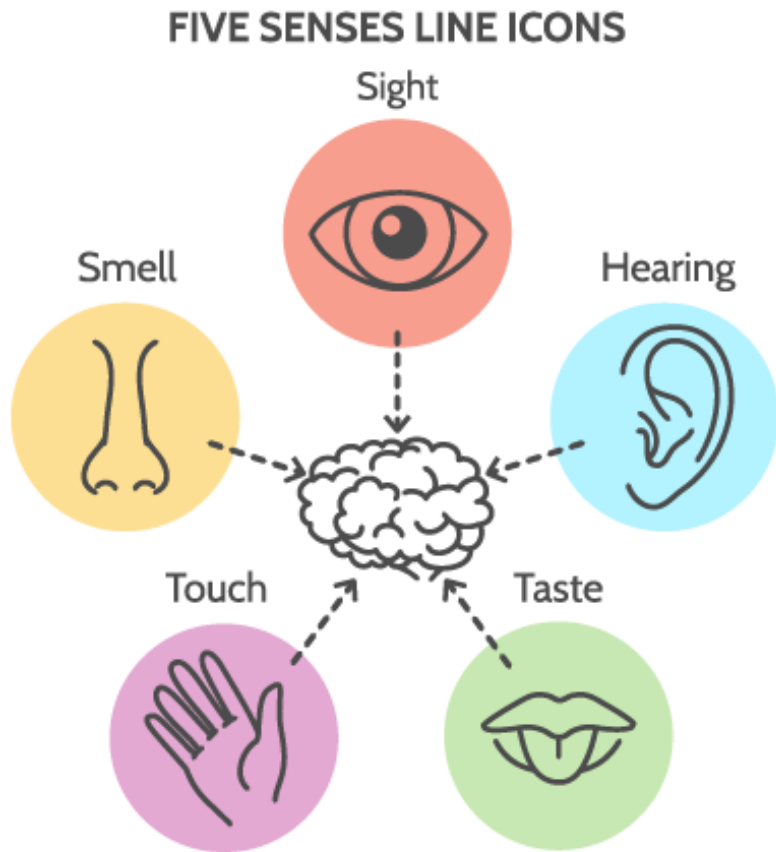




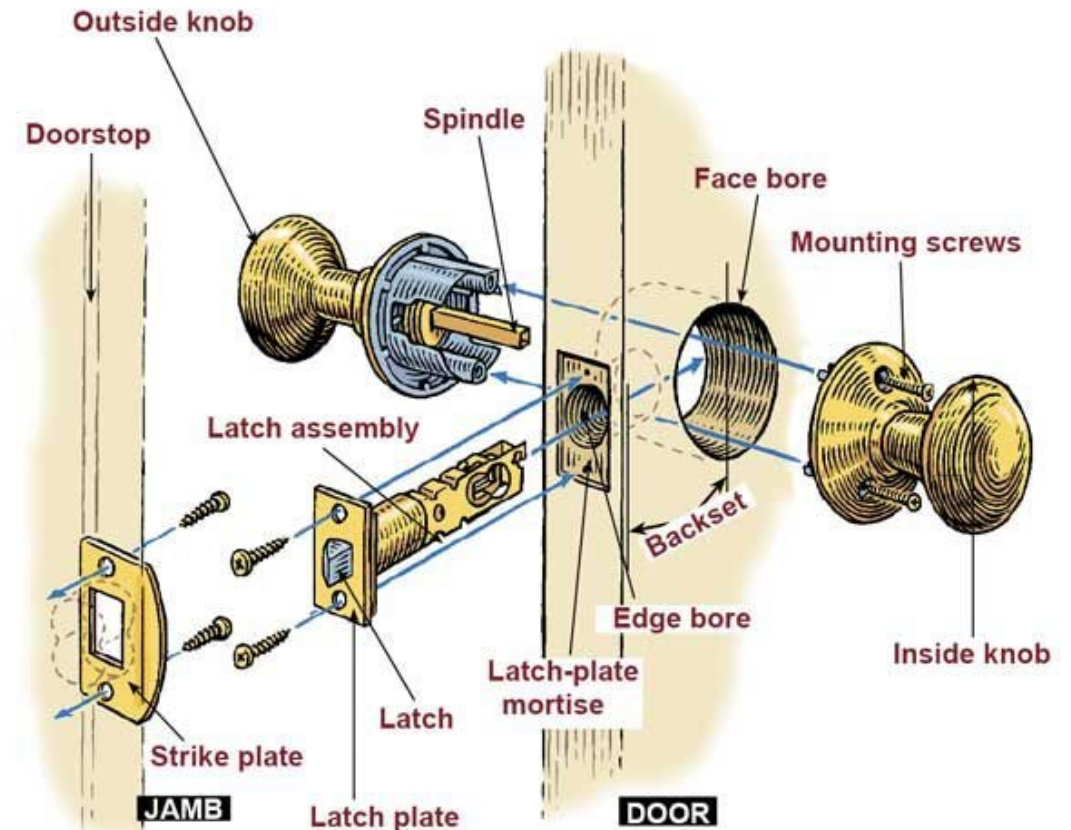
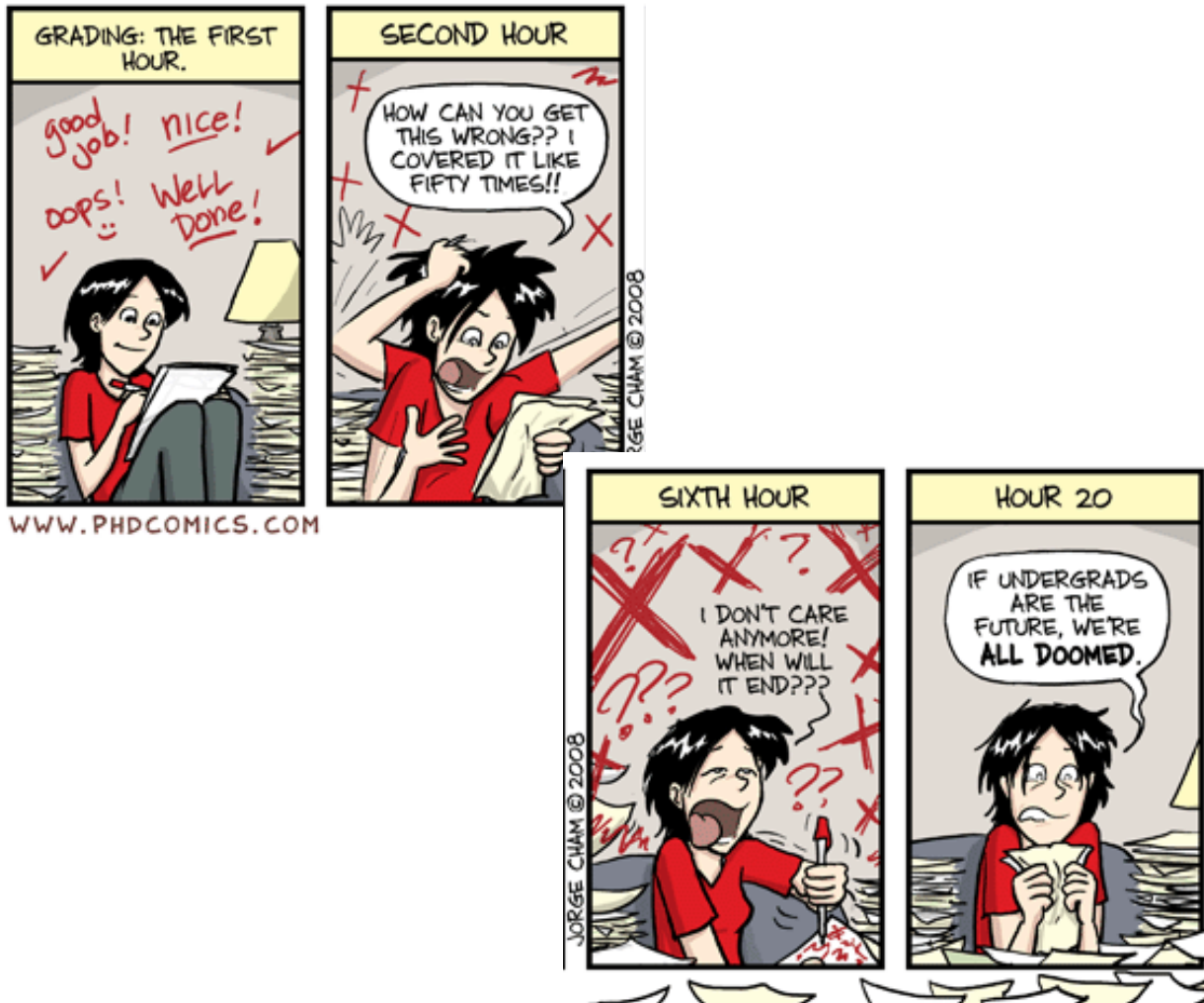
Multimodal Machine Learning in the Era of Gigantic Pretrained Models

