JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

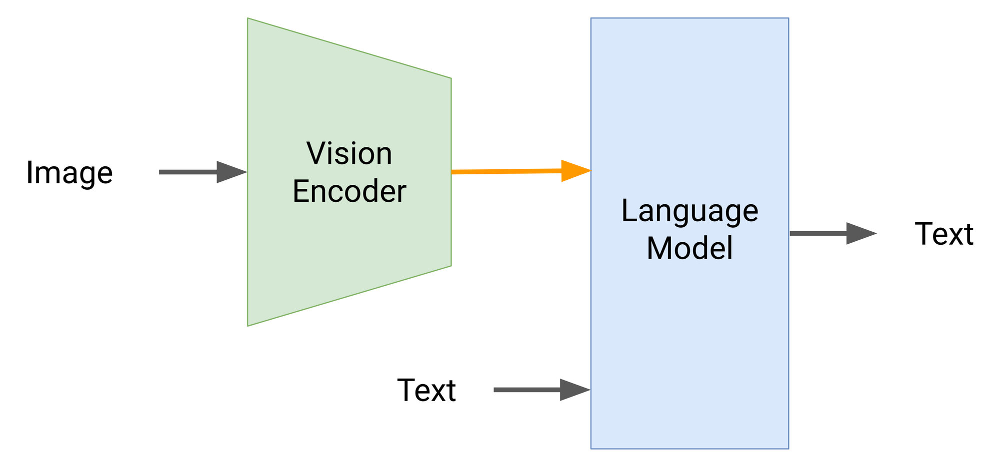
MAPL🍁: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting

Oscar Mañas, Pau Rodriguez*, Saba Ahmadi*, Aida Nematzadeh, Yash Goyal, Aishwarya Agrawal

