our dataset based on the Flickr30K Entities with dense annotations between groundable phrases and bounding boxes of objects. 60 seen words and 31 unseen words are chosen.

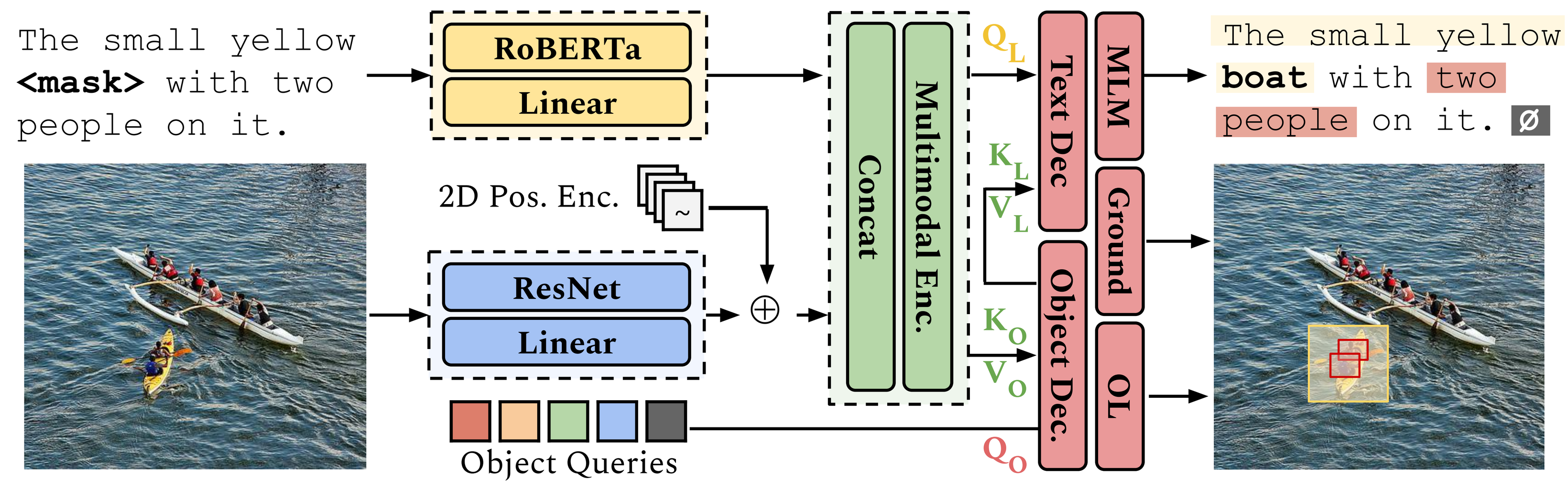
References

[1] Susan Carey and Elsa Bartlett. 1978. Acquiring a single new word. Papers and Reports on Child Language Development, 15:17–29.

Links

Computational Model

We introduce Object-oriented BERT (`OctoBERT`), a dual-stream VLM. The object decoder produces an object embedding for each learnable object query and we perform language modeling explicitly on the representations of the perceived objects.



As a visually grounded language model, `OctoBERT` is pre-trained with three objectives: masked language modeling (MLM), object localization (OL), and grounding by word-region alignment.

Bootstrap Grounded Pre-training

`OctoBERT` shows strong performance in terms of both grounded metrics, significantly outperforming the groundless baseline $\texttt{OctoBERT}_{\mathrm{w/o\,G}}$ and pre-trained baselines, even for systems pre-trained with significantly more data and computation.

Metrics	G-HR@1	log G-PPL	HR@1	log PPL	Acc@0.5	IoU
Models	Seen					
`ViLT+MDETR`	19.8 / 19.3	2.53 / 2.43	64.7	1.27	31.1 / 30.4	28.5 / 31.2
`VisualBERT` (FT)	28.5 / -	2.96 / -	42.3	2.33	**68.1** / -	53.3 / -
$\texttt{OctoBERT}_{\mathrm{w/o\,G}}$ (FT)	28.9 / 27.8	2.33 / 2.38	63.9	1.41	44.0 / 43.0	40.0 / 38.2
`OctoBERT`	**47.0** / **46.3**	**1.79** / **1.81**	**66.9**	**1.26**	**66.8** / **66.3**	**58.8** / **57.6**
Models	Unseen					
$\texttt{OctoBERT}_{\mathrm{w/o\,G}}$ (FT)	1.1 / 1.1	11.89 / 12.04	3.7	10.87	38.7 / 31.9	36.2 / 31.0
`OctoBERT`	2.3 / 2.3	11.58 / 11.74	4.2	11.01	**61.3** / **53.1**	**56.3** / **48.0**

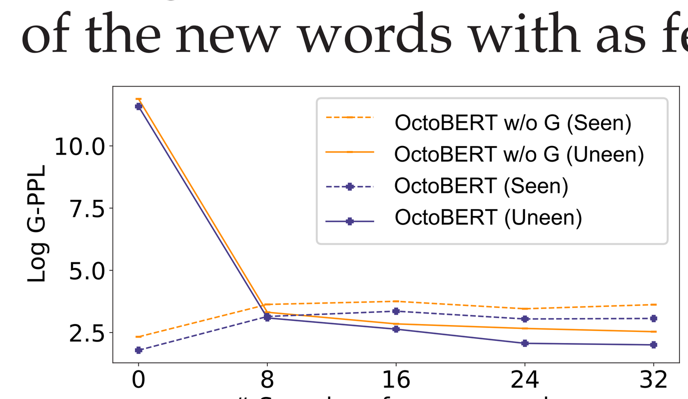
`OctoBERT` has a surprising performance in localizing unseen words behind the `MASK`s. This performance disparity in language modeling and localization on unseen words suggests the ability of **word-agnostic grounding**: to locate the most likely referent of a word through both the linguistic context and the visual context, even if the word itself is never seen during pre-training.

`Three men seated on a <MASK> in a small village.`

- W2W Prediction: **`animal`**
- Unseen Ground Truth: **`elephant`**

Few-Shot New Word Acquisition

We explore the multi-class and single-class incremental learning settings. `OctoBERT` is able to quickly acquire grounded meanings of the new words with as few as 8 examples.



# Samples	log G-PPL (`pizza`) w/ G	log G-PPL (`pizza`) w/o G	log G-PPL (`circular`) w/ G	log G-PPL (`circular`) w/o G
0	10.70	9.59	15.21	15.12
8	**1.47**	2.21	**1.59**	2.25
16	**1.07**	2.54	**1.07**	2.25
24	**1.19**	1.25	**1.55**	1.81
32	**0.90**	1.18	**1.23**	1.61

Predictors of Performance

- A strong correlation between frequency and perplexity, indicating that `OctoBERT` still heavily relies on distributional statistics.
- Visually salient and less perceptually ambiguous are easier to localize and acquire, consistent with human learners.
- A misalignment between the human perceived familiarity of words and the machine's perplexities, *i.e.*, the more familiar humans are with a word, the more perplexed models get.
- Aligns well with human intuition for imageability but not concreteness, indicating the lack of physical interaction.