Boyang “Albert” Li
Nanyang Associate Professor
Nanyang Technological University

Why Multimodal Learning



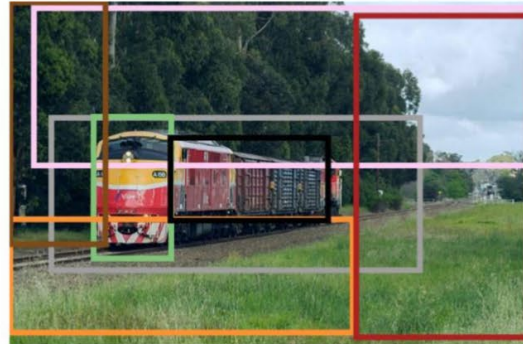
Humans Excel at Understanding Multimodal Information



Visual Captioning



A horse carrying a large load of hay and two people sitting on it.



train on the tracks. trees are green. front of the train is yellow. grass is green. green trees in the background. photo taken during the day. red train car.

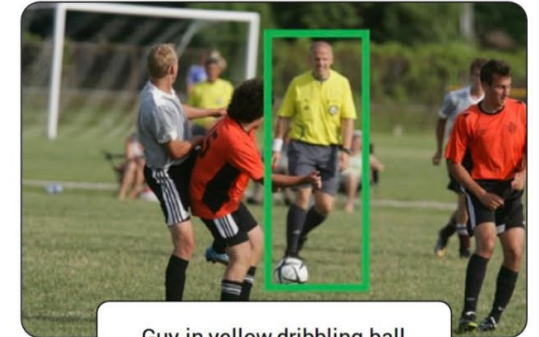
- **Popular Topics:** Advanced attentions, RL/GAN-based model training, Style diversity, Language richness, Evaluation
- **Popular Tasks:** Image/video captioning, Dense captioning, Storytelling

Visual QA/Grounding/Reasoning



Is there something to cut the vegetables with?

VQA



Guy in yellow dribbling ball

Referring Expressions

- **Popular Topics:** Multimodal fusion, Advanced attentions, Use of relations, Neural modules, Language bias reduction
- **Popular Tasks:** VQA, GQA, VisDial, Ref-COCO, CLEVR, VCR, NLVR2

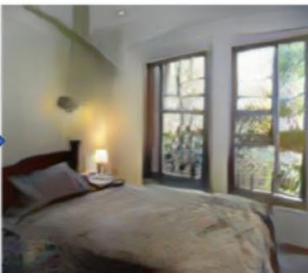
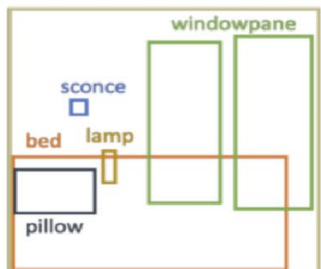
Text-to-image Synthesis

This bird is red with white belly and has a very short beak

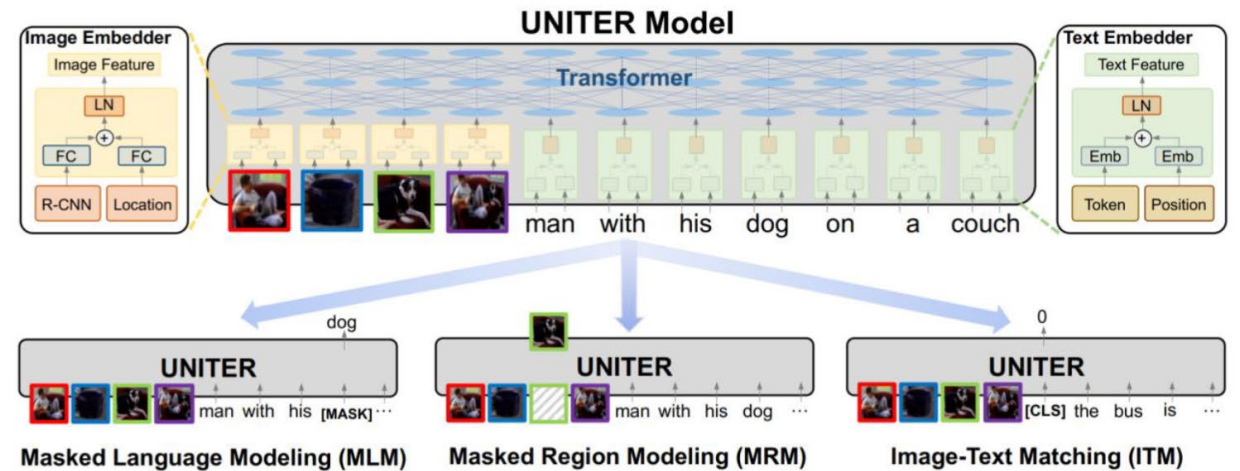


Popular Tasks:

