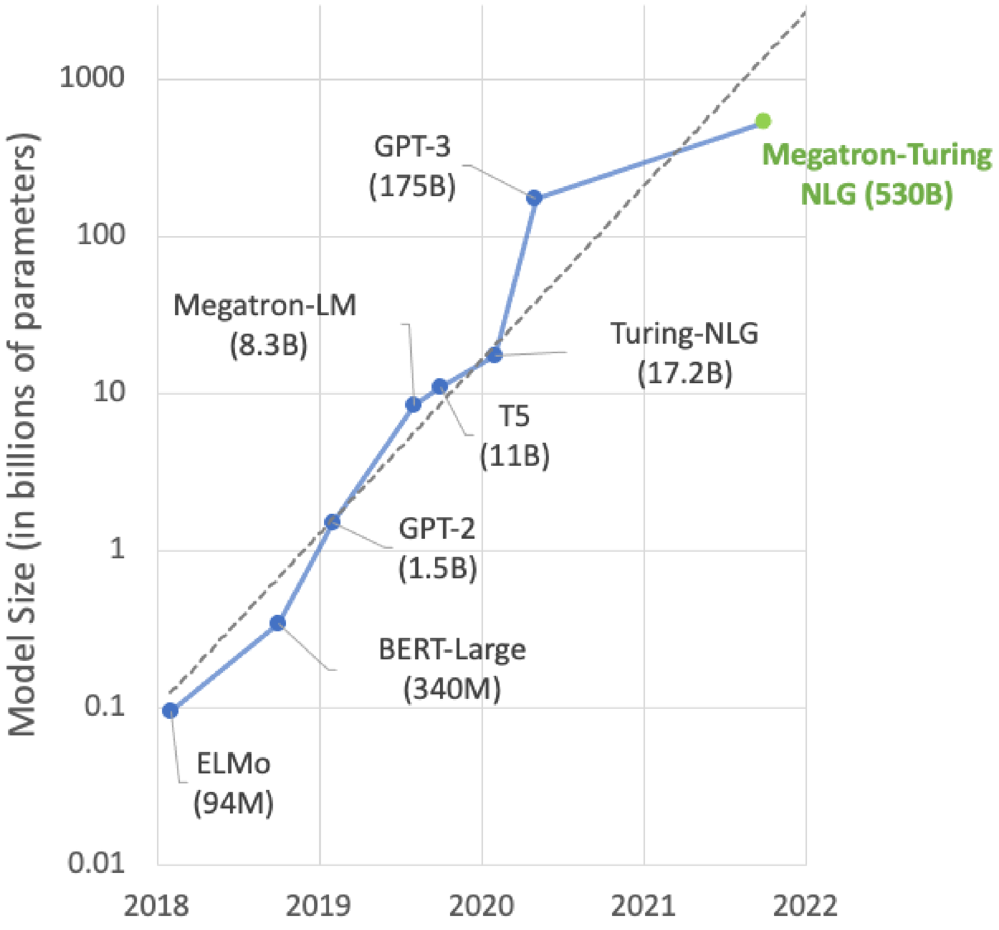
Multimodal Learning in the Era of Gigantic Pretrained Models

Boyang "Albert" Li
Nanyang Associate Professor
Nanyang Technological University

The University of British Columbia. June 26, 2023

Era of Large Language Models (LLMs)



Model Name	Year	# Parameters
T0	2021	11B
LaMDA	2021	137B
InstructGPT	2022	175B
GPT-NeoX	2022	20B
OPT	2022	175B
PaLM	2022	**540B**
GLM-130B	2022	130B
BLOOM	2022	176B
Galactica	2022	120B
ChatGPT	2022	**1760B**

