Linearly Mapping from Image to Text Space

Jack Merullo, Louis Castricato, Carsten Eickhoff, Ellie Pavlick (Brown Univ.)

ICLR 2023

https://arxiv.org/abs/2209.15162

09. 14. 2023
EffL @POSTECH
Hagyeong Lee

Problem of Language Model

To appear at ACL 2020

Climbing towards NLU:
On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

Emily M. Bender and Alexander Koller., "Climbing towards NLU: on meaning form and understanding in the age of data", ACL 2020

A System exposed only to form in its training cannot in principle learn meaning

Form & Meaning in Language

Form

- Anything we can find in a language (e.g., symbols, mouth movements)

Meaning

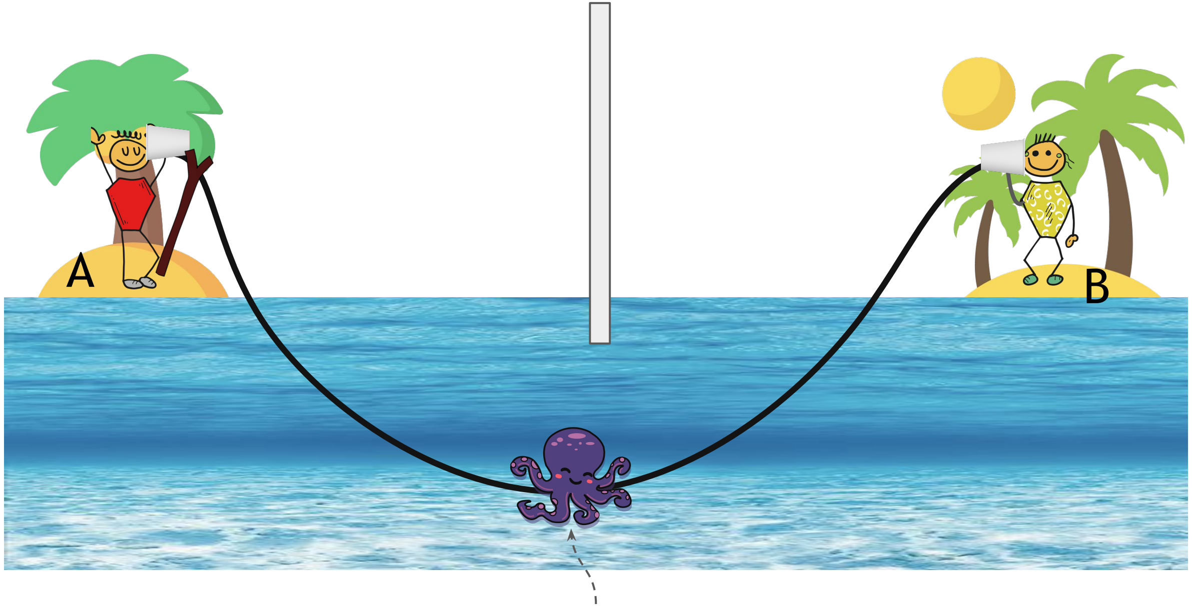
- Relationship between form and non-linguistic parts
- Including Communicative intent

Is **form** alone **meaningful?**

⇒ Octopus Thought exp.

Emily M. Bender and Alexander Koller., “Climbing towards NLU: on meaning form and understanding in the age of data”, ACL 2020