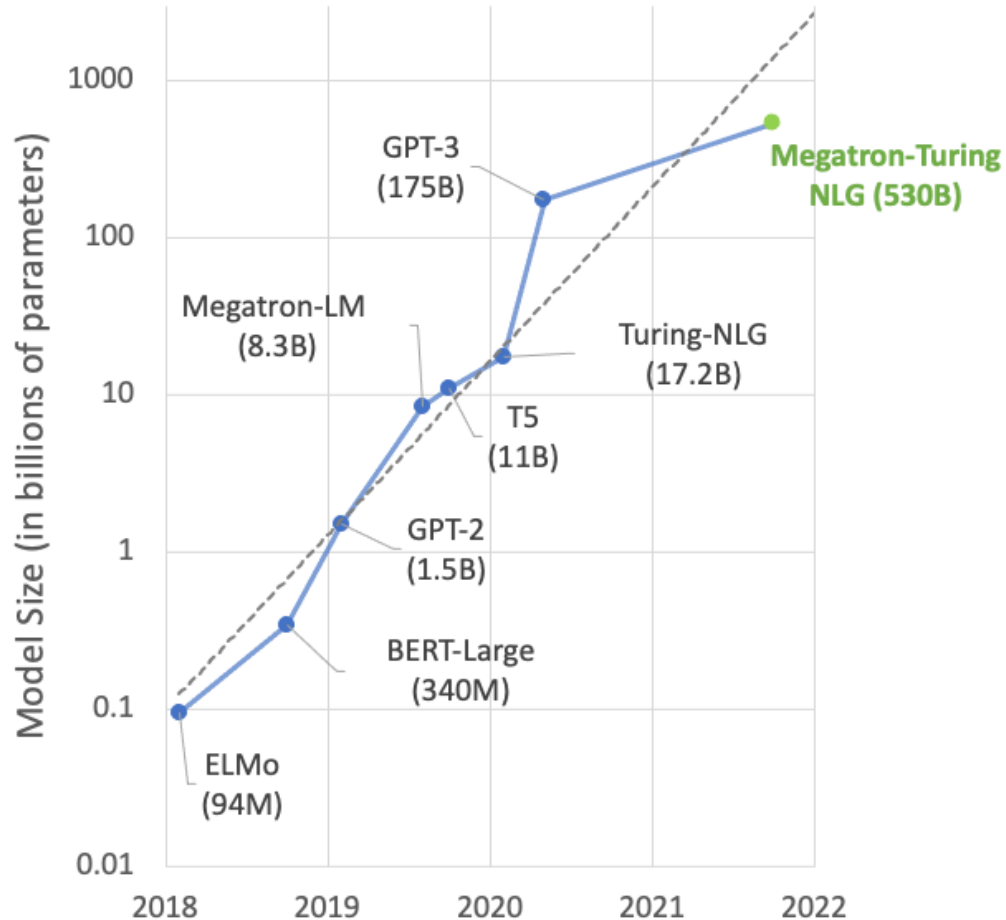
- Text-to-image
- Layout-to-image
- Scene-graph-to-image
- Text-based image editing
- Story visualization



Self-supervised Learning



Era of Large Pretrained Language Models (LPLMs)



Model Name	Year	# Parameters
T0	2021	11B
LaMDA	2021	137B
InstructGPT	2022	175B
GPT-NeoX	2022	20B
OPT	2022	175B
PaLM	2022	540B

LPLM: Training

- Given a context of words immediately before, predict the next word.

Wikipedia is a multilingual free online encyclopedia written and maintained by a community of _____

Correct answer: volunteers

$$\theta^* = \max_{\theta} Pr(w_t | w_{t-1}, w_{t-2}, \dots, w_1, \theta)$$

LPLMs: In-context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

LPLMs: Rationales & Prompt Engineering

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

LPLMs: Rationales & Prompt Engineering

LPLMs demonstrate strong abilities to perform reasoning with natural language as the intermediate representation.

“In-context learning” may be a misnomer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. 2022.

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

**Is it possible to reason about
visual content using language?**

Yes. Kinda of.

Visual Question Answering

- Object Detection and Attribute Identification
- Action Recognition
- Spatial Understanding
- Commonsense Reasoning