🤗 Open LLM Leaderboard

As of June 25, 2023

Model	Average ⬆️	ARC (25-s) ⬆️	HellaSwag (10-s) ⬆️	MMLU (5-s) ⬆️	TruthfulQA (MC) (0-s
tiiuae/falcon-40b-instruct	63.2	61.6	84.4	54.1	52.5
timdettmers/guanaco-65b-merged	62.2	60.2	84.6	52.7	51.3
CalderaAI/30B-Lazarus	60.7	57.6	81.7	45.2	58.3
tiiuae/falcon-40b	60.4	61.9	85.3	52.7	41.7
timdettmers/guanaco-33b-merged	60	58.2	83.5	48.5	50
ausboss/llama-30b-supercot	59.8	58.5	82.9	44.3	53.6
huggyllama/llama-65b	58.3	57.8	84.2	48.8	42.3
pinkmanlove/llama-65b-hf	58.3	57.8	84.2	48.8	42.3
llama-65b	58.3	57.8	84.2	48.8	42.3
MetaIX/GPT4-X-Alpasta-30b	57.9	56.7	81.4	43.6	49.7
Aeala/VicUnlocked-alpaca-30b	57.6	55	80.8	44	50.4
digitous/Alpacino30b	57.4	57.1	82.6	46.1	43.8

Broad Competence

Acing human exams

“Unparalleled mastery of natural language”

Sparks of AGI

Severe Hallucination

Can’t do simple math

Yann LeCun:
Nobody will be interested in LLMs in 5 years

Image created by Midjourney v5.2

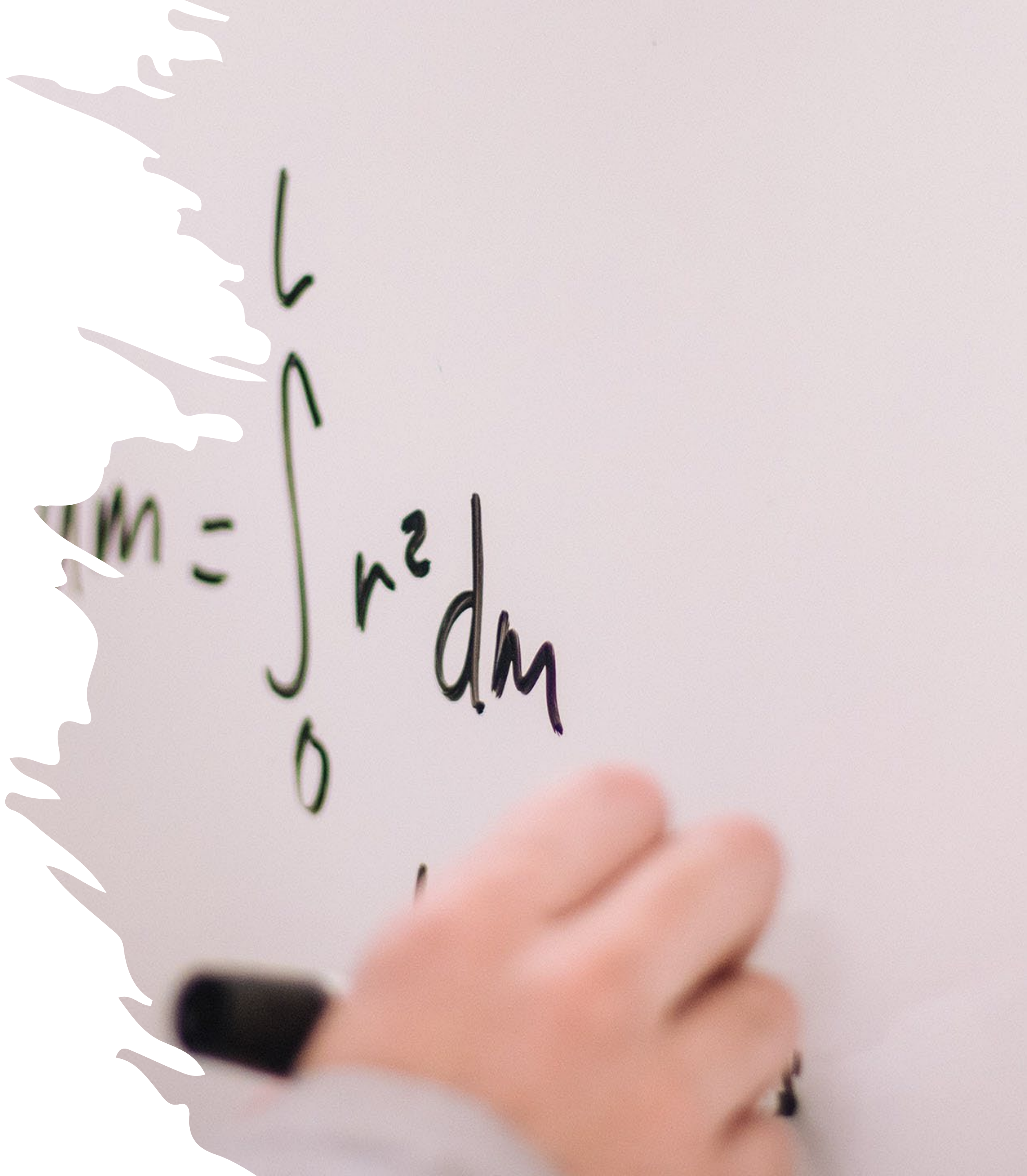
How do we think about LLMs?