Octopus Thought Experiment

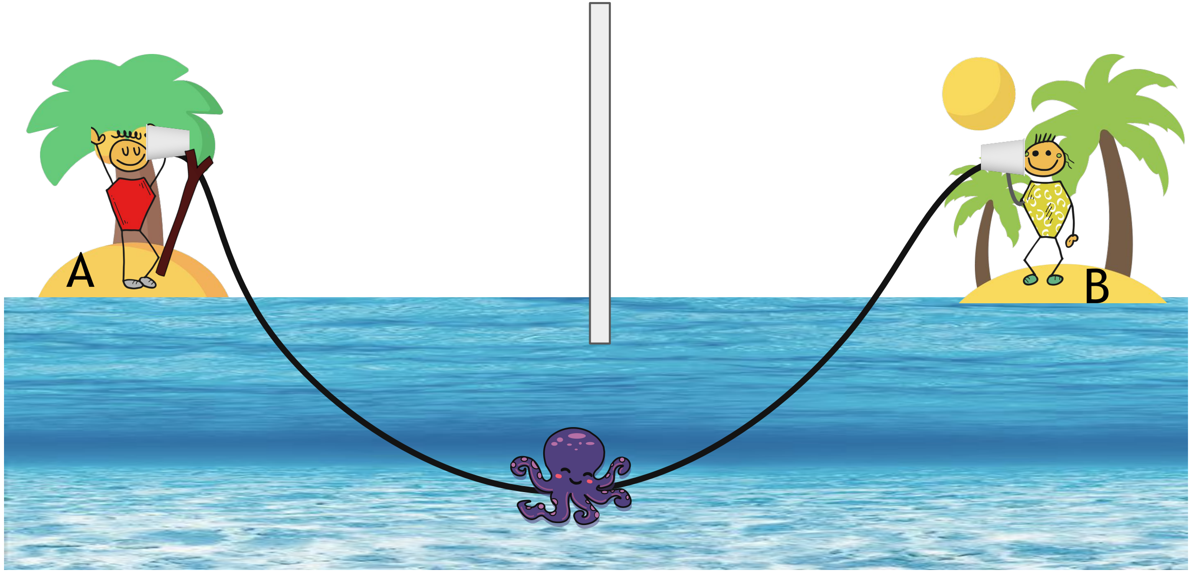


A highly intelligent octopus that knows nothing about Human language

- Excellent at spotting *statistical* patterns

Emily M. Bender and Alexander Koller., “Climbing towards NLU: on meaning form and understanding in the age of data”, ACL 2020

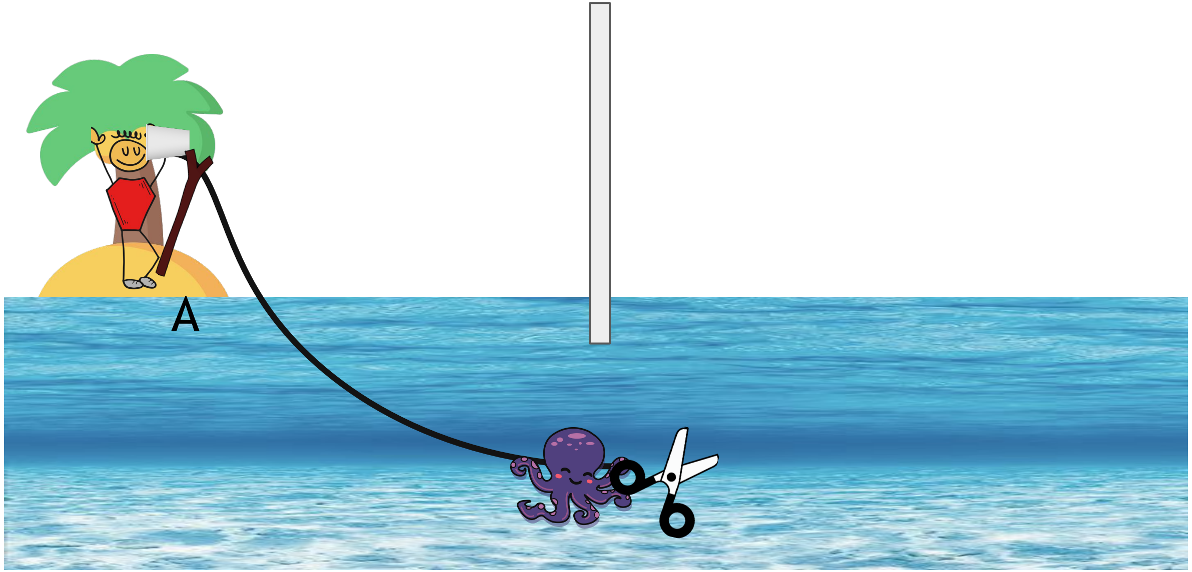
Octopus Thought Experiment



- Observed the use of certain words in similar **forms**
- Maybe noticed a common lexical pattern

