Advancing Multimodal Vision-Language Learning / Faire Progresser L'apprentissage Multimodal de la Vision et du Langage

Aishwarya Agrawal

Assistant Professor @ UdeM and Mila

Research Scientist @ DeepMind (one day a week)

Mila

Multimodal AI Research (MAIR) Lab

Oscar Mañas

Saba Ahmadi

Le Zhang

Sarvjeet Singh Ghotra

Rabiul Awal

Joining in Fall this year:

Kanishk Jain

Qian Yang

P Shravan Nayak

Vision-Language Tasks

Image Captioning

"A group of young people playing a game of Frisbee."

Visual Question Answering

Q: *"What is the mustache made of?"*

A: *"bananas"*

Vision-Language Tasks

Image Retrieval

"Grey haired man in black and yellow tie."

Image Generation

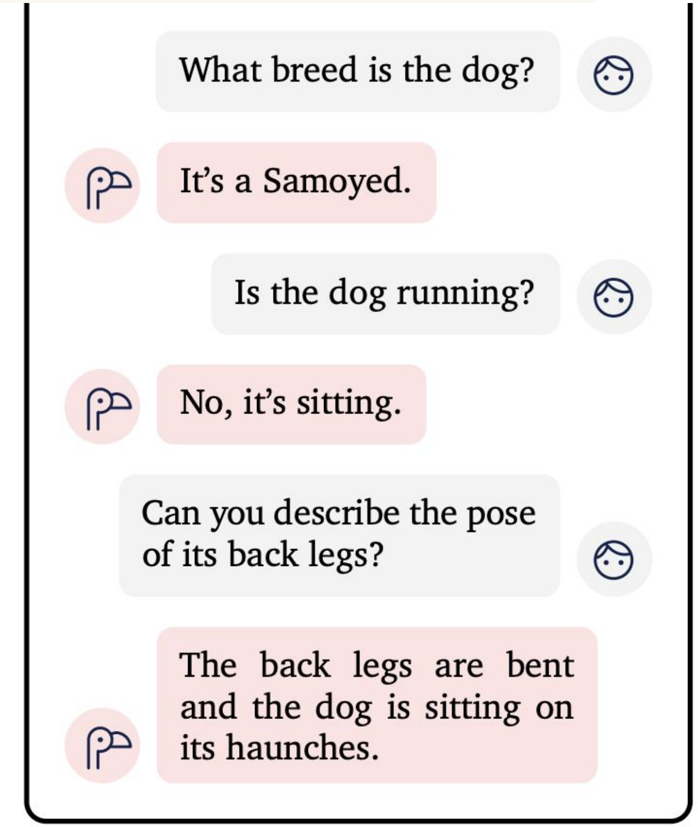
"Grey haired man in black and yellow tie."

Why vision and language?

- Intuitive:
  - Humans learn in multimodal settings
- Applications:
  - Aid to visually impaired users
  - Online shopping and organizing photos
  - Grounded virtual assistants
- Scientific:
  - Visual recognition
  - Language understanding
  - Combining information across modalities
  - Visio-linguistic compositional reasoning
  - Commonsense and factual knowledge reasoning