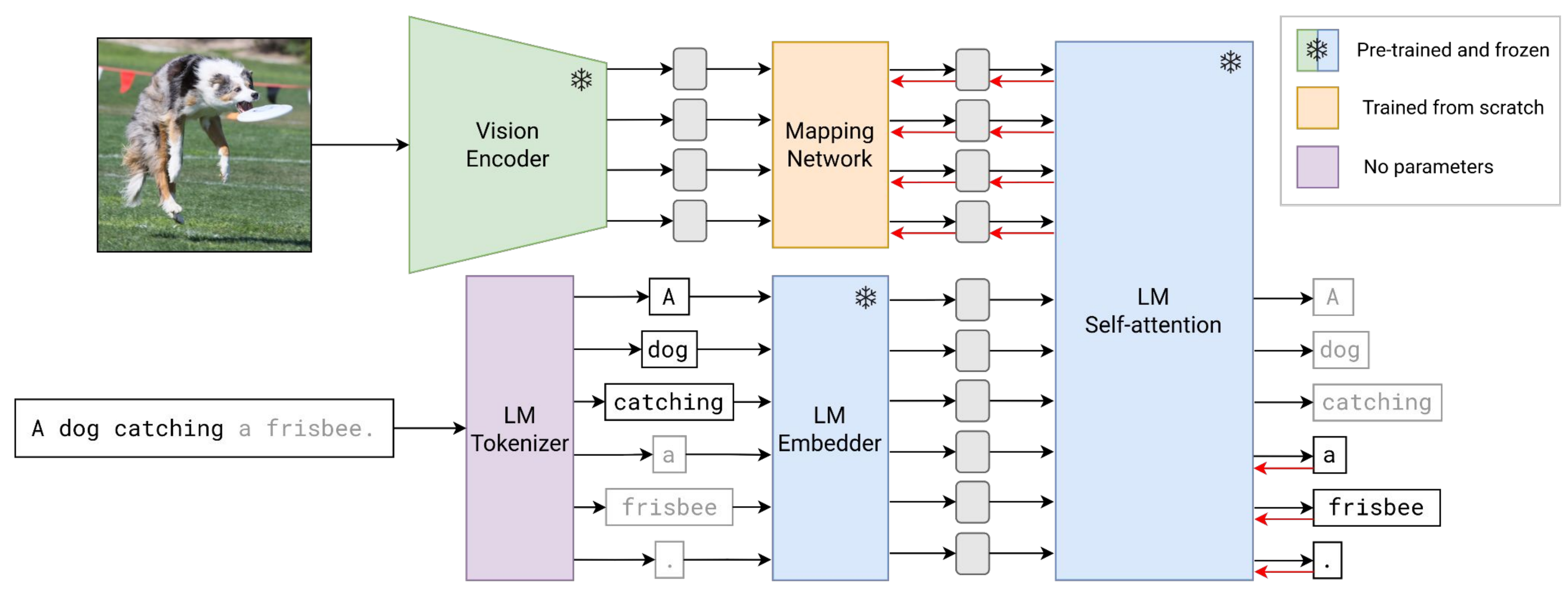
*denotes equal contribution

Motivation

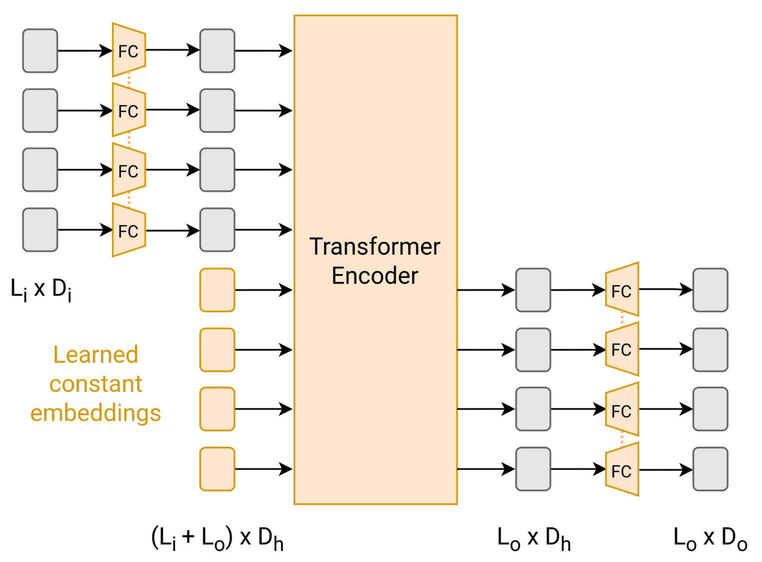
- Problem definition: processing images and text to generate text (e.g., image captioning, VQA)
- Impressive recent progress in learning vision-only and language-only pre-trained models (e.g., CLIP, ALIGN, GPT, OPT, LLaMA)
- Research question: can we reuse such powerful unimodal models and *efficiently adapt* them for multimodal vision-language downstream tasks?
- Issues with existing approaches (e.g., Frozen, MAGMA, Flamingo):
  - **Large** number of **trainable parameters** (~40M to ~10B)
  - Inserting adapter layers is **not straightforward**
  - Learning vision encoders from scratch **does not scale well**