Emily M. Bender and Alexander Koller., “Climbing towards NLU: on meaning form and understanding in the age of data”, ACL 2020

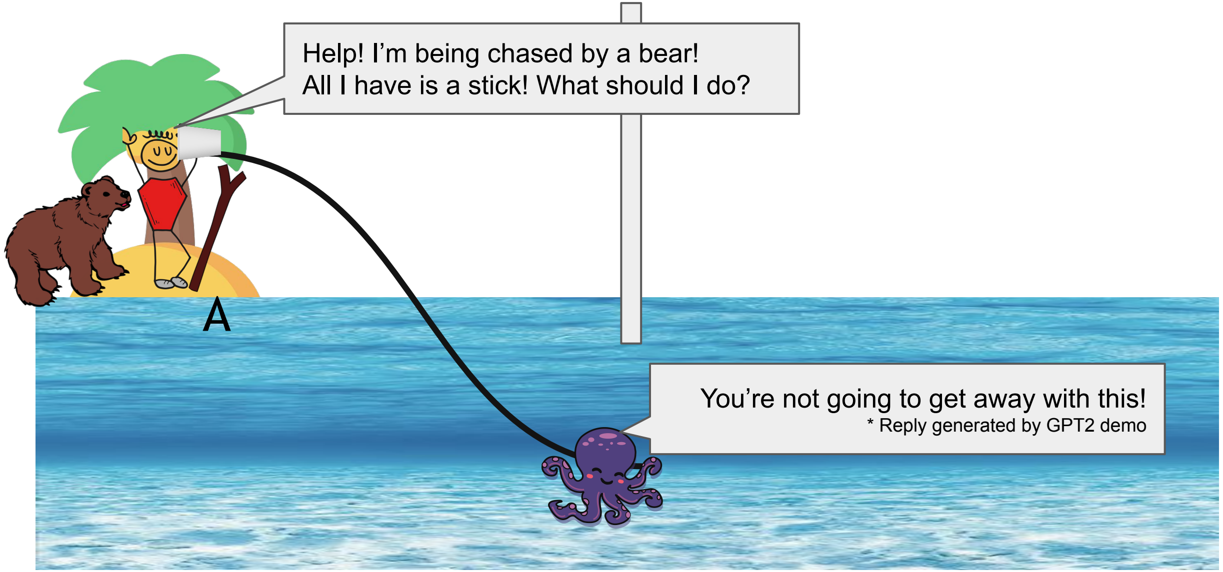
Octopus Thought Experiment



starts impersonating B and replying to A

Emily M. Bender and Alexander Koller., “Climbing towards NLU: on meaning form and understanding in the age of data”, ACL 2020

Octopus Thought Experiment



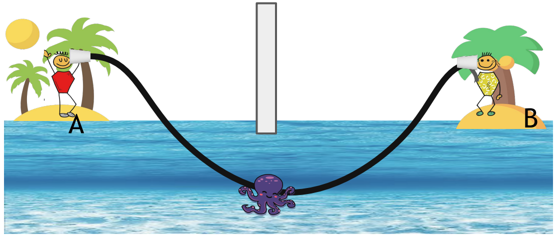
The octopus doesn't know the referents of the words

- no idea what bears or sticks are

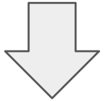
⇒ Octopus = LM

Emily M. Bender and Alexander Koller., "Climbing towards NLU: on meaning form and understanding in the age of data", ACL 2020

Octopus Thought Experiment - Conclusion



- LMs do not tend to learn conceptual representations (meanings) of language.
  - Humans acquire language not only through the **form** (representation) but also through the **interaction** of various factors in physical world.