Current State of Vision-Language Research

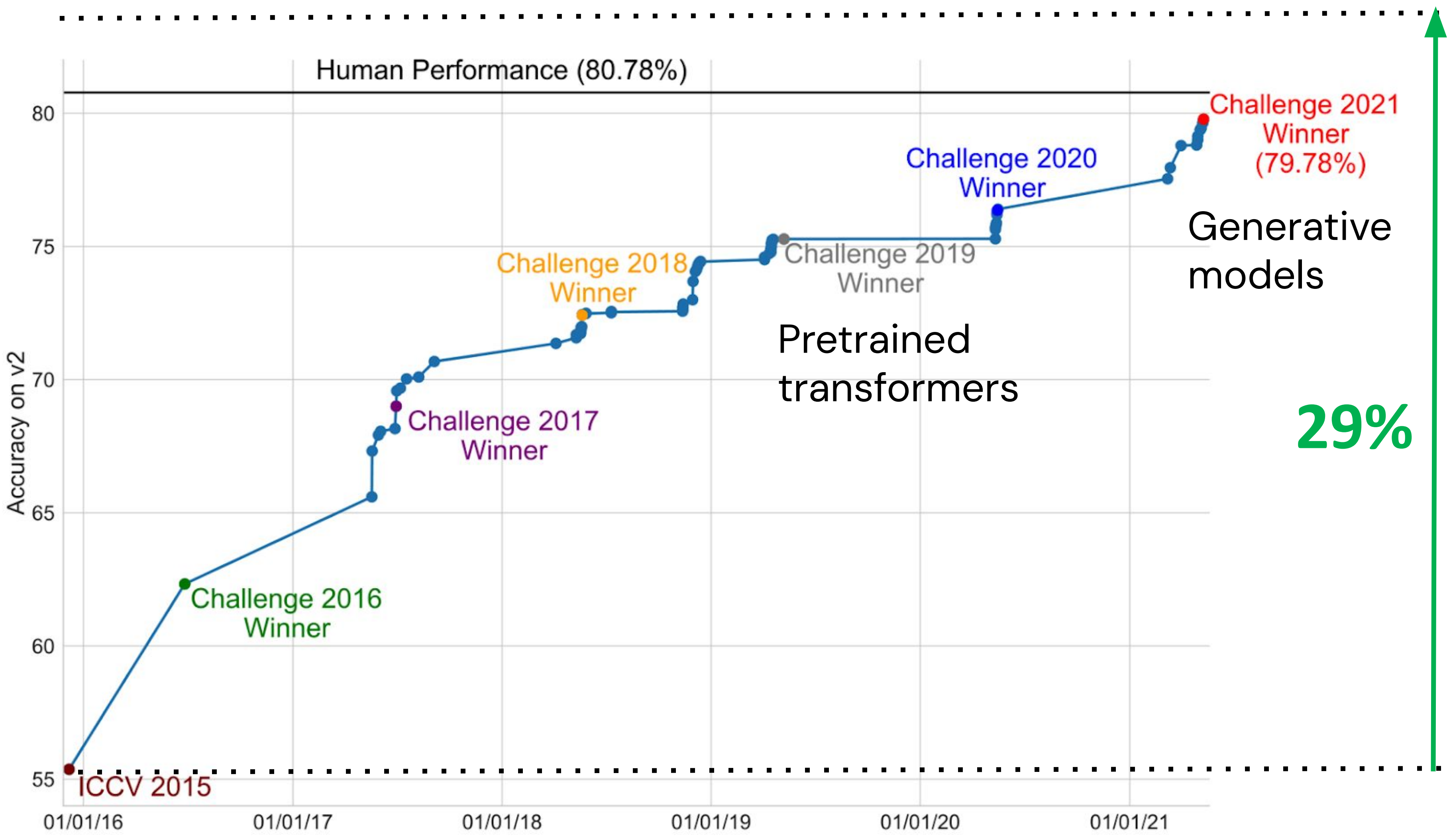
DeepMind's Flamingo
Link



Vision-Language Progress

PaLI: 84.3% [Chen et al., 2022]

Flamingo: 82.1%
[Alayrac et al., 2022]



Vision-Language Challenges

- Out-of-distribution generalization

DeepMind

Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization


Aishwarya Agrawal[*,$]

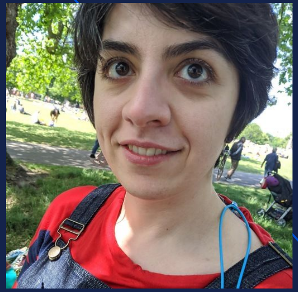

Ivana Kajić[*]


Emanuele Bugliarello[*]


Elnaz Davoodi[^]


Anita Gergely[^]


Phil Blunsom


Aida Nematzadeh[*,$]

EACL 2023

*, ^ denote equal contribution, $ denotes equal senior contribution

Experimental Setup
Datasets

VQAv2
(Goyal et al., 2017)

Q: What is the color of the hydrant?

A1: orange
A2: yellow
A3: orange
[...]

VG
(Krishna et al., 2017)

Q: What are these zebras doing?

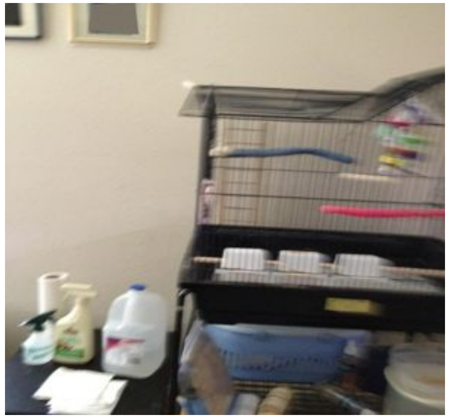
A: Eating

GQA
(Hudson and Manning, 2019)

Q: What is the large container made of?

A: cardboard

VizWiz
(Gurari et al., 2018)



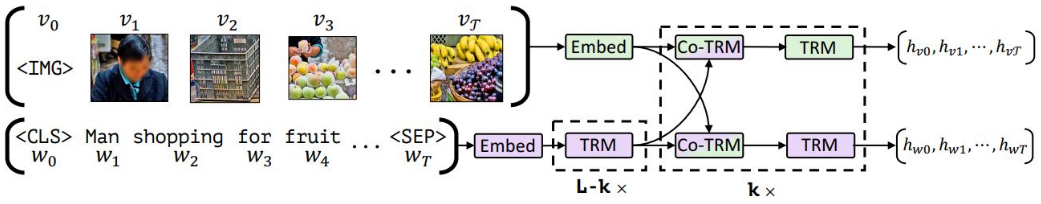
Q: Please fully describe what you see in this image, thank you.

A1: bird cage bottles paper towels
A2: birdcage cleaning supplies
A3: unanswerable
[...]

