ePALM: Efficient Perceptual Augmentation of Language Models

Mustafa Shukor, Corentin Dancette, Matthieu Cord

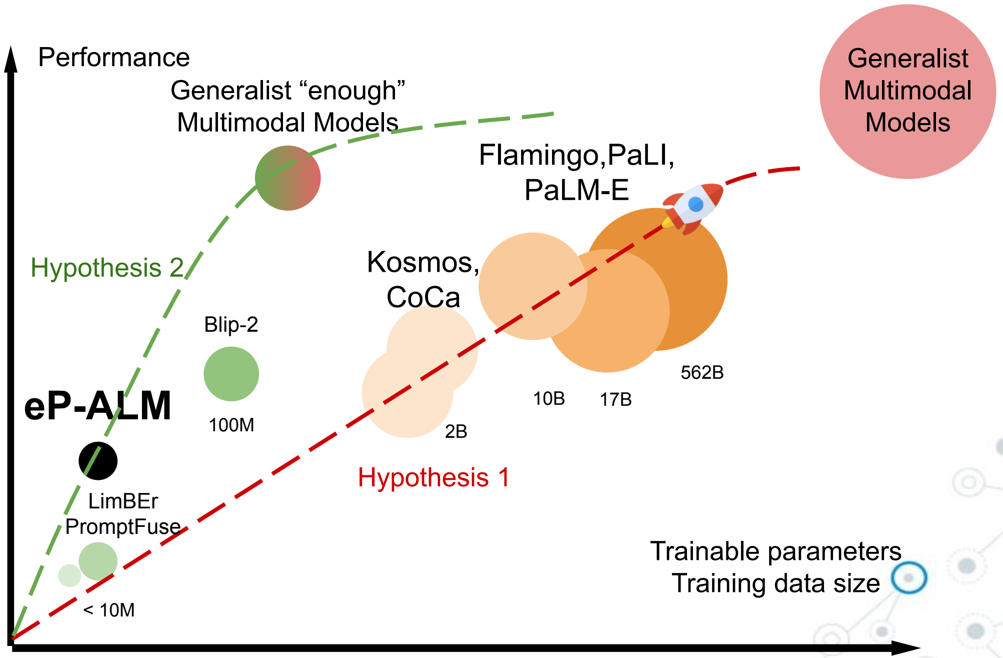
Sorbonne University

Towards Generalist Multimodal Models

Assumption: very powerful unimodal models (e.g. LLMs)

Objective: how to build powerful Multimodal Models?

- **Hypothesis 1**: large-scale multimodal training
- **Hypothesis 2**: Efficient Adaptation of pretrained unimodal models