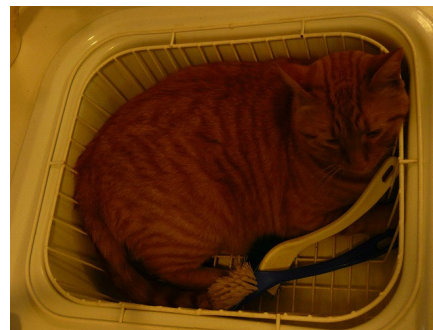
What animal is in the window? **Bird**



What is hanging above the toilet? **Teddy Bear**



Is the animal sleeping? **No**



Why are the men jumping? **to catch frisbee**



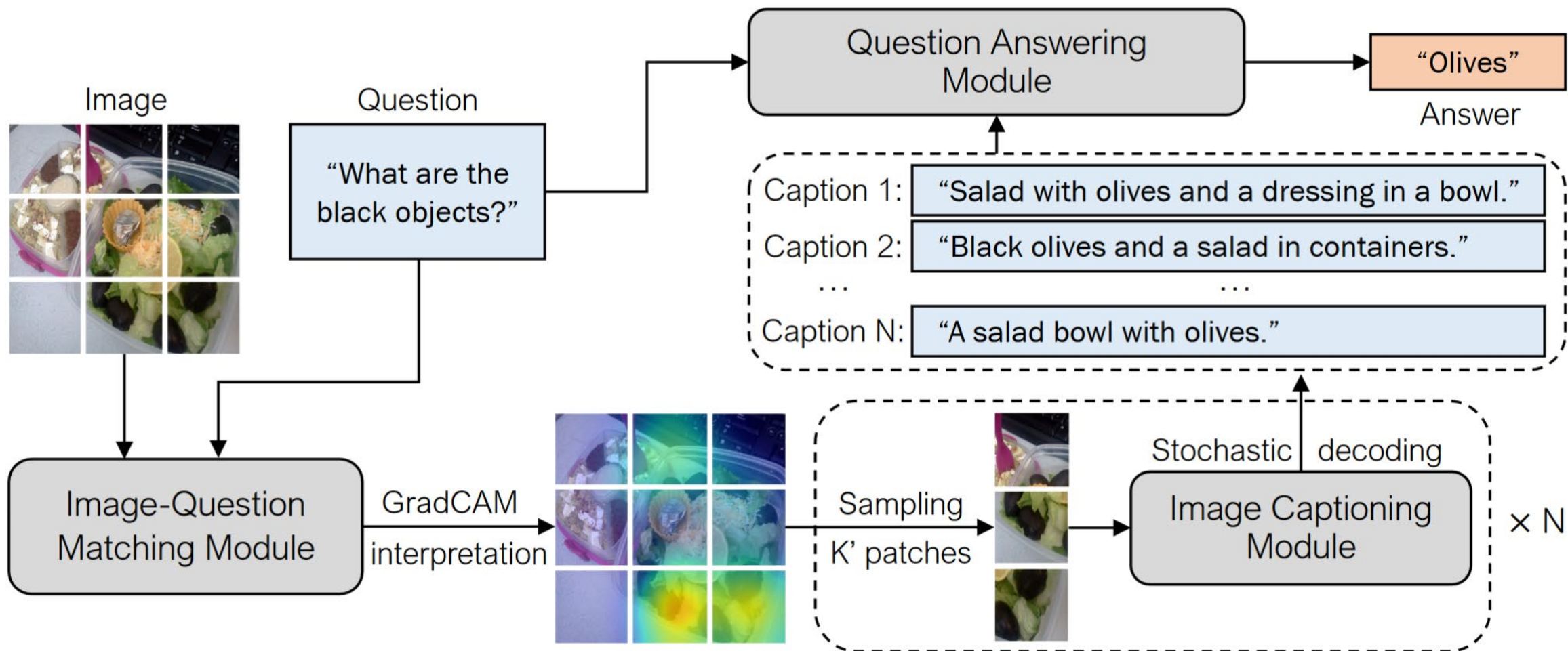
Plug-and-Play VQA

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training. EMNLP Findings. 2022.



- Conventional wisdom suggests that in order to connect pretrained models, we should perform some end-to-end training. Otherwise, performance will probably be low.
- We connect several pretrained models to perform VQA using language and saliency maps as the intermediate representation.
- NO training is required.
- We outperform Deepmind's Flamingo on zero-shot VQAv2 with fewer parameters

System Architecture



Q: what utensil is this?

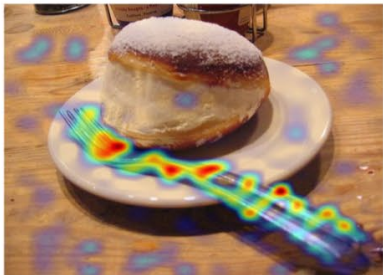
A: fork



Generic captions:

1. a spoon and fork are sitting on a white plate on a wooden table
2. a round cake with cream on it on a plate

Prediction: a spoon

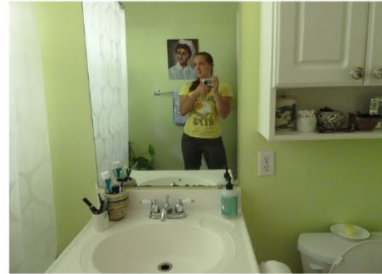


Question-guided captions:

1. a fork, silverware, fork and a spoon are shown
2. utensil on the plate which seems to have a fork and the fork

Prediction: fork

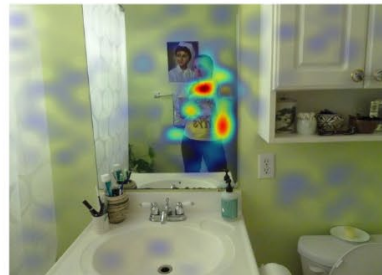
Q: what is the popular name for the type of photo this lady is taking? A: selfie



Generic captions:

1. a smiling teen girl taking a picture in a mirror
2. a person standing in a small bathroom taking a photo

Prediction: self-portrait



Question-guided captions:

1. a woman is taking a selfie and taking a selfie
2. a woman is taking a picture in a mirror and taking a picture

Prediction: selfie



Q: is there any art hanging on the walls? A: yes



Generic captions:

1. two beds in a suite with luggage in a bag on top of them
2. two large beds sitting in a room with suitcases

Prediction: no



Question-guided captions:

1. three pictures in a frame above two beds
2. a hotel room with 2 double beds and pictures on the wall

Prediction: yes

Q: what is the name of the theater? A: grand



Generic captions:

1. a very tall tower with a little clock on it
2. there is an old clock tower at this town

Prediction: the palace



Question-guided captions:

1. a white grand theatre, on a bright day
2. the grand store, grand in grand, is seen