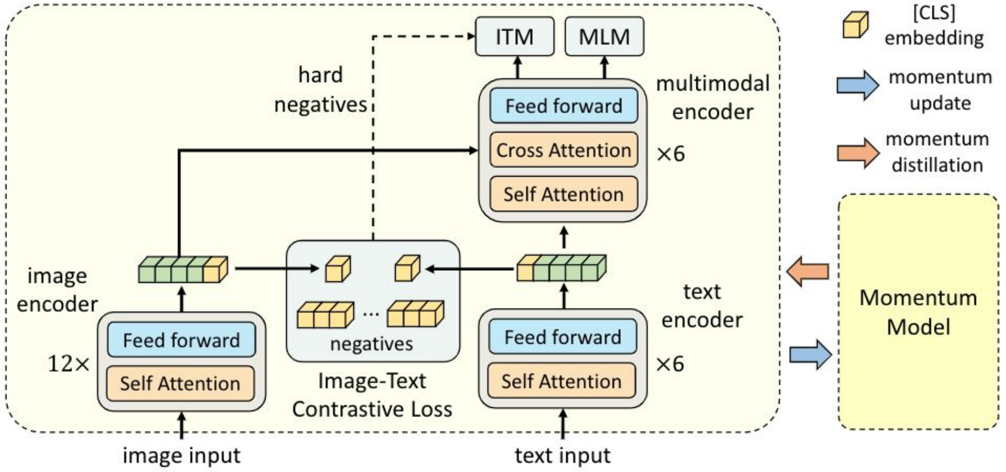
Slide credits: Ivana Kajić

Experimental Setup
Models

- Two representative, widely-used pretrained models achieving strong performance in V&L tasks:



ViLBERT (Lu et al., 2019)

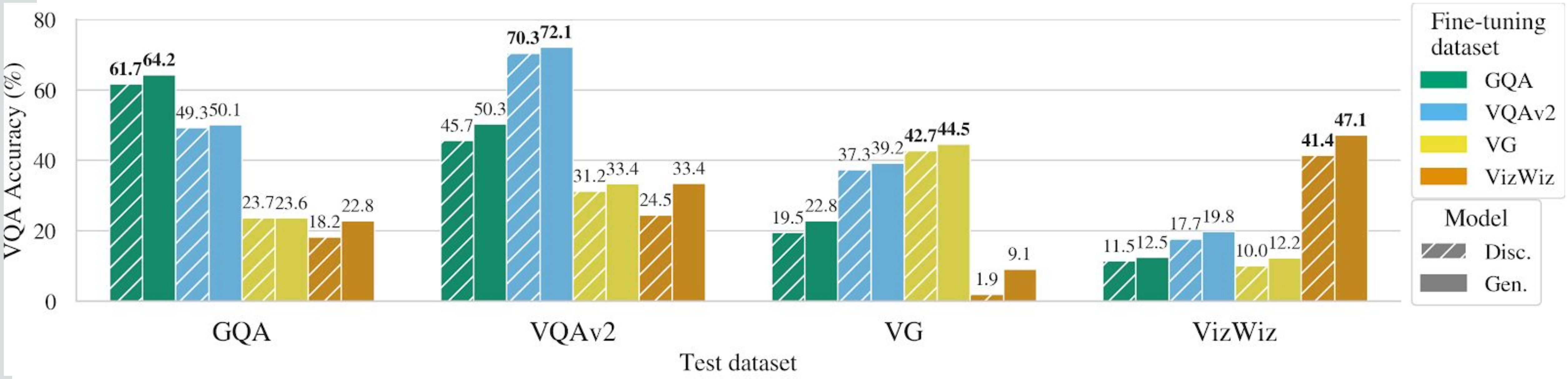


ALBEF (Li et al., 2021)

- Total: 128 experiments

Slide credits: Ivana Kajić

IID vs OOD performance



- Large drop in performance for OOD evaluation

Slide credits: Ivana Kajić

Potential factors causing poor OOD generalization: A qualitative analysis

- Poor reasoning skills (logical, spatial, compositional)

  E.g., *"Is the cheese to the right or to the left of the empty plate?"*

- Overfitting to answer priors

  E.g., *"What is the skateboarder wearing to protect his head?"* → *"helmet"*

- Overfitting to question format

  E.g., *"What animal … ?", "What kind of animal … ?"* (GQA)

  ⇩ 45% accuracy drop

  "Who is … ?", "What is … ?" (VG)

Vision-Language Challenges

- Out-of-distribution generalization
- Data-efficient adaptation to new tasks

Data-efficient adaptation to new tasks

- If a model can *caption images* (VL task-1), can we adapt it to *answer questions about images* (VL task-2) with *few examples*?

- Can we use few-shot capabilities of *pre-trained language models* such as GPT-3?