Method



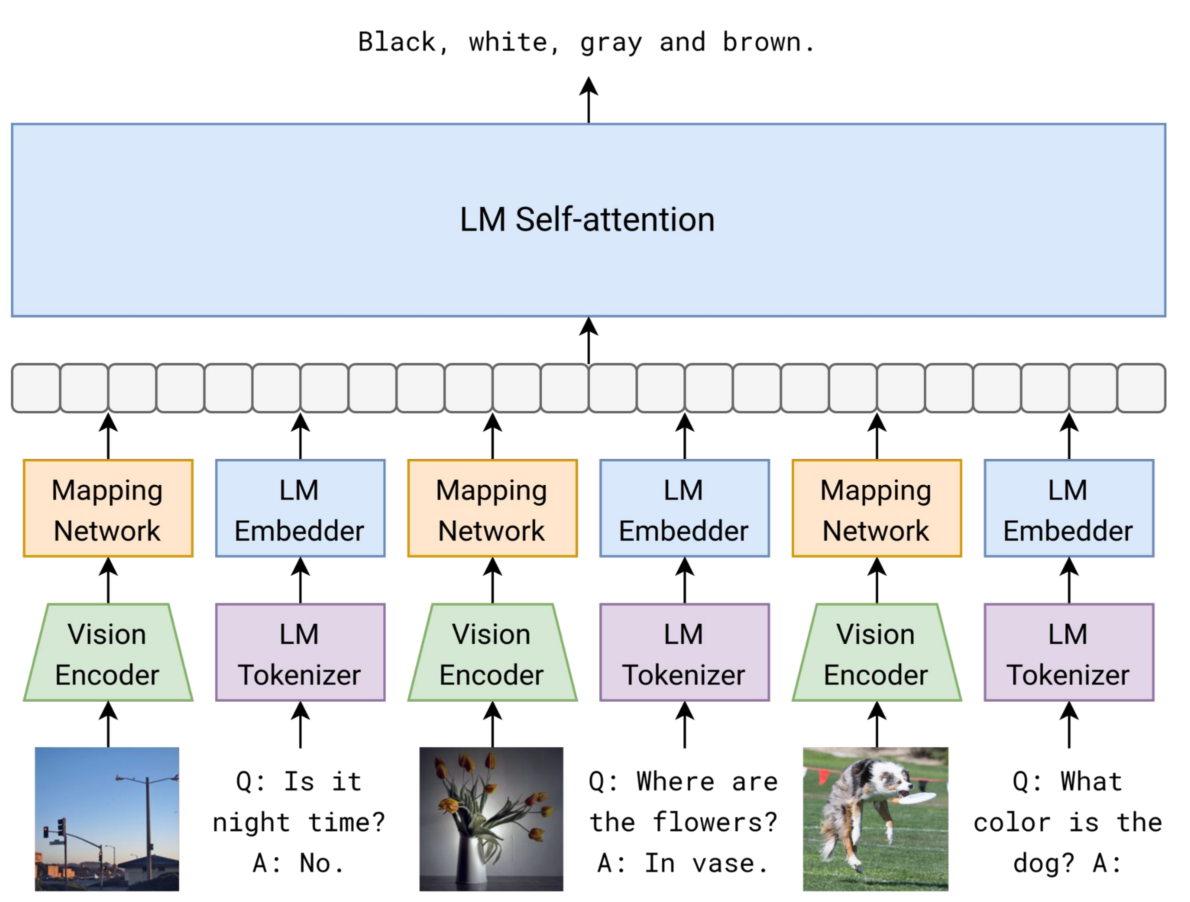
- Idea: learn a lightweight vision-language mapping between unimodal representation spaces
- CLIP-ViT-L/14 vision encoder (303M ❄️ params)
- GPT-J language model (6.1B ❄️ params)
- Transformer-based mapping network (3.4M 🔥 params)
- Image-conditioned language modeling loss



Benefits of our approach:

- Orders of magnitude fewer trainable 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