- A different type of general intelligence from humans
  - Therefore, hard to understand
  - Implicit anthropomorphic thinking is a common pitfall
- A lot of memorization and pattern matching
  - Huge input/output bandwidth
  - Sufficient to compensate for the lack of reasoning
  - No sense of humor (Jentzsch and Kersting, 2023)
  - Solving compositional problems using memorization (Dziri et al. 2023)

Sophie Jentzsch and Kristian Kersting. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. arXiv 2306.04563. 2023

Dziri et al. Faith and Fate: Limits of Transformers on Compositionality. 2023

Intermittent Performance and Prompt Brittleness are Consistent with Memory-based Generalization



A Gigantic Treasure Box in Need of Keys

Image created by Midjourney v5.2

Keys to Unlock LLM Capabilities

- Chain-of-thought Prompting (Wei et al. 2022)
- Let's think step by step (Kojima et al. 2022)
- Instruction Tuning (FLAN by Wei et al. 2021; T0 by Sanh et al. 2021; InstructGPT by Ouyang et al. 2022)
- And so on…

- But the content of the treasure box is not easily simulated (Gudibande et al. 2023)

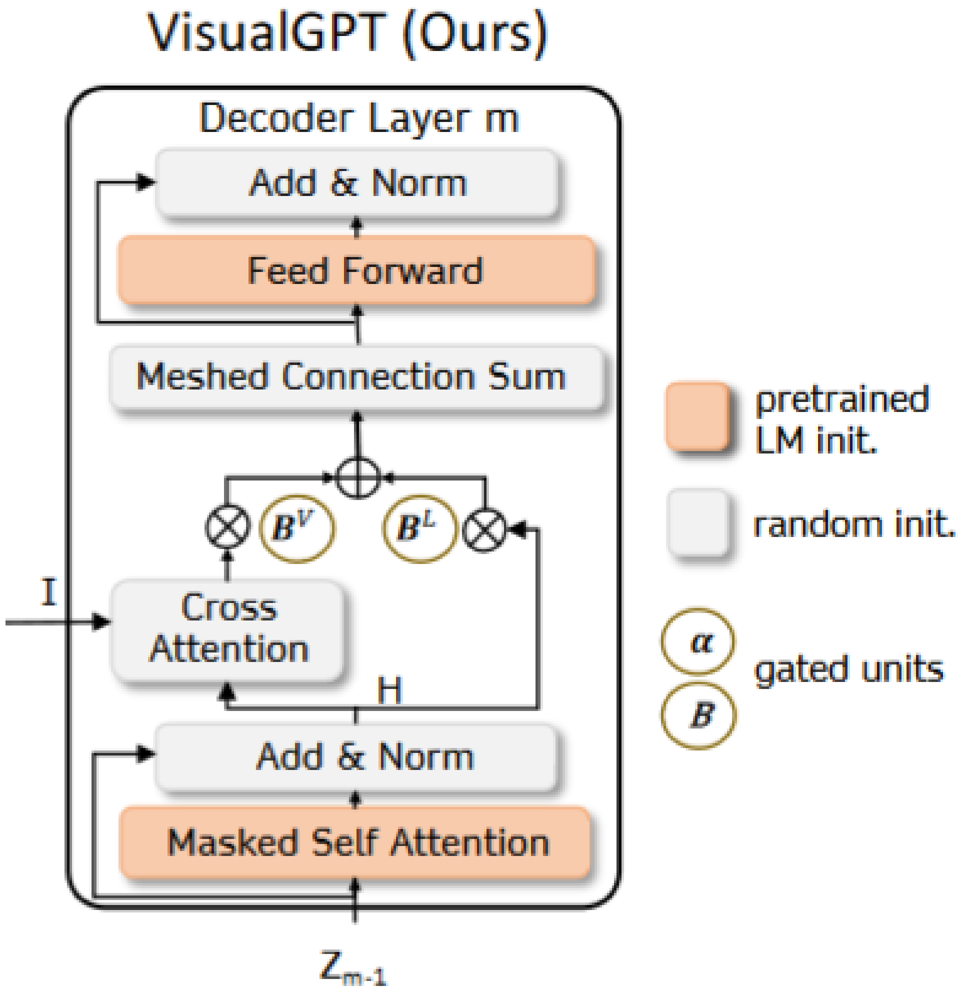
Gudibande et al. The False Promise of Imitating Proprietary LLMs. arXiv 2305.15717. 2023.

Leveraging LLMs for Multimodal Purposes

VisualGPT (2021)

Jun Chen, Han Guo, Kai Yi, **Boyang Li**, and Mohamed Elhoseiny. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. arXiv 2102.10407. 2021.

- One of the early works for adapting pretrained LLMs for multimodal tasks



InstructBLIP (2023)

Wenliang Dai, Junnan Li, Dongxu Li, Anthony M. H. Tiong, Junqi Zhao, Weisheng Wang, **Boyang Li**, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv 2305.06500