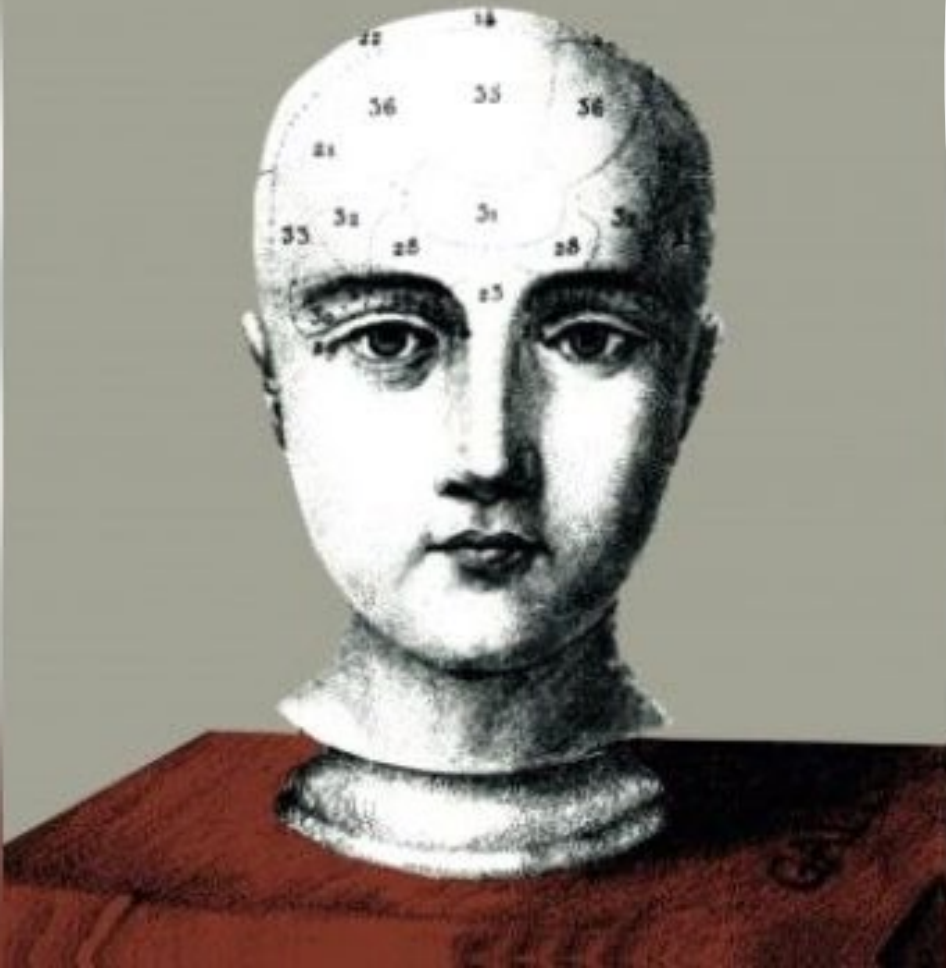
Prediction: grand

Method	Language			Vision			VQA _{v2}		OK-VQA	GQA
	Model	#Params	VL-aware	Model	#Params	VL-aware	Val	Test-dev	Test	Test-dev
<i>Pretrained models conjoined by end-to-end VL training.</i>										
VL-T5 _{no-vqa}	T5	224M	✓	Faster R-CNN	64M	✗	13.5	-	5.8	6.3
FewVLM _{base}	T5	224M	✓	Faster R-CNN	64M	✗	43.4	-	11.6	27.0
FewVLM _{large}	T5	740M	✓	Faster R-CNN	64M	✗	47.7	-	16.5	29.3
VLKD _{ViT-B/16}	BART	407M	✓	ViT-B/16	87M	✓	38.6	39.7	10.5	-
VLKD _{ViT-L/14}	BART	408M	✓	ViT-L/14	305M	✓	42.6	44.5	13.3	-
Flamingo _{3B}	Chinchilla-like	2.6B	✓	NFNet-F6	629M	✓	-	49.2	41.2	-
Flamingo _{9B}	Chinchilla-like	8.7B	✓	NFNet-F6	629M	✓	-	51.8	<u>44.7</u>	-
Flamingo _{80B}	Chinchilla	80B	✓	NFNet-F6	629M	✓	-	56.3	50.6	-
Frozen	GPT-like	7B	✗	NF-ResNet-50	40M	✓	29.5	-	5.9	-
<i>Pretrained models conjoined by natural language and zero training.</i>										
PICa	GPT-3	175B	✗	VinVL-Caption	259M	✓	-	-	17.7	-
PNP-VQA _{base}	UnifiedQAv2	223M	✗	BLIP-Caption	446M	✓	54.3	55.2	23.0	34.6
PNP-VQA _{large}	UnifiedQAv2	738M	✗	BLIP-Caption	446M	✓	57.5	58.8	27.1	38.4
PNP-VQA _{3B}	UnifiedQAv2	2.9B	✗	BLIP-Caption	446M	✓	<u>62.1</u>	<u>63.5</u>	34.1	42.3
PNP-VQA _{11B}	UnifiedQAv2	11.3B	✗	BLIP-Caption	446M	✓	63.3	64.8	35.9	<u>41.9</u>

Table 2: Comparison with state-of-the-art models on zero-shot VQA. Flamingo (Alayrac et al., 2022) inserts additional parameters into the language model and perform training using billion-scale vision-language data. The best accuracy is bolded and the second best is underlined.

THE MODULARITY OF MIND

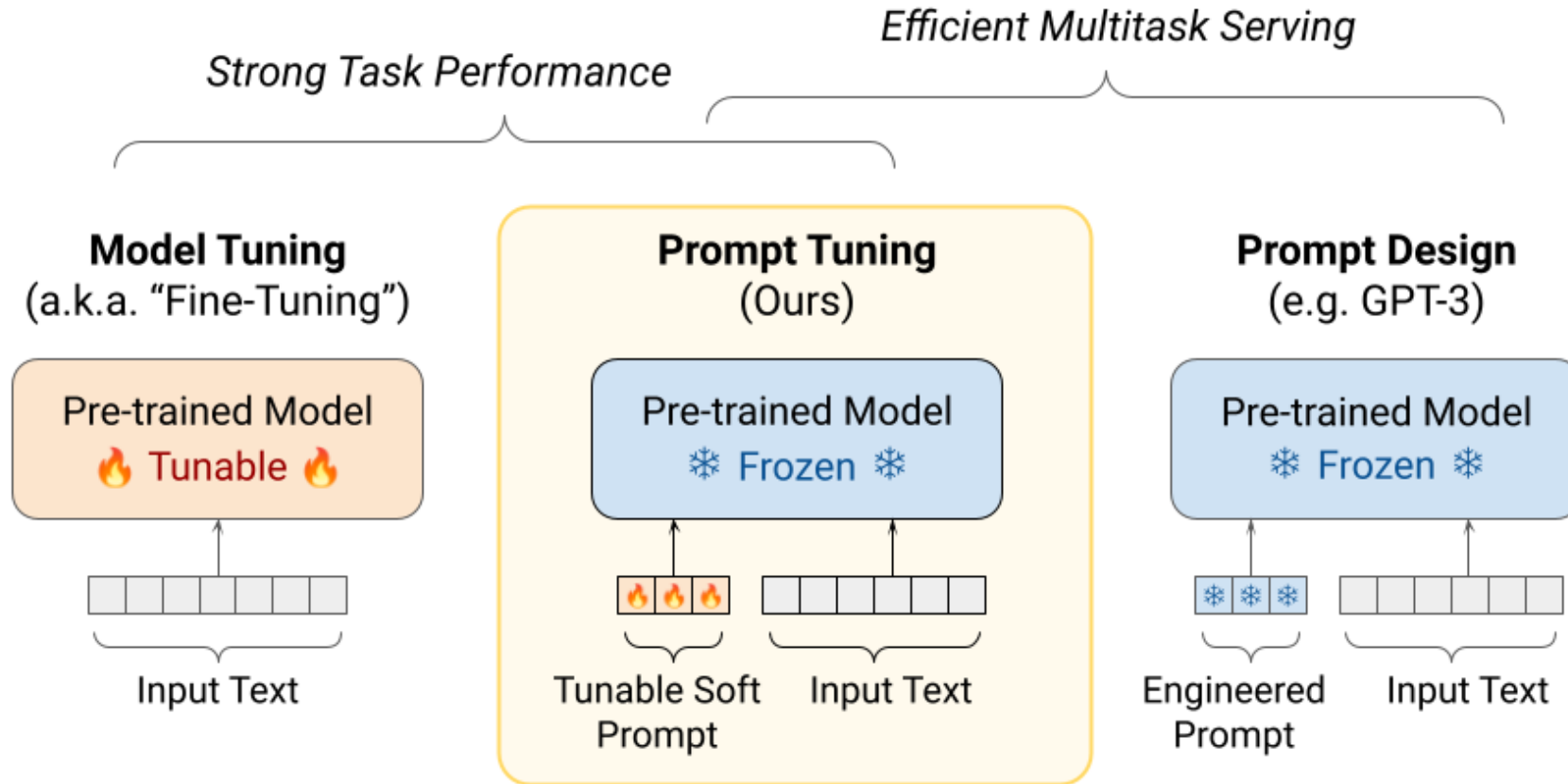
Jerry A. Fodor



Modular System Design?