EACL 2023

MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting


Oscar Mañas


Pau Rodriguez*


Saba Ahmadi*

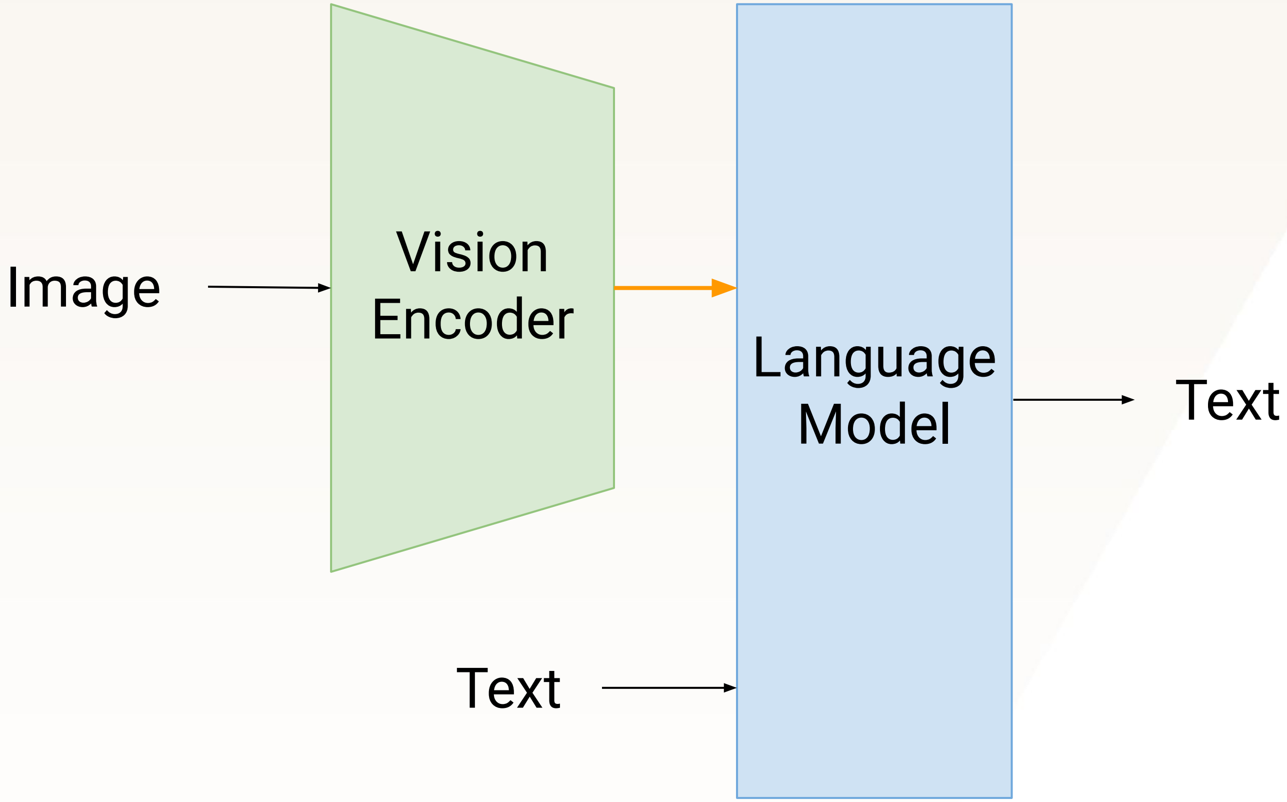

Aida Nematzadeh


Yash Goyal


Aishwarya Agrawal

*denotes equal contribution

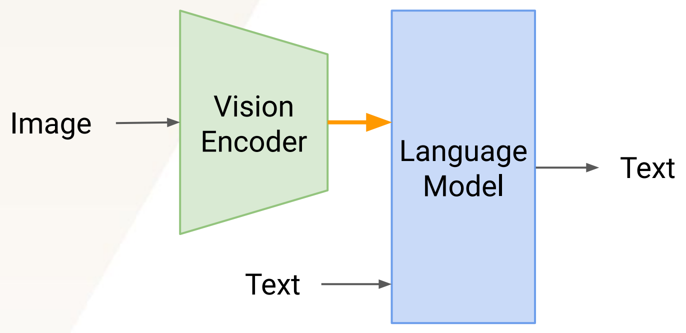
Goal: Adapt pre-trained language model for visual inputs



Slide adapted from Oscar Mañas

What existing approaches do

- Finetune the entire language model [Dai et al. 2022, Hao et al. 2022]
- Insert and train adapter layers in the language model [Eichenberg et al. 2021, Alayrac et al. 2022]
- Learn vision encoder from scratch [Tsimpoukelli et al. 2021]



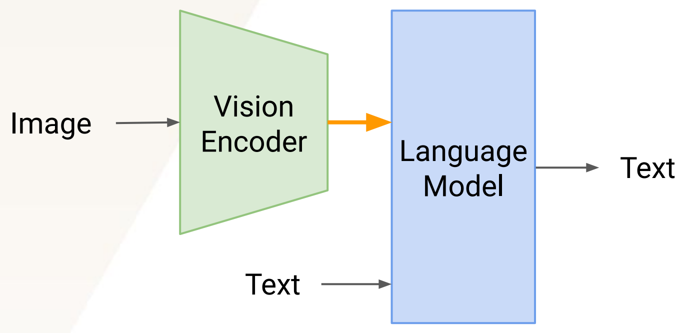
Issues with existing approaches:

- Large number of trainable parameters (~40M to ~10B)
- Inserting adapter layers is not straightforward
- Learning vision encoder from scratch does not scale well with larger vision encoders

Figure credits: Oscar Mañas

What we propose

- **Reuse large** pre-trained unimodal **models** while keeping them **completely frozen** and free of adapter layers
- **Learn** a **lightweight mapping** between the representation spaces of pretrained unimodal models.