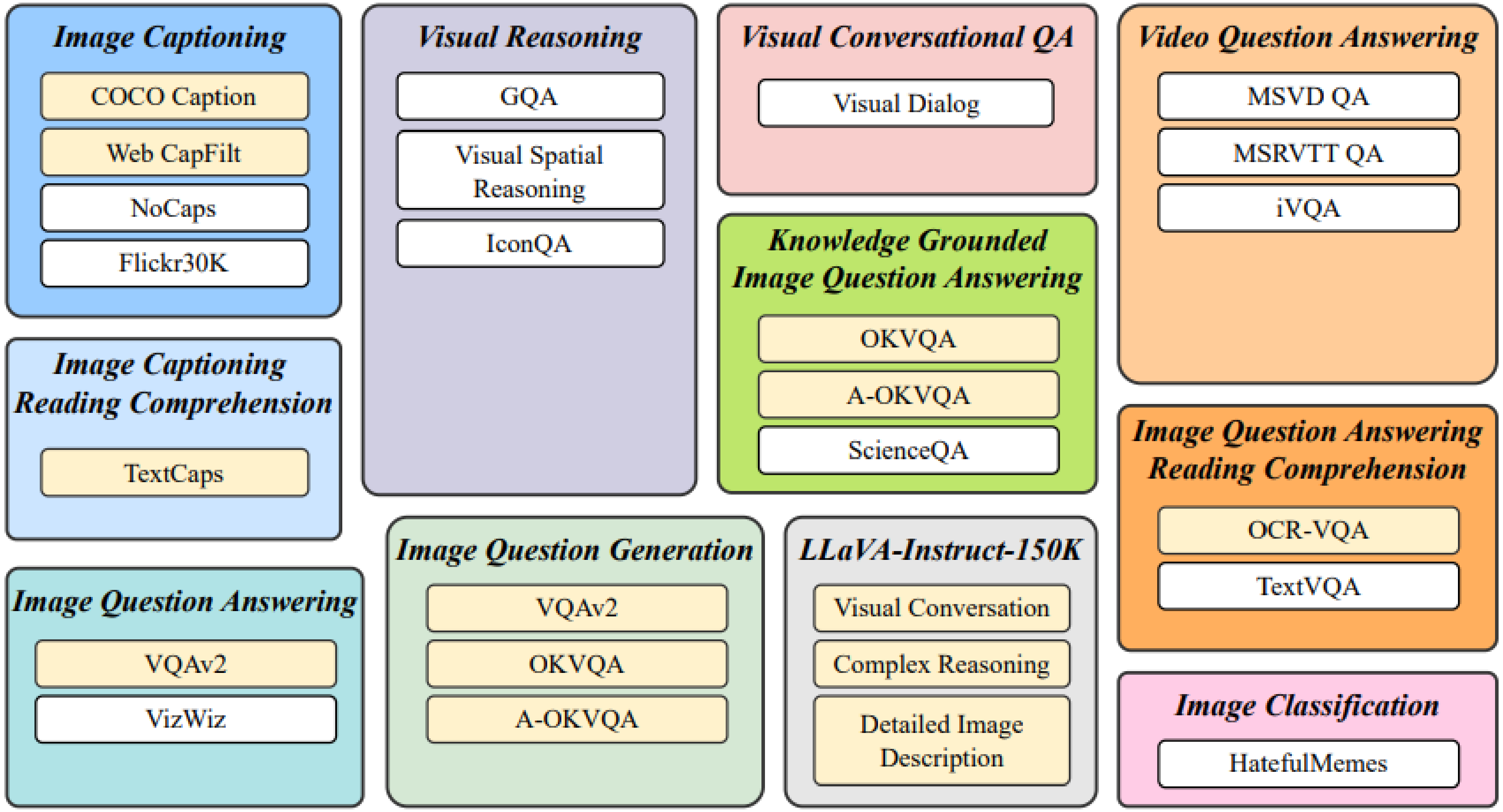


Figure 2: Tasks and their corresponding datasets used for vision-language instruction tuning. The held-in datasets are indicated by yellow and the held-out datasets by white.

InstructBLIP (2023)