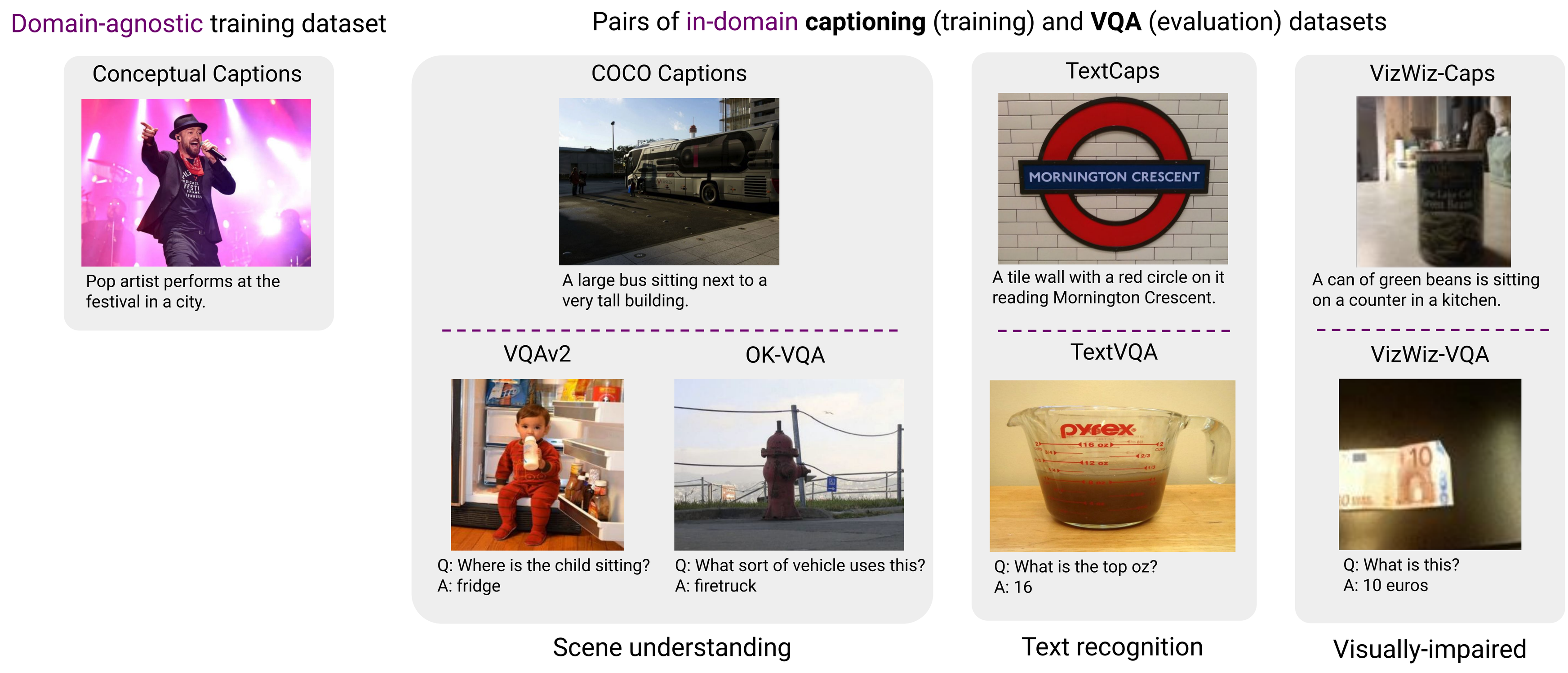
Vision-language few-shot prompting



- We can leverage the **in-context learning** capabilities of the frozen language model to transfer to **unseen vision-language tasks** (e.g., VQA)
- The learned mapping network allows us to feed **visual context** to the language model, enabling **multimodal prompting**
- Now the language model is able to "see"!

Experimental setting

- We evaluate our method on several vision-language benchmarks across two tasks:
  - Image captioning (CC, COCO, TextCaps, VizWiz-Caps)
  - Few-shot VQA (VQAv2, OK-VQA, TextVQA, VizWiz-VQA)