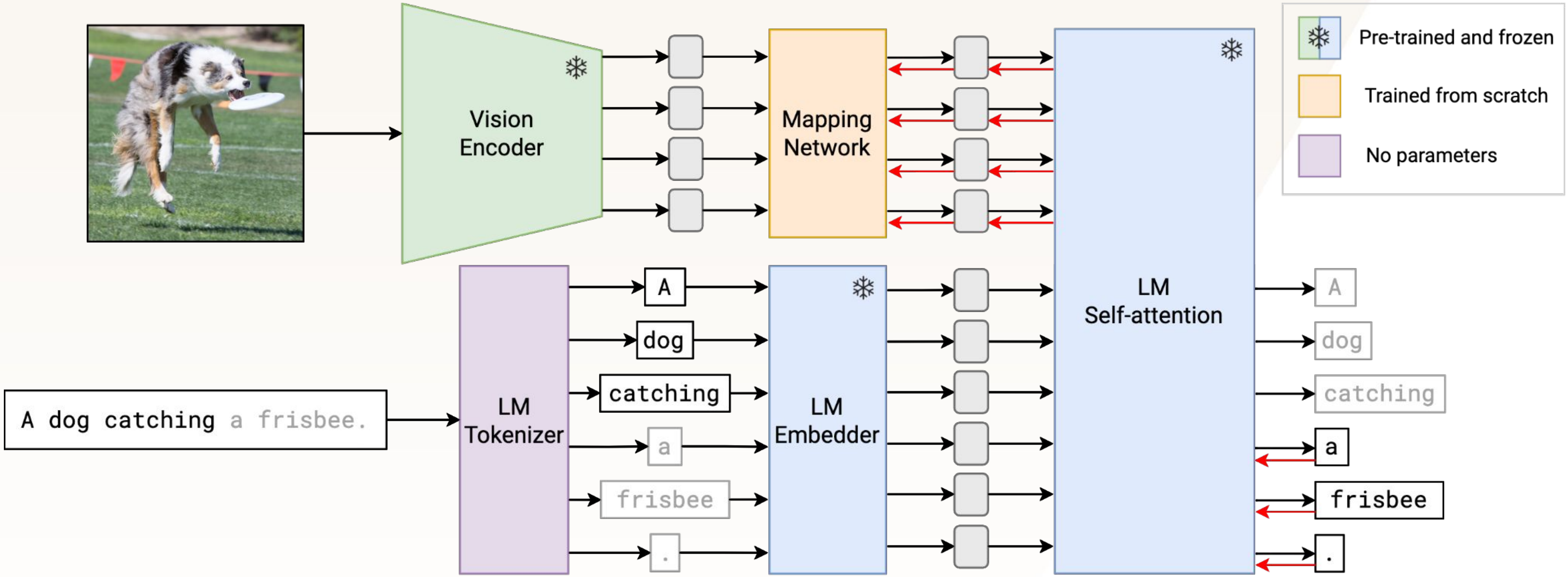


Benefits of our approach:

- Orders of magnitude fewer parameters
- Can be trained in just a few hours
- Uses modest computational resources and public datasets
- Modular, hence easily extensible to newer/better pretrained unimodal models