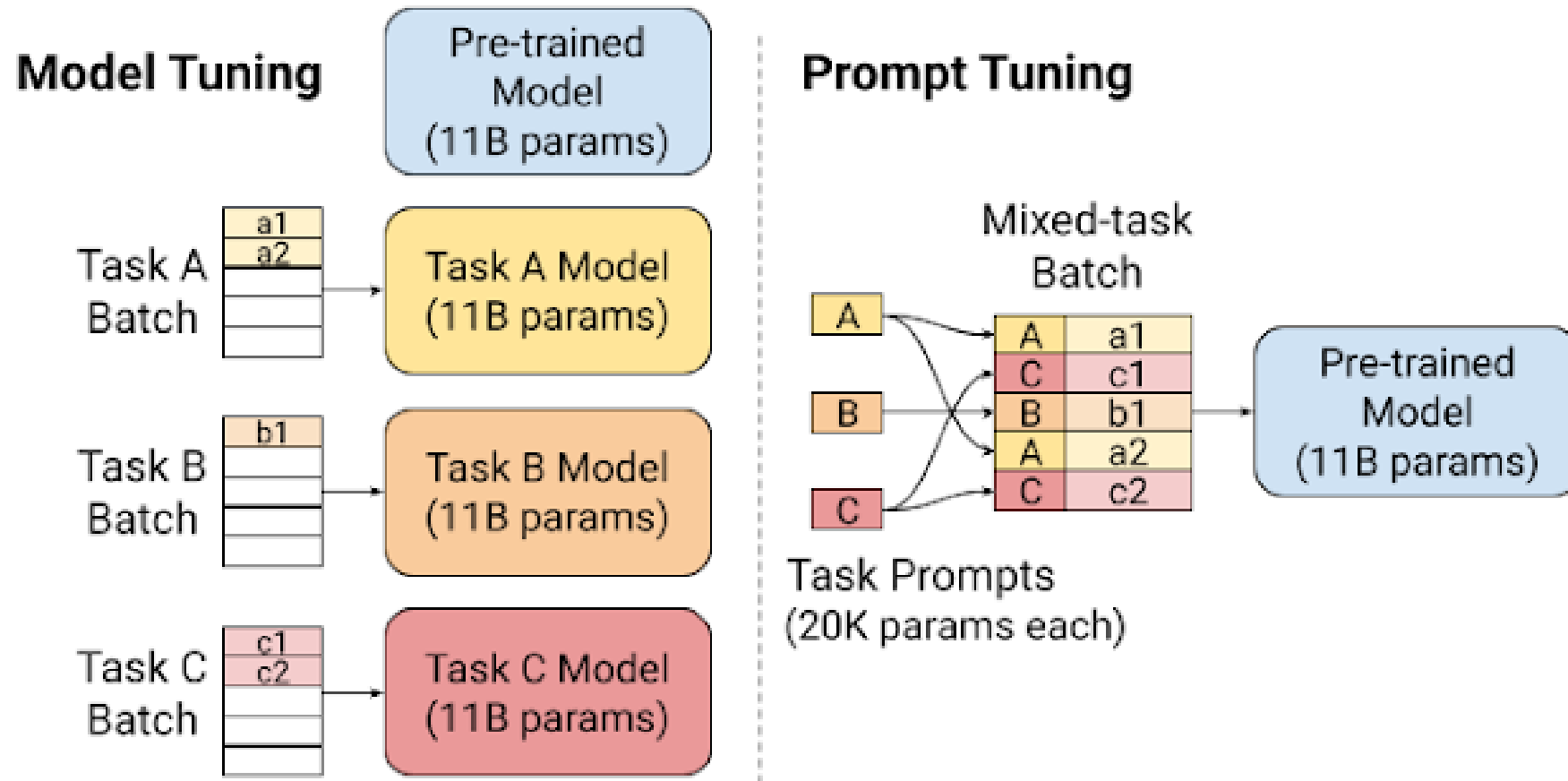
- Practically, switching modules without affecting the rest.
- Speculatively, on the path toward Artificial General Intelligence?
- Maybe modularity only makes sense when the modules scale up.

LPLMs: Learning Soft Prompts



Typically about 100 words, each having about 1024 dimensions.

LPLMs: Learning Soft Prompts

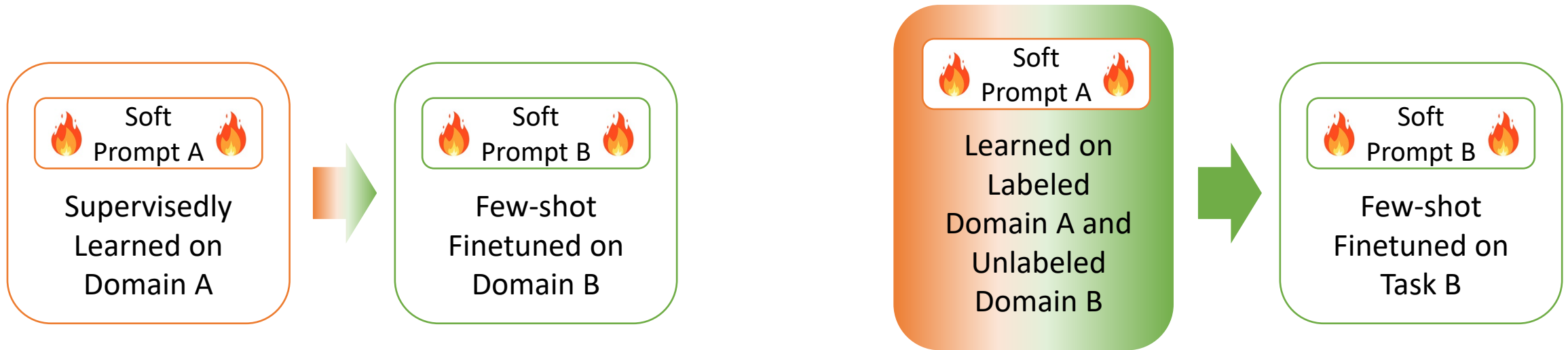


- However, prompt tuning requires a large number of training examples (Su et al., 2021).
- Its performance under few-shot learning is not as good as full-model finetuning.

How can we improve the sample efficiency of prompt tuning?

Xu Guo, Boyang Li, and Han Yu. Improving the Sample Efficiency of Prompt Tuning with Domain Adaptation. EMNLP Findings 2022.