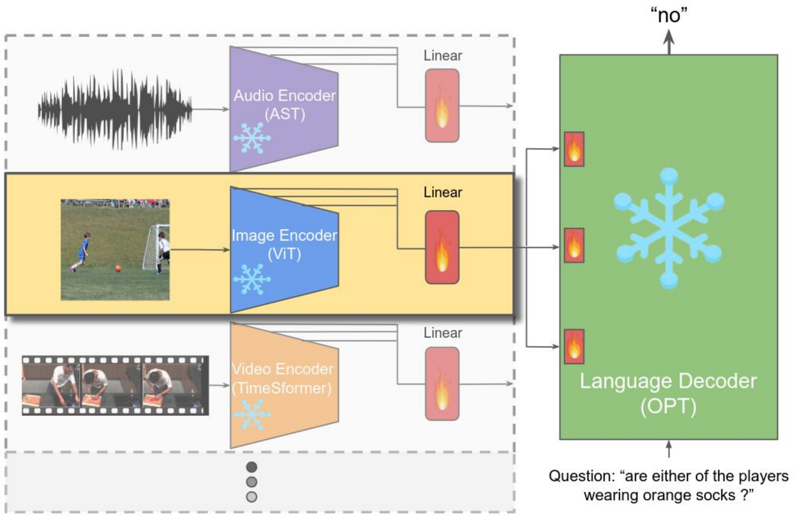


TL;DR

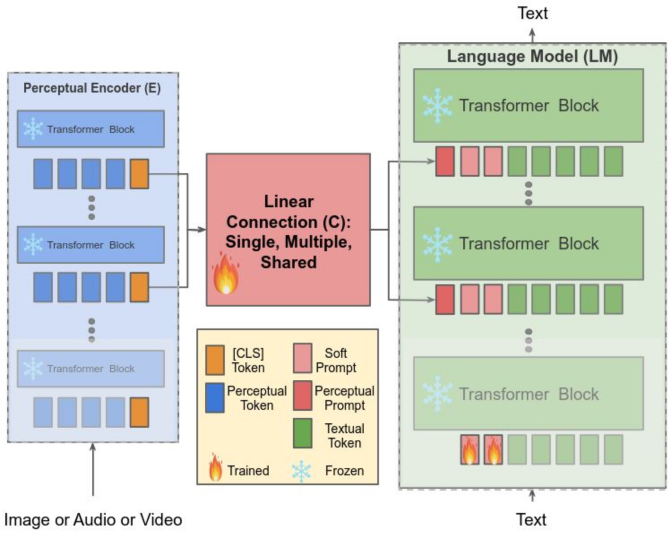
Summary of the work: we efficiently (**<0.06%** train. param./**No pretraining**) adapt frozen, pretrained, **unimodal** models (e.g OPT and ViT) to solve multimodal tasks (VQA, Captioning) across image, **video** and **audio** modalities

3 main takeaways:

1. Parameter-Efficiency: Training only a **linear projection**
2. **Late cross-modal interaction** mechanism
3. Data-Efficiency: **without multimodal pretraining**, **few-shot learning**



Our Recipe:



- **Model**:
  - *Language Model*: e.g. OPT
  - *Unimodal Encoders*: ViT-Base (ImageNet), TimeSformer-B (kinetics), AST-B (audioset)
  - *Adaptation parametres*:
    - ***Cross-Modal Connection***: **linear projection** that projects the visual/audio **[CLS]** tokens extracted from the **encoders' last layers** and inject them in the **OPT's last layers**
    - ***Soft Prompt***: 10 learnable tokens prepended to the text input
- **Data**: target datasets (e.g. COCO, VQAv2, AudioCaps, MSR-VTT)
- **Training**: training only adaptation parameters on target dataset

Results: Image-Text Tasks

Last layer's visual tokens fed to **input OPT layer**

- **Lin. proj.**
- **Prompt Tuning (PT)+Lin. proj.**
- **Adapters+Lin proj.**

Late cross interaction mechanism

- **PT+Shared Lin. proj.**
- **PT+Multiple Lin. proj.**
- **PT+Shared Lin proj.**

Method	VQA v2		GQA		COCO	
	Val	Test	Val	Test	B@4	CIDEr
PromptFuse† [55]	34.1†	–	–	–	–	–
B_{LimBEr}	34.1	33.5	30.81	29.4	–	–
$B_{PromptFuse}$	40.4	39.5	33.74	31.51	15.05	48.26
B_{MAGMA}	32.2	31.8	30.98	28.93	–	–
eP-ALM$_{pt}$	48.8	47.8	43.8	40.3	27.52	91.92
eP-ALM	**50.7/53.3**†	**50.2**	**45.0**	**40.4**	**29.47**	**97.22**
eP-ALM$_{pt}$-L*	54.58/54.47†	54.47	46.86	42.7	31.24	107.0

- Consistently better than other baselines that prepend visual tokens to the input layer and use adapters or prompt tuning
- Better data-efficiency and zero-shot generalization (details in the paper)

Results: Comparison with SoTA

Dataset (Metric)	SoTA (ZS)	eP-ALM (FT)	SoTA (FT)
AudioCaps (CIDEr)	–	<u>63.6</u>	**66.7** (Liu et al. [59])
MSRVTT-QA (Acc)	17.4 (Flamingo80B [2])	<u>36.7</u>	**44.1** (OmniVL [88])
MSR-VTT (CIDEr)	–	<u>50.7</u>	**60** (MV-GPT [73])
COCO (CIDEr)	84.3 (Flamingo80B [2])	<u>107.0</u>	**145.3** (OFA [89])
VQAv2 (Acc)	<u>56.3</u> (Flamingo80B [2])	53.3	**84.3** (PaLI [14])
GQA (Acc)	29.3 (FewVLM [43])	<u>42.7</u>	**60.8** (VL-T5 [17])

- Comparison with SoTA, trained with large number of parameters, and most often include large-scale pretraining

Conclusion

Direct Finetuning (eP-ALM)

👍 Efficient to train
👍 Generally better performance
👍 Easy to adapt to new tasks/datasets
👍 Efficient to adapt to new LLMs
👎 Task-specific finetuning

Pretrain-Zeroshot (e.g. LimBEr, Flamingo)

👎 Costly pretraining
👎 Limited performance, saturation with FS ICL
👎 Finetuning is needed for “new” datasets/tasks
👎 Pretraining is needed for a new LLM
👍 One training for many tasks

- **Future directions:**
  - **Constraint relaxation**: beyond linear projection, more trainable parameters, better zero-shot capabilities, efficient Multimodal pretraining, better/larger LLMs
  - **Better cross-modal interaction mechanisms**

Code

https://github.com/mshukor/eP-ALM

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access