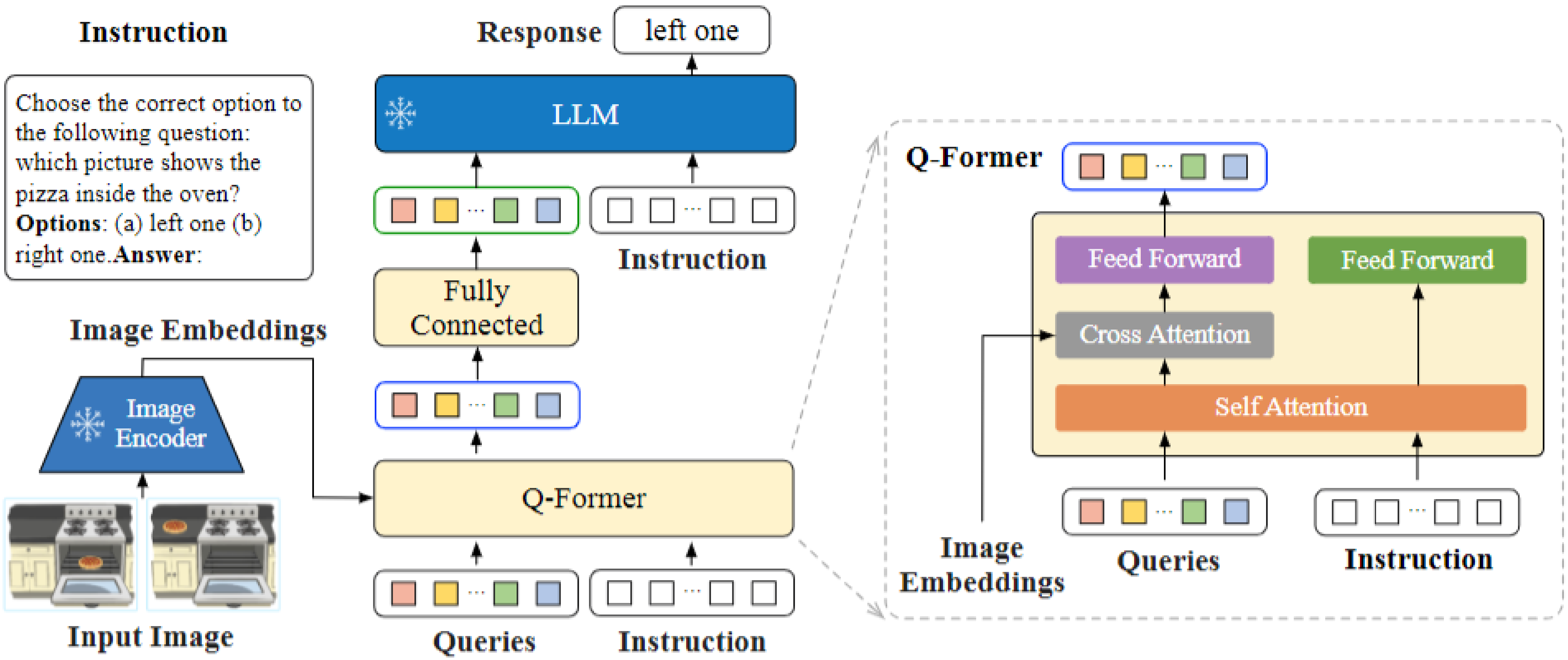


Figure 3: Model architecture of InstructBLIP. The Q-Former extracts instruction-aware visual features from the output embeddings of the frozen image encoder, and feeds the visual features as soft prompt input to the frozen LLM. We instruction-tune the model with the language modeling loss to generate the response.


Model Finetuning


Model Deployment

How to acquire new multimodal capabilities without finetuning?

We demonstrate a system for visual question answering.

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

Examples from VQAv2 (Goyal et al. 2017)

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.

Paper

- Conventional wisdom suggests that in order to connect pretrained models, end-to-end training is necessary.
- We connect pretrained models 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

Pretrained Modules

Paper

BLIP (Li et al, 2022)

Image-Question Matching Module

Image Captioning Module

UnifiedQA v2 (Khashabi et al. 2022)