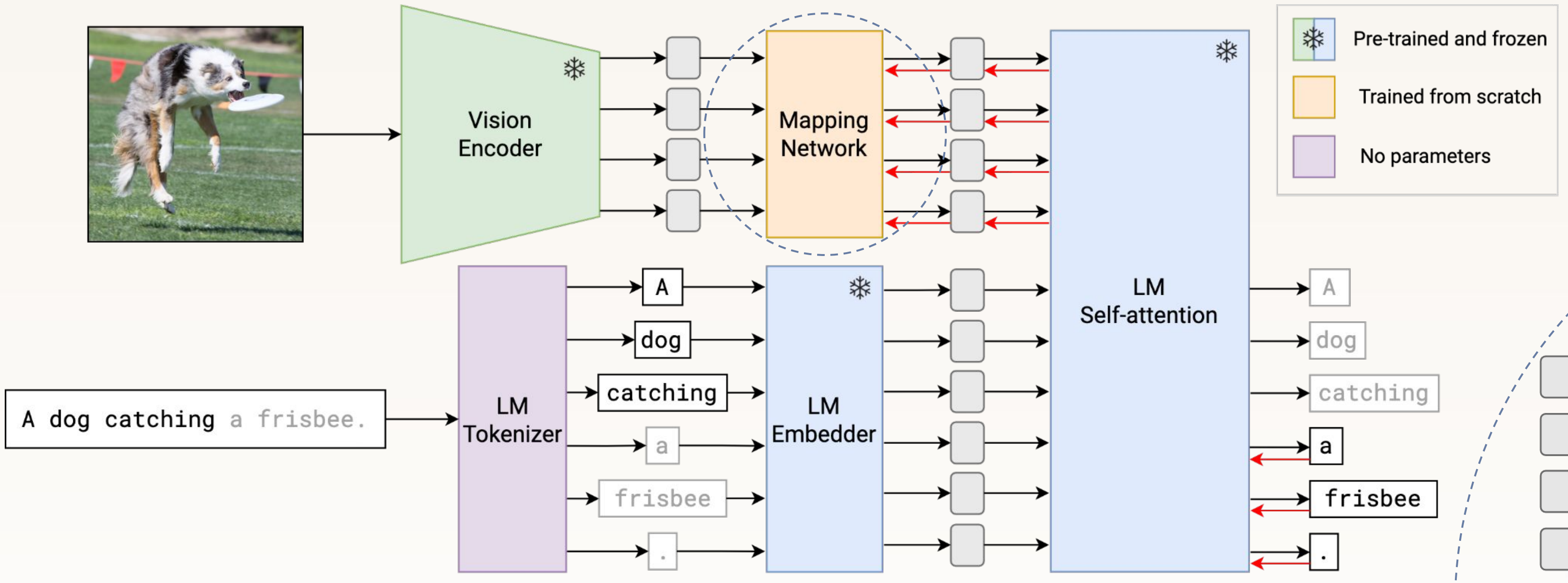
Figure credits: Oscar Mañas

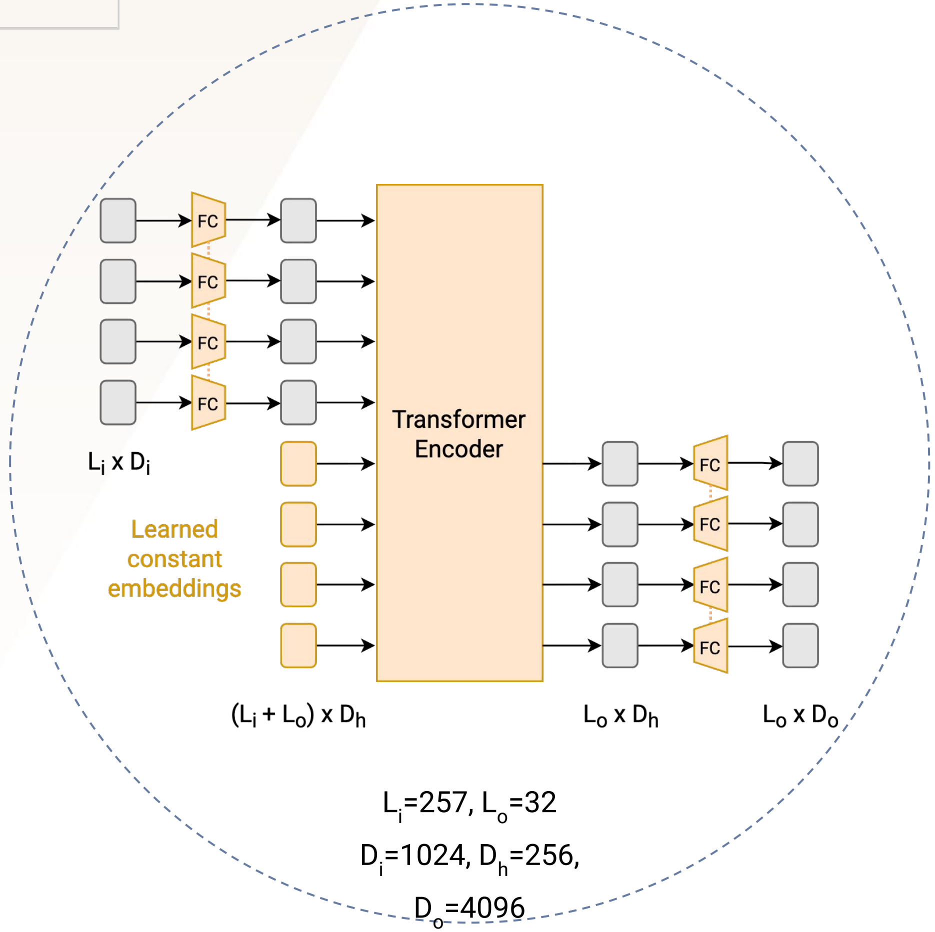
MAPL 🍁: method



Slide credits: Oscar Mañas

MAPL🍁: method

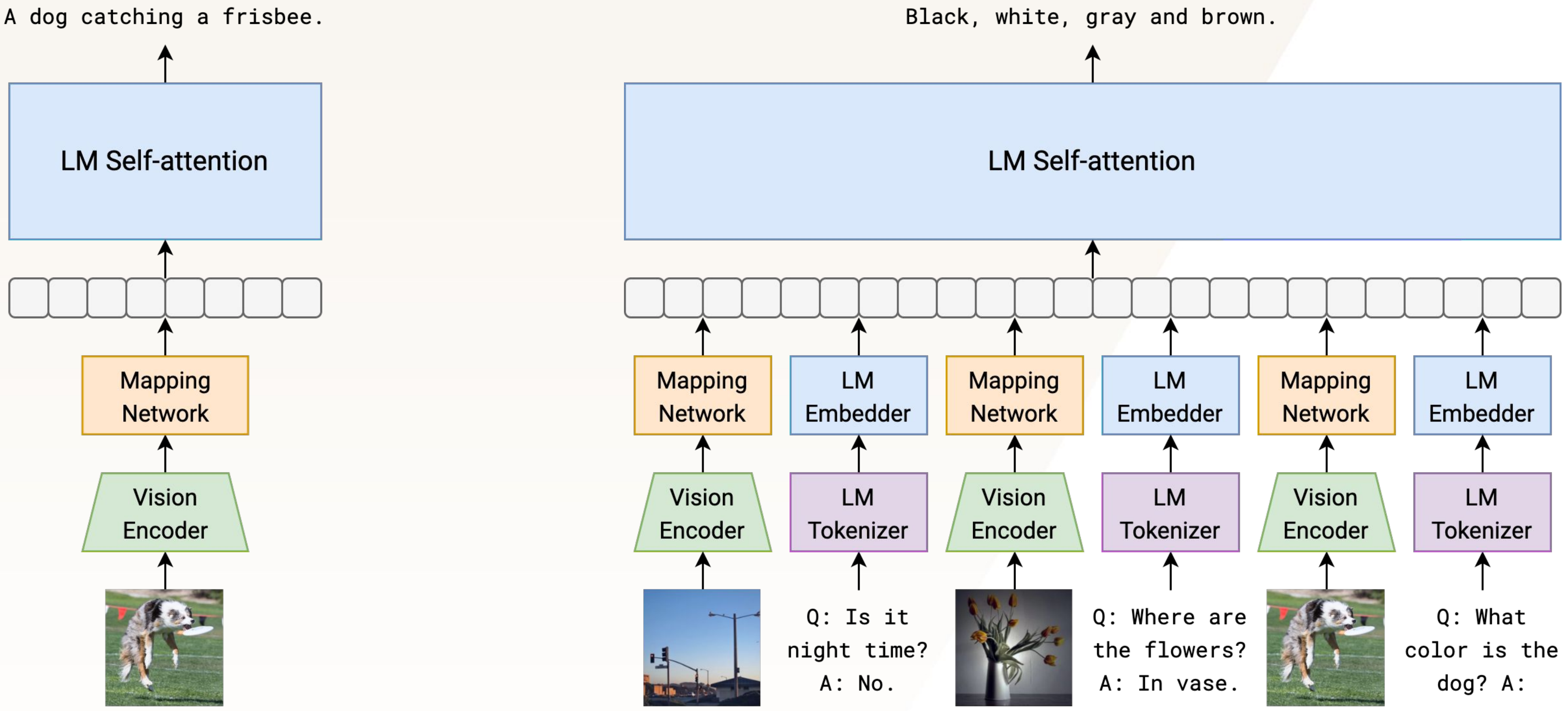




- Dimensionality bottleneck.
- Shared projection layers.
- Learned constant embeddings.

Slide credits: Oscar Mañas

MAPL: inference



0-shot image captioning.

2-shot VQA.

Slide credits: Oscar Mañas

MAPL : experimental results

- MAPL achieves **superior or competitive** performance compared to similar methods while training orders of magnitude **fewer parameters**.
- MAPL is **more effective** than the baseline in **low-data** settings.