Improving the Sample Efficiency of Prompt Tuning with Domain Adaptation



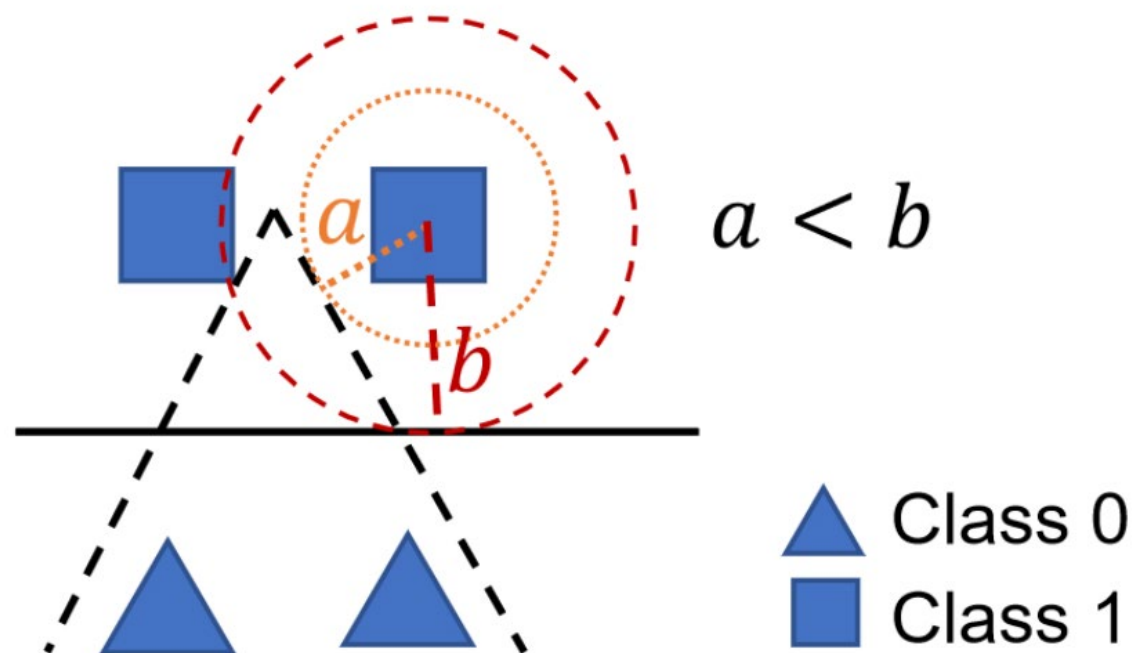
Transfer Learning for Prompts (Gu et al., 2022)

We propose **Boosting Prompt Tuning with Domain Adaptation (OPTIMA)**

OPTIMA



- Adversarial Training
 1. We find a small perturbation δ to the input that causes the network to change prediction.
 2. With that perturbation fixed, we train the network to predict the correct label.
 3. This leads to a decision boundary passing through regions with low data density



OPTIMA



- Adversarial Domain Similarity
 - We only care about perturbation vectors in the regions where the two domains are similar.

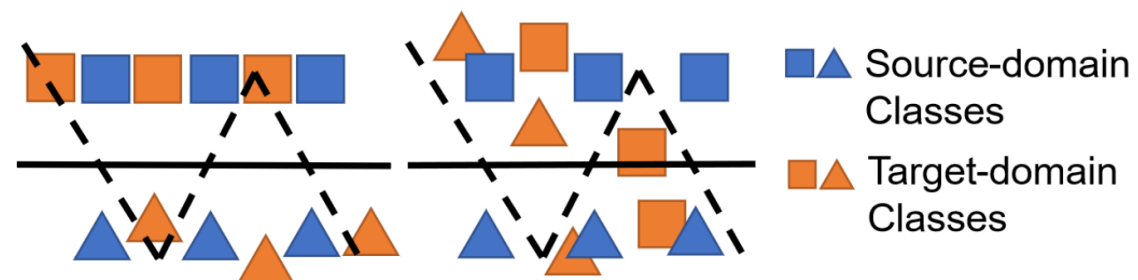


Figure 1: Smooth vs. zigzag decision boundaries. Left: When the distribution of the target-domain data (orange) are similar to the source domain (blue), the smooth decision boundary (solid line) generalizes better than the zigzag boundary. Right: When the distributions are different, it is not clear if the smooth decision boundary is the better choice.

Few-shot Results



Paper

Method	Params	PLM	Source	QQP		MRPC		MNLI
				Acc.	F1	Acc.	F1	Acc.
Frozen	0		✗	45.5	54.9	33.8	11.8	41.7
PT	102K		✗	48.4 ± 4.9	52.5 ± 5.5	53.1 ± 11.4	55.9 ± 23.4	33.4 ± 1.6
FT	770M	T5-Large	✗	55.1 ± 6.7	52.0 ± 6.0	<u>59.5</u> ± 7.8	<u>67.9</u> ± 12.6	35.6 ± 2.4
PFT	770M		✗	<u>55.1</u> ± 5.1	<u>57.8</u> ± 3.1	58.9 ± 11.0	65.3 ± 11.8	35.6 ± 3.6
PPT	410K	T5-XXL	✓	52.1 ± 11.1	56.2 ± 21.1	52.1 ± 11.1	56.2 ± 21.1	34.4 ± 1.4
				MRPC → QQP		QQP → MRPC		SNLI → MNLI
				Acc.	F1	Acc.	F1	Acc.
SPOT	102K		✓	64.5 ± 2.7	64.5 ± 0.8	68.7 ± 2.5	77.1 ± 2.9	74.3 ± 0.9
FreeLB	102K	T5-Large	✓	65.0 ± 2.4	64.5 ± 1.5	68.5 ± 2.2	77.6 ± 2.2	75.0 ± 1.0
VAT	102K		✓	66.2 ± 2.0	64.9 ± 0.7	69.6 ± 1.9	79.0 ± 2.1	74.9 ± 1.1
DANN	102K		✓	63.4 ± 2.5	62.5 ± 2.7	68.0 ± 3.5	76.2 ± 5.1	73.1 ± 1.4
OPTIMA	102K		✓	69.1* ± 1.7	65.8* ± 1.9	71.2* ± 1.7	79.9* ± 1.7	78.4* ± 0.6