Question-Answering Module

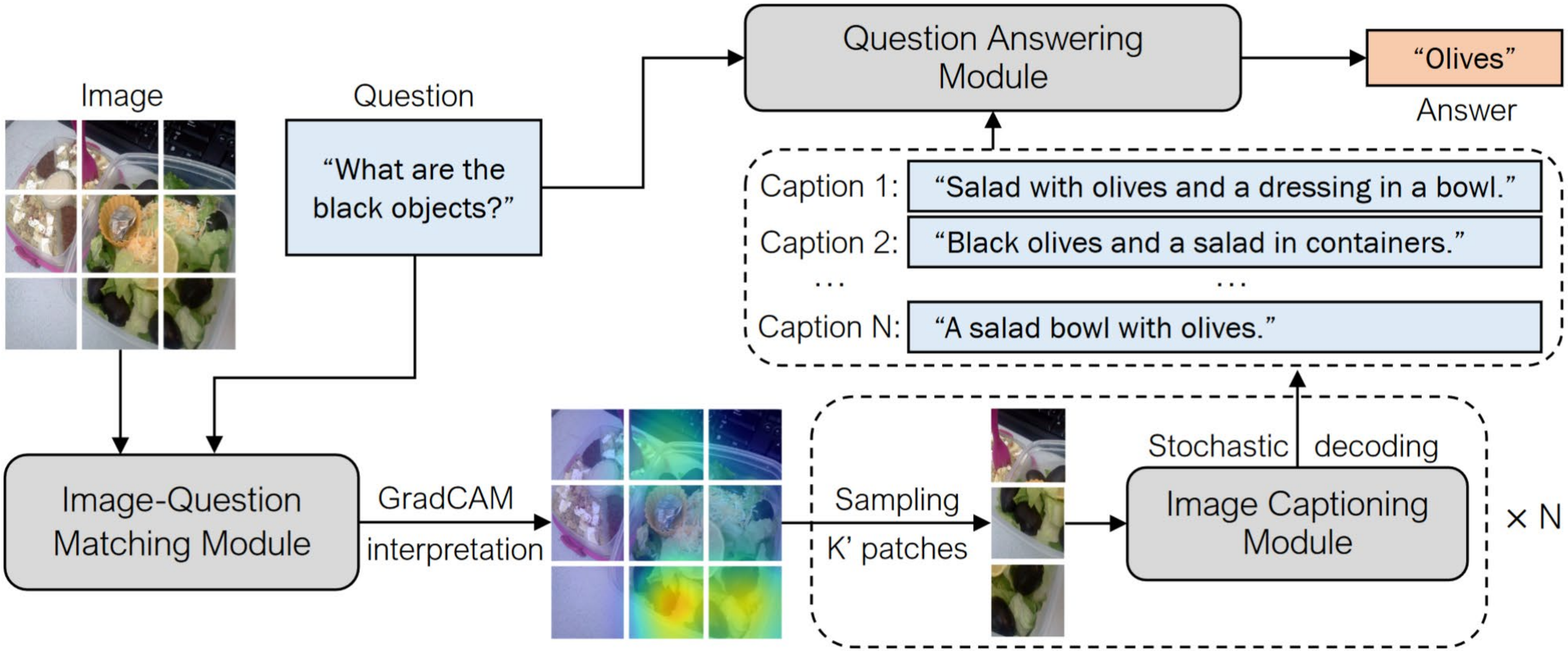
Pretrained to classify an image-caption pair as Matching or Not Matching.

Pretrained to write a caption for an image, which consists of 14x14 image patches.

Pretrained to perform textual question answering.

System Architecture

Paper



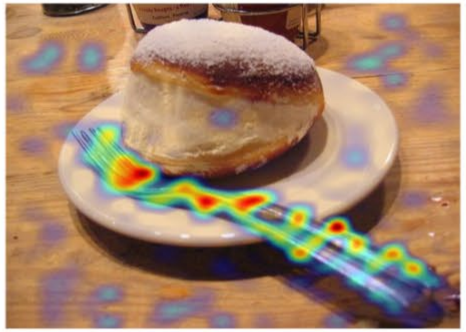
Case Studies

Paper

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

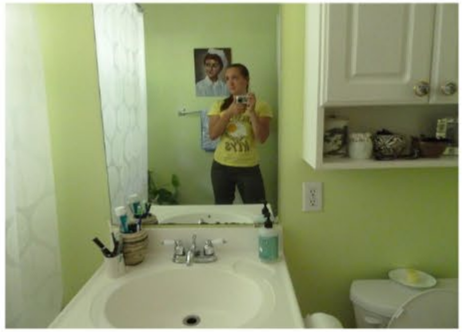
Prediction: a spoon



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

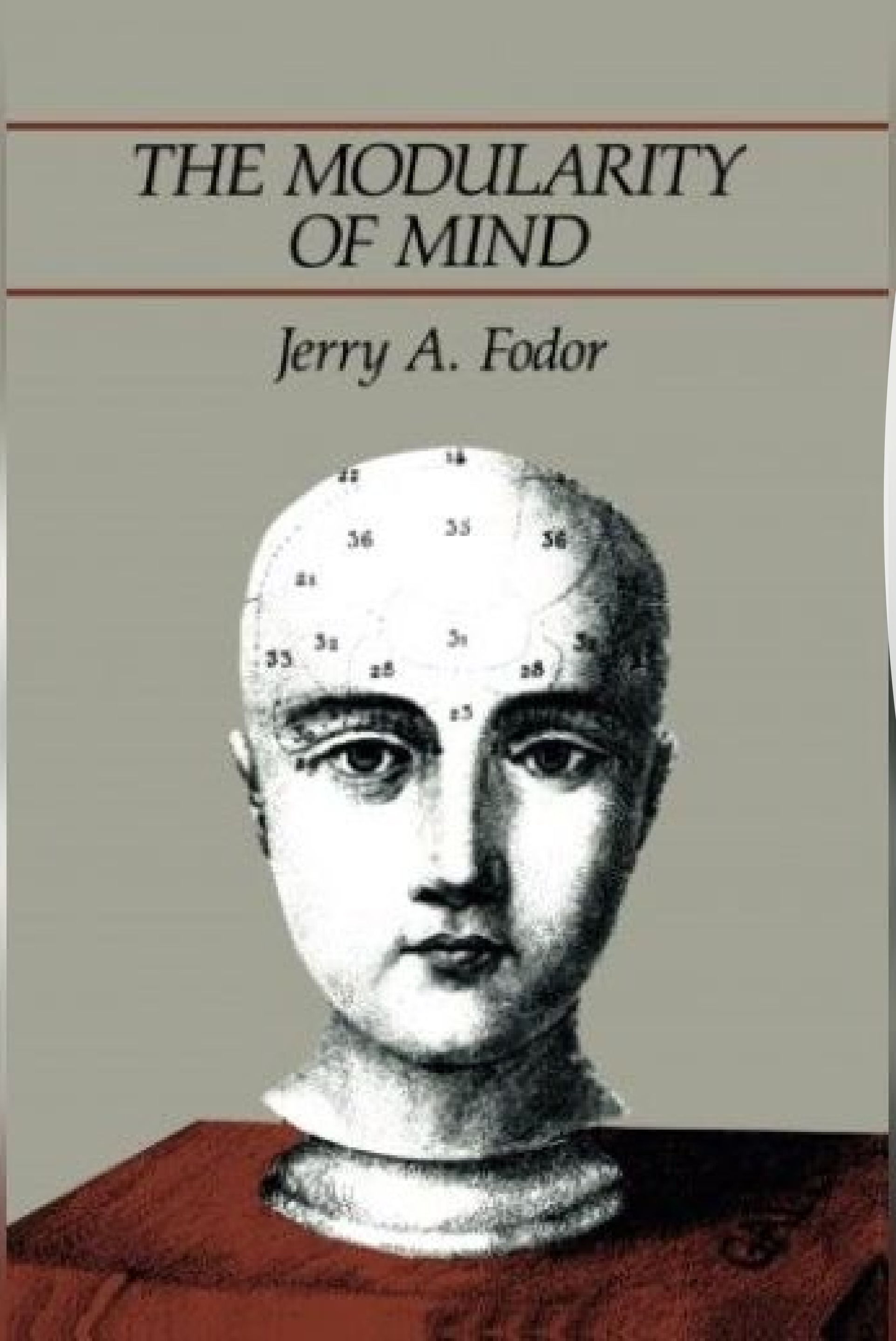
Prediction: self-portrait

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

Method	Language Model	Language #Params	Language VL-aware	Vision Model	Vision #Params	Vision VL-aware	VQAv2 Val	VQAv2 Test-dev	OK-VQA Test	GQA Test-dev
Pretrained models conjoined by end-to-end VL training.										
VL-T5$_{\text{no-vqa}}$	T5	224M	✓	Faster R-CNN	64M	✗	13.5	-	5.8	6.3
FewVLM$_{\text{base}}$	T5	224M	✓	Faster R-CNN	64M	✗	43.4	-	11.6	27.0
FewVLM$_