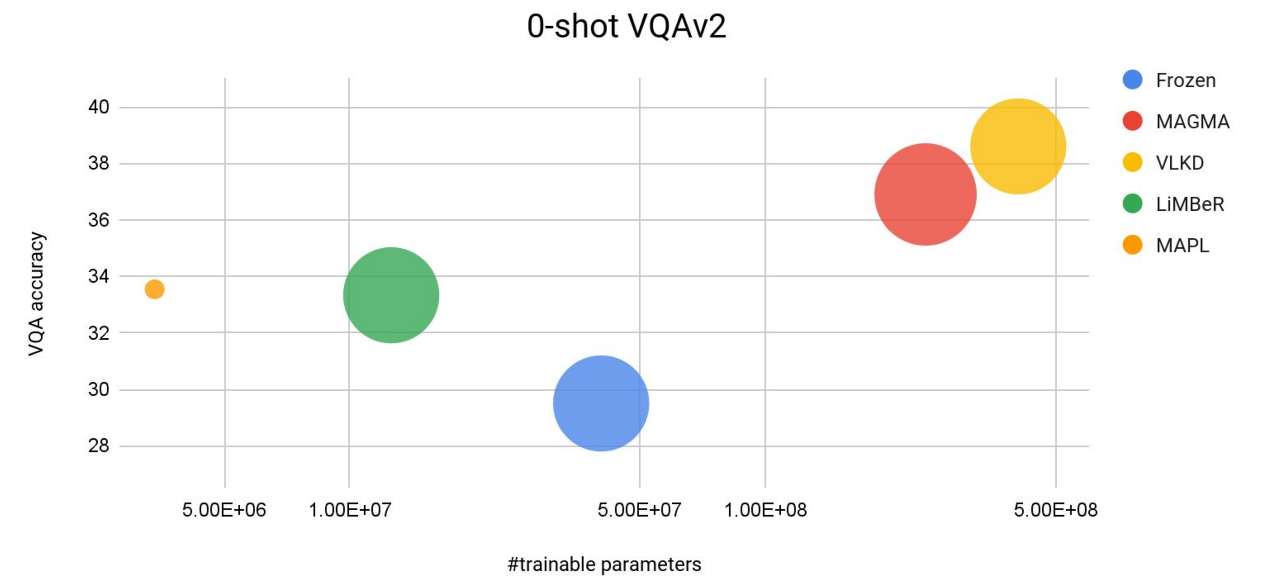


Quantitative results

Domain-agnostic training

Image-caption pairs come from a domain agnostic of the downstream task

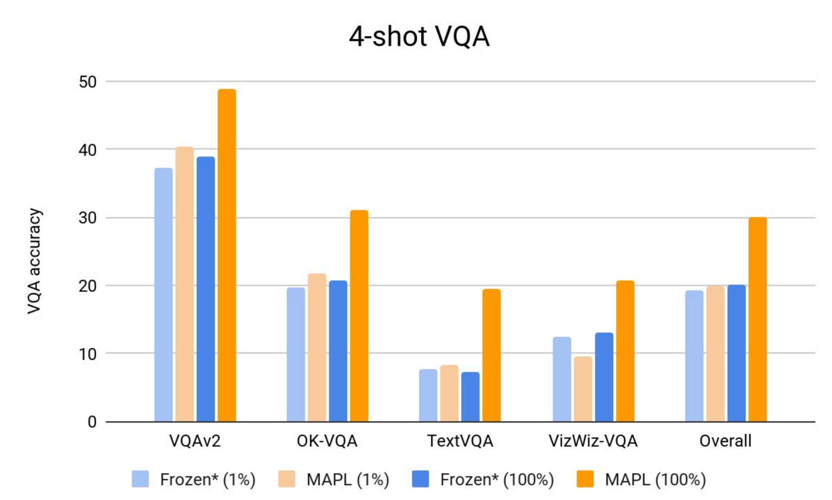
- MAPL is competitive with existing and concurrent methods while training orders of magnitude fewer parameters on less multimodal data (bubble size)

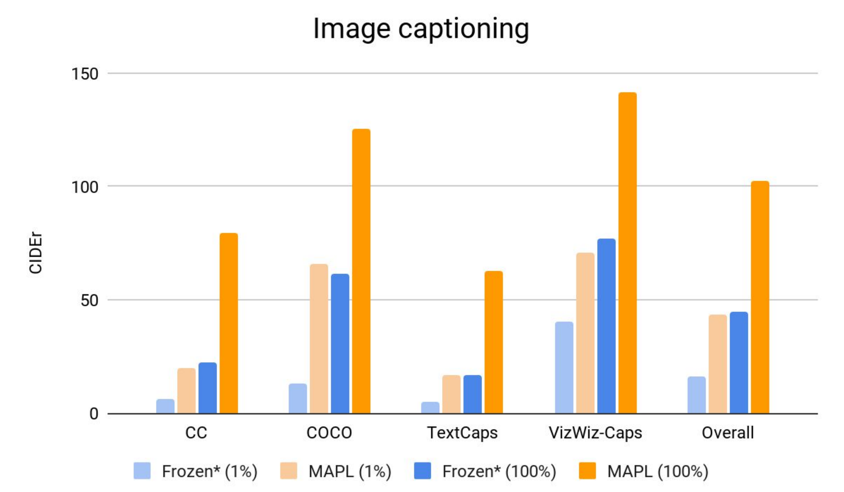


In-domain training

Image-caption pairs come from same domain as the downstream task

- Both MAPL and Frozen benefit from directly training on in-domain data, but the gap is larger for MAPL
- MAPL outperforms Frozen on all considered tasks and benchmarks when training on 100% of in-domain data
- MAPL trained on 1% of in-domain data generally outperforms Frozen trained on 100% of in-domain data on 4-shot VQA





Qualitative results

Conclusion

- Pretrained vision and language models can be repurposed for new VL tasks with modest computational resources and public datasets
- MAPL matches or outperforms similar methods on several VL benchmarks with fewer trainable parameters and less training data
- MAPL is effective in low-data and in-domain settings, useful when training with large-scale datasets is difficult
- Effective recycling of large pretrained models is becoming increasingly important

MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting. Mañas O., Rodriguez P.*, Ahmadi S.*, Nematzadeh A., Goyal Y. & Agrawal A. arXiv preprint arXiv:2210.07179 (2022)