Few-shot Results



Paper

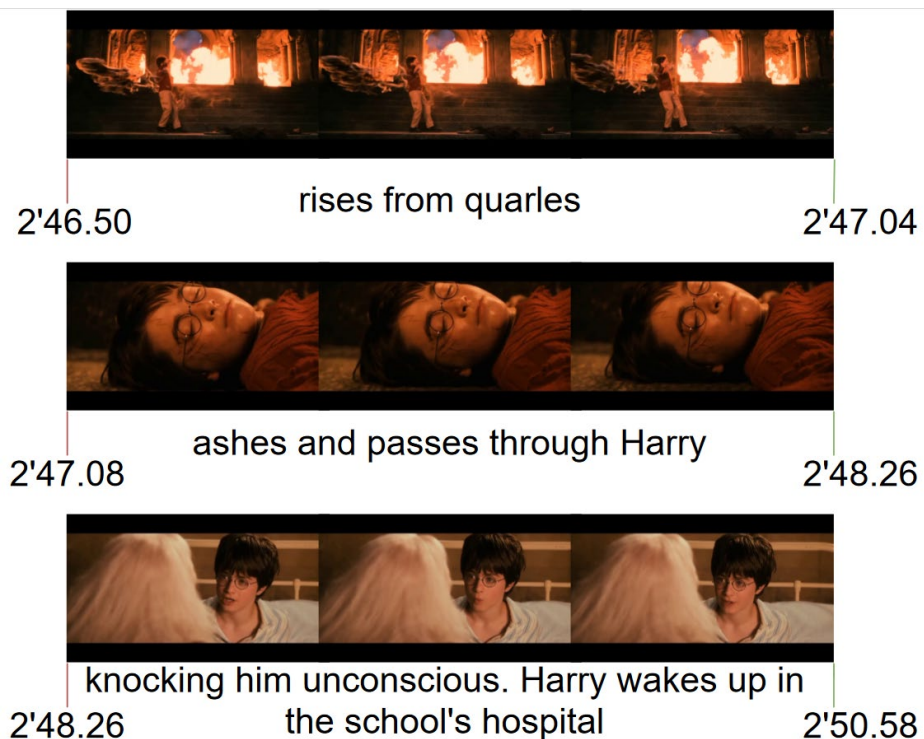
Method	Params	PLM	Source	SNLI Acc.	SICK Acc.		CB Acc.	
Frozen	0		✗	35.9	37.1		55.4	
PT	102K		✗	34.6 ± 2.4	61.5 ± 7.8		38.3 ± 13.6	
FT	770M	T5-Large	✗	<u>41.6</u> ± 3.8	67.6 ± 6.3		51.2 ± 7.8	
PFT	770M		✗	38.6 ± 5.1	<u>71.3</u> ± 6.4		<u>57.3</u> ± 9.2	
PPT	410K	T5-XXL	✓	34.7 ± 2.8	54.6 ± 14.0		43.0 ± 14.6	
				MNLI → SNLI Acc.	SNLI → SICK Acc.	MNLI → SICK Acc.	SNLI → CB Acc.	MNLI → CB Acc.
SPOT	102K		✓	78.8 ± 1.1	69.9 ± 5.3	72.9 ± 5.9	61.7 ± 5.0	65.3 ± 3.4
FreeLB	102K	T5-Large	✓	81.5 ± 0.7	69.5 ± 6.8	73.1 ± 4.8	61.6 ± 4.2	66.1 ± 3.3
VAT	102K		✓	80.9 ± 0.9	68.6 ± 6.4	72.7 ± 6.3	59.0 ± 5.5	68.7 ± 4.8
DANN	102K		✓	71.1 ± 3.2	69.0 ± 6.7	73.4 ± 3.7	55.7 ± 5.5	66.9 ± 4.6
OPTIMA	102K		✓	82.1* ± 0.8	73.3 ± 6.8	74.8 ± 4.4	64.8* ± 1.1	71.2* ± 3.1

Source-domain & Zero-shot Results



Method	MRPC	MRPC \rightarrow QQP		QQP	QQP \rightarrow MRPC		MNLI \rightarrow CB
	Acc.	Acc.	F1	Acc.	Acc.	F1	Acc.
SPOT	82.5 \pm 1.5	60.9 \pm 4.6	63.6 \pm 2.0	80.9 \pm 2.2	65.7 \pm 3.4	73.2 \pm 5.7	63.2 \pm 5.7
FreeLB	85.5 \pm 0.3	63.1 \pm 3.7	63.9 \pm 1.0	82.2 \pm 2.7	69.4 \pm 1.1	78.7 \pm 1.3	67.8 \pm 3.9
VAT	84.7 \pm 0.8	64.8 \pm 4.6	64.1 \pm 1.7	81.9 \pm 0.7	68.9 \pm 1.5	78.5 \pm 1.5	67.8 \pm 5.8
DANN	81.5 \pm 2.1	63.9 \pm 1.8	57.6 \pm 3.3	81.4 \pm 0.7	63.6 \pm 4.8	71.5 \pm 9.7	59.8 \pm 4.4
OPTIMA	85.7 \pm 0.7	68.9 \pm 0.8	66.3 \pm 0.6	82.7 \pm 1.3	71.2 \pm 0.4	80.0 \pm 0.6	68.3 \pm 2.6
Method	MNLI	MNLI \rightarrow SNLI	MNLI \rightarrow SICK	SNLI	SNLI \rightarrow MNLI	SNLI \rightarrow SICK	SNLI \rightarrow CB
	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.
SPOT	83.4 \pm 0.8	79.2 \pm 1.0	51.8 \pm 0.7	88.9 \pm 0.1	75.6 \pm 0.4	52.7 \pm 1.9	47.6 \pm 3.7
FreeLB	84.8 \pm 0.8	81.8 \pm 0.7	52.2 \pm 0.2	89.9 \pm 0.1	77.5 \pm 0.5	52.9 \pm 1.9	47.5 \pm 4.7
VAT	83.7 \pm 0.3	81.0 \pm 0.2	51.4 \pm 1.4	88.7 \pm 0.1	77.1 \pm 1.3	51.8 \pm 2.1	45.8 \pm 0.8
DANN	80.4 \pm 2.7	72.4 \pm 5.9	61.9 \pm 2.7	85.3 \pm 3.2	70.3 \pm 3.6	51.5 \pm 1.2	42.3 \pm 2.2
OPTIMA	84.6 \pm 0.3	82.1 \pm 0.8	55.2 \pm 1.0	89.2 \pm 0.1	79.1 \pm 0.1	53.8 \pm 0.5	49.4 \pm 4.2

New Dataset: Synopses of Movie Narratives



- “Watch a movie in 5 minutes” videos
- 869 hours, 683,611 sentences
- Texts are not literal descriptions
 - Mental state descriptions
 - Reporting bias
 - 16-40% text are well matched with videos using near-SOTA (2020) models
- Calls for
 - Understanding: events, causal relations, theory of mind, long-term identity tracking, etc.
 - Commonsense reasoning

Conclusions

- Large Pretrained Language Models are transforming AI
- We design systems that
 - Exploit new capabilities (language-based reasoning)
 - Solve new challenges (few-shot prompt tuning)
- We propose a new dataset that poses greater challenges to these models

Contact: Boyang “Albert” Li, boyangli@ntu.edu.sg