	Trainable params	Training examples	n-shot VQAv2 0	n-shot VQAv2 4	n-shot VQAv2 8	n-shot OK-VQA 0	n-shot OK-VQA 4	n-shot OK-VQA 8	n-shot TextVQA 0	n-shot TextVQA 4	n-shot TextVQA 8	n-shot VizWiz-VQA 0	n-shot VizWiz-VQA 4	n-shot VizWiz-VQA 8	n-shot Overall 0	n-shot Overall 4	n-shot Overall 8
Existing methods using domain-agnostic training																	
Frozen	40.3M†	3.3M	29.50	38.20	-	5.90	12.60	-	-	-	-	-	-	-	-	-	-
MAGMA $_{\text{CC12M}}$	243M†	3.8M	36.90	45.40	-	13.90	23.40	-	-	-	-	5.60	10.60	-	-	-	-
VLKD $_{\text{CC3M}}$	406M	3.3M	38.60	-	-	10.50	-	-	-	-	-	-	-	-	-	-	-
Flamingo	10.2B	>2.1B	-	-	-	50.60	57.40	57.50	35.00	36.50	37.30	-	-	-	-	-	-
100% domain-agnostic training																	
MAPL-blind $_{\text{CC-clean}}$	3.4M	374K	20.62	35.01	35.11	4.84	14.68	14.28	3.68	5.43	5.82	3.18	8.65	9.55	8.08	15.94	16.19
Frozen* $_{\text{CC-clean}}$	40.3M	374K	25.98	37.80	38.52	5.51	18.86	19.91	5.11	6.15	6.30	4.33	11.28	16.68	10.23	18.52	20.35
MAPL $_{\text{CC-clean}}$	3.4M	374K	**33.54**	**45.13**	**45.21**	**13.84**	**24.25**	**23.93**	**8.26**	**8.88**	**8.77**	**11.72**	**18.46**	**19.52**	**16.84**	**24.18**	**24.36**
1% domain-agnostic training																	
Frozen* $_{\text{CC-clean}}$	40.3M	3.7K	26.22	36.69	37.41	5.50	**18.76**	**20.51**	5.71	**7.19**	7.53	3.83	**11.71**	**16.66**	10.31	**18.58**	**20.53**
MAPL $_{\text{CC-clean}}$	3.4M	3.7K	**30.80**	**37.38**	**37.95**	**8.77**	18.18	19.15	**6.40**	7.07	**7.74**	**5.68**	9.26	10.58	**12.91**	17.97	18.85
100% in-domain training																	
PICa*	0	0	20.61	46.86	47.80	11.84	**31.28**	**33.07**	-	-	-	-	-	-	-	-	-
Frozen* $_{\text{COCO}}$	40.3M	414K	32.09	38.90	39.42	9.81	20.72	21.83	7.54	6.82	6.74	5.87	12.07	17.35	13.82	19.63	21.33
Frozen* $_{\text{TextCaps}}$	40.3M	103K	32.49	37.39	38.03	11.34	19.87	20.82	8.83	7.33	7.51	6.25	12.26	16.86	14.73	19.21	20.80
Frozen* $_{\text{VizWiz}}$	40.3M	110K	26.93	37.38	37.91	5.85	19.12	20.64	6.38	7.44	7.47	5.57	13.06	18.06	11.18	19.25	21.02
MAPL $_{\text{COCO}}$	3.4M	414K	**43.51**	**48.75**	**48.44**	**18.27**	31.13	31.63	10.99	11.10	11.08	**14.05**	17.72	19.18	21.70	**27.17**	**27.58**
MAPL $_{\text{TextCaps}}$	3.4M	103K	38.83	43.34	43.43	16.33	25.07	25.92	**22.27**	**19.53**	**19.75**	12.31	16.69	18.18	**22.43**	26.15	26.82
MAPL $_{\text{VizWiz}}$	3.4M	110K	32.80	42.94	43.20	11.70	24.91	25.73	9.27	10.36	10.23	10.42	**20.63**	**23.10**	16.05	24.71	25.56
1% in-domain training																	
Frozen* $_{\text{COCO}}$	40.3M	4.1K	30.18	37.23	37.89	9.33	19.60	20.71	7.43	7.65	7.67	4.37	12.00	16.48	12.83	19.12	20.69
Frozen* $_{\text{TextCaps}}$	40.3M	1.0K	32.09	36.72	37.25	10.75	18.85	19.51	8.17	7.57	7.28	5.39	11.79	16.20	14.10	18.73	20.06
Frozen* $_{\text{VizWiz}}$	40.3M	1.1K	29.62	37.30	37.87	7.57	19.36	20.60	7.16	7.17	7.25	4.53	**12.51**	**17.56**	12.22	19.08	**20.82**
MAPL $_{\text{COCO}}$	3.4M	4.1K	**37.69**	**40.42**	**40.84**	**13.92**	**21.66**	**22.41**	8.30	6.96	6.84	**6.94**	10.72	12.43	**16.71**	**19.94**	20.63
MAPL $_{\text{TextCaps}}$	3.4M	1.0K	33.57	36.70	36.87	12.46	17.45	18.21	**9.34**	**8.29**	**8.62**	6.54	9.58	11.62	15.48	18.00	18.83
MAPL $_{\text{VizWiz}}$	3.4M	1.1K	31.88	36.81	37.04	9.59	17.64	17.64	7.25	5.99	6.04	4.73	9.48	11.33	13.36	17.48	18.01

ArXiv: https://arxiv.org/abs/2210.07179

MAPL🍁 : qualitative results

Input:

Completion:

a man watches the sea birds as they fly over the beach.

a rail crossing with a sign warning of trains.

a boy playing soccer in the field.

Input:

What kind of leaf is this?

What does this animal eat?

What type of cheese is on these vegetables?

Completion:

A maple leaf.

Squirrels eat nuts, seeds, berries, and insects.

broccoli.

Slide credits: Oscar Mañas

Vision-Language Challenges

- Out-of-distribution generalization
- Data-efficient adaptation to new tasks

Thanks!
Questions?