How well can a text-only language model learn aspects of the physical world?

Previous Works

- Show success in mapping images to language model soft prompts as a method for multimodal pre-training (e.g., *MAGMA*, *Frozen*)

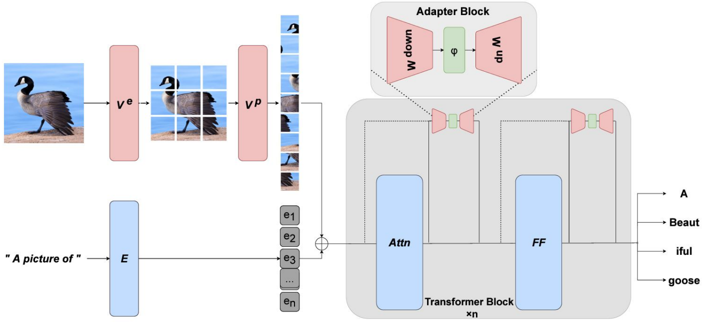


Figure 2: MAGMA's architecture. The layers in red are trained, and the layers in blue remain frozen.

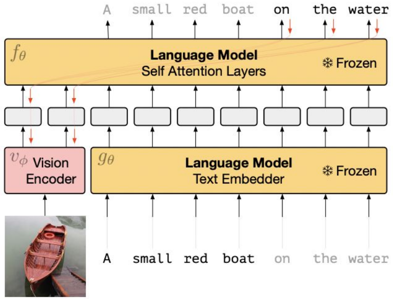


Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

⇒ However, no attempts to restrict the mechanism behind this mapping and understand how it works.

Constantin Eichenberg et al., "MAGMA-Multi modal Augmentation of Generative Models through Adapter-based Finetuning", EMNLP 2022
Maria Tsimpoukelli et al., "Multimodal Few-Shot Learning with Frozen Language Models", NeurIPS 2021

Language & Image representation

- **Hypothesis.**

  Conceptual representations (between language and image embeddings) can be approximately <u>mapped to one</u> through <u>a linear transformation</u>

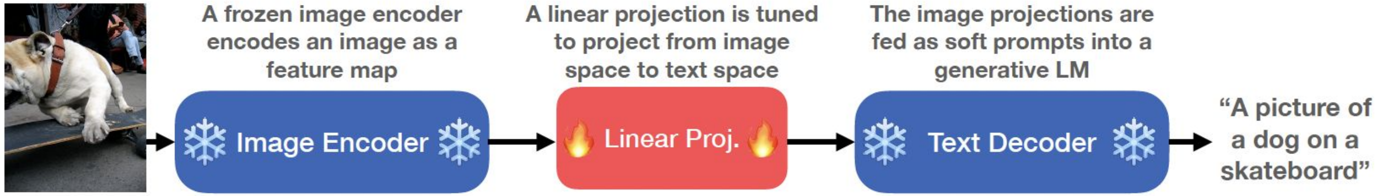
  - Why train on linear transformation?
    - because of the simplicity !

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Train linear projections from image representations into the text space of a language model to produce image-to-text tasks

= transform an image representation into “soft prompts”
(do not correspond to discrete language tokens)



Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Linear projection layer
  - To project from h_I (hidden size of a pre-trained image encoder) to text space e_L (text embedding size of the LM)
- Pre-trained Language Model
  - GPT-J model (open source weights of 6B param.)

Constantin Eichenberg et al., "MAGMA–Multi modal Augmentation of Generative Models through Adapter-based Finetuning", EMNLP 2022

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Image Encoders : Different *E*s
  - To determine the consistency between encodings from *E* and LM
  - Choice of *E*
    - the degree of linguistic supervision (saw in pre-training)

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Image Encoders : Different *E*s

Degree of accessibility to linguistic labeled data	Image Encoder *E*	
Strong	CLIP RN50x16	Trained to learn multi-modal image-text embeddings
Weak	NFRN50	Trained on an image classification (on labeled WordNet hyper/hyponym) e.g., hyper: Vehicle, hypo: car, train, bus
None	BEIT-Large	Trained using a self-sup.masked visual token (on ImageNet)

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Image Encoders : Different *E*s

Degree of accessibility about linguistic labeled data	Encoder *E*
Strong	CLIP RN50x16
Weak	NFRN50
None	BEIT-Large

1) MAGMA_released 🔥

→ using MAGMA's adapter (not linear projection)

2) MAGMA _ours 🔥

→ using linear projection

3) **CLIP** ❄️

*** 🔥: Update the visual encoder (and LM both; MAGMA)**
/ ❄️: Freeze the visual encoder and LM (Released pre-trained model)

Constantin Eichenberg et al., "MAGMA–Multi modal Augmentation of Generative Models through Adapter-based Finetuning", EMNLP 2022

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Image Encoders : Different *E*s

Degree of accessibility about linguistic labeled data	Encoder *E*
Strong	CLIP RN50x16
Weak	NFRN50
None	BEIT-Large

1) Pre-trained
(for image classification on the WordNet)
: NFRN50 ❄️

2) Fine-tuning
(**update the pre-trained image encoder**)
: NFRN50 Tuned 🔥
→ *Frozen* model

3) Randomly initialized
: NFRN50 Random ❄️

* 🔥: **Update the visual encoder**
/ ❄️: **Freeze the visual encoder and LM (Released pre-trained model)**

Maria Tsimpoukelli et al., “Multimodal Few-Shot Learning with Frozen Language Models”, NeurIPS 2021

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Image Encoders : Different *E*s

Degree of accessibility about linguistic labeled data	Encoder *E*
Strong	CLIP RN50x16
Weak	NFRN50
None	BEIT-Large

1) Pre-trained (for masked visual on the ImageNet 22K) & Fine-tuning : BEIT FT ❄️

2) Randomly initialized : BEIT Random ❄️

* 🔥: **Update the visual encoder**
/ ❄️: **Freeze the visual encoder and LM (Released pre-trained model)**

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Training procedure
  - Mapping Linear Projection layer
  - Dataset : CC3M (Conceptual Captions 3M)

Method

LiMBeR (Linearly Mapping Between Representation spaces)

- Evaluation
  - Task : Image captioning / Visual Question Answering
  - Datasets : MS-COCO, NoCaps / VQA2
  - Captioning Metrics
    - CIDEr-D
      - rewards generating accurate words which are more likely to be visually informative
    - CLIPScore / Ref-CLIPScore
      - evaluate similarity between image and caption without/with references

Experiments : Captioning

Image Captioning

CLIP	a giraffe in the lobby of the building
NFRN50	the giraffe in the zoo.
BEIT	a peacock in the garden
NFRN50 Random	a man and a woman in a field of flowers

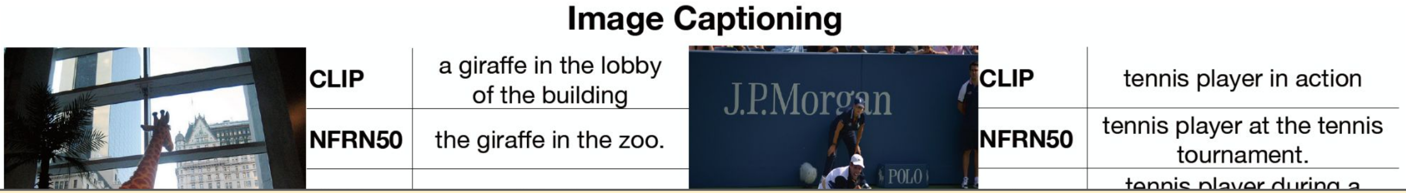
CLIP	tennis player in action
NFRN50	tennis player at the tennis tournament.
BEIT	tennis player during a tennis match.
NFRN50 Random	the new logo for the team

*** 🔥: Update the visual encoder (and LM both; MAGMA) / ❄️: Freeze the visual encoder and LM**

	Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		**CoCo**	CoCo	
		In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
Jointly-tuned	🔥NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
	🔥MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
	🔥MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
Just training the projection layer	❄️BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
	❄️NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
	❄️BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
	❄️NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
	❄️BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
	❄️CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

no linguistic supervision transfers well to the LM for captioning

Experiments : Captioning



There is in fact a relationship between
the **linguistic supervision** of the pre-training task and perf. on transferring to the LM !

	Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		CoCo	CoCo	
		In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
Jointly-tuned	🔥NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
	🔥MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
	🔥MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
Just training the projection layer	❄BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
	❄NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
	❄BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
	❄NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
	❄BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
	❄CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

no linguistic supervision transfers well to the LM for captioning

Experiments : VQA (Visual Question Answering)

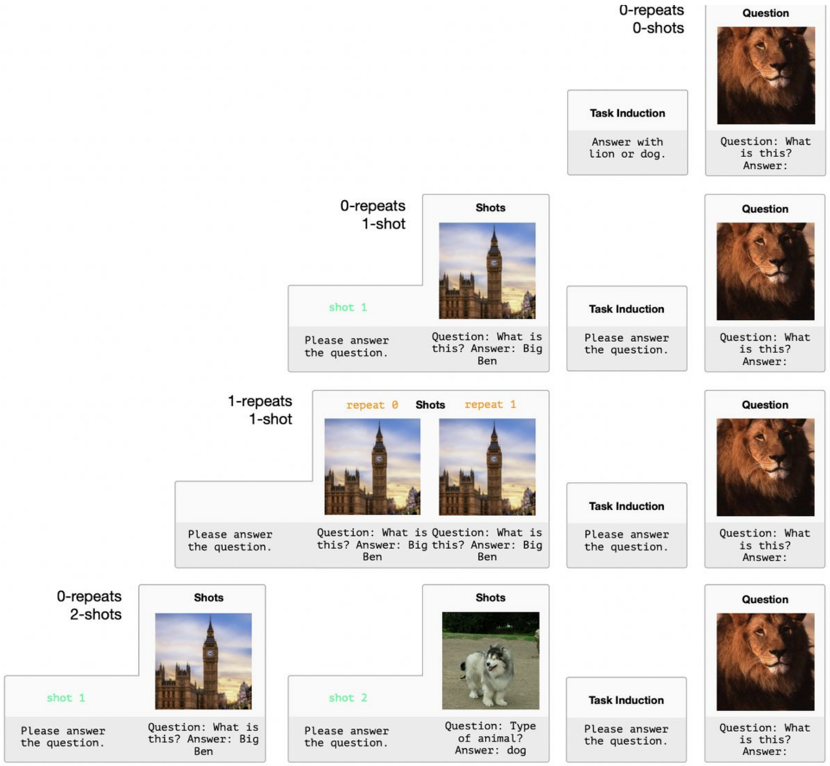


Figure 5: Examples of few-shot learning vocabulary.

Maria Tsimpoukelli et al., “Multimodal Few-Shot Learning with Frozen Language Models”, NeurIPS 2021

Experiments : VQA (Visual Question Answering)

Visual Question Answering

CLIP	He is surfing a wave.
NFRN50	He is surfing the waves.
BEIT	He is jumping into the water.
NFRN50 Random	He is swimming in the pool.

Q: What is the person doing?
A: `surfing`

CLIP	A tennis racket
NFRN50	A tennis racket
BEIT	A baseball bat.
NFRN50 Random	A tree

Q: What is the person holding?
A: `tennis racket`

VQA n-shots	0	1	**2**	4
Blind	20.60	35.11	36.17	36.99
🔥NFRN50 Tuned	27.15	37.47	38.48	39.18
🔥MAGMA (ours)	24.62	39.27	40.58	41.51
🔥MAGMA (reported)	32.7	40.2	**42.5**	43.8
❄️NFRN50 Random	25.34	36.15	36.79	37.43
❄️BEIT	24.92	34.35	34.70	31.72
❄️NFRN50	27.63	37.51	38.58	39.17
❄️CLIP	33.33	39.93	**40.82**	40.34

BEIT does not encode visual info. that corresponds to lexical categories

Experiments : Visual Concepts

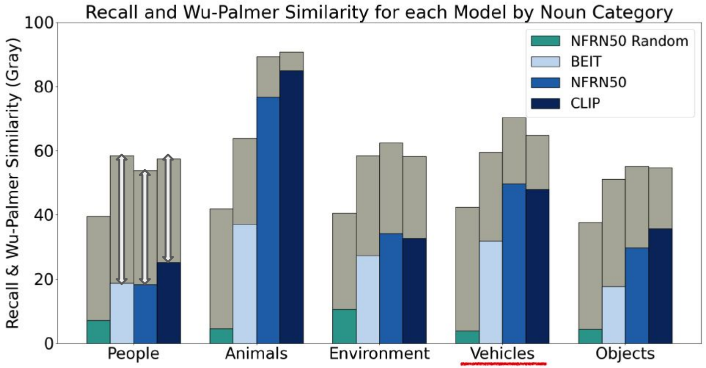
Why BEIT prompts perform so poorly for VQA despite performing decently for captioning?

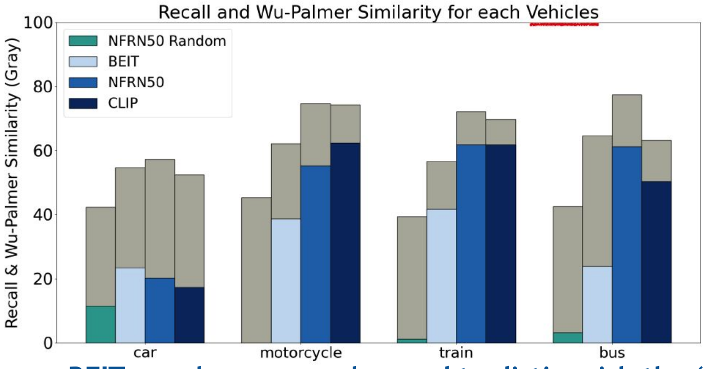
- **Hypothesis.** BEIT does not encode visual info. that corresponds to lexical categories
- Metrics
  - Wu-Palmer similarity
    - Calculate the distance between the GT and the generated word in the WordNet taxonomy
    - Measure **how close** a word was to the correct answer

Experiments : Visual Concepts

Why BEIT prompts perform so poorly for VQA despite performing decently for captioning?

- (On average) Recall (blue and green bar) score follows : CLIP > NFRN50 > BEIT >> Random
- However, judging by Wu-Palmer similarity (gray bar): **BEIT** performs nearly the same or better than NFRN50 and CLIP on 4/5 of the noun categories.
    - **BEIT does not learn conceptual differences between two similar looking objects that use different words. ⇒ transferring a related one based on visual similarity**





BEIT may have never learned to distinguish the 'bus' concept

Conclusion

- Show the linguistic supervision of the vision model pretraining objective correlates with the degree of similarity
    - Verified a hypothesis : training only a linear layer is enough for mapping visual pre-trained knowledge to text space.
    - And it can enable downstream tasks (such as few/zero-shot VQA, image captioning) utilizing stored knowledge from both worlds

- Future work (or Question)
    - Could it be improved by considering different model sizes ?
      (e.g. larger or smaller CLIP models or supervised resnets or BEITs)
        - whether the probing results get better or worse with image encoder size

Q & A

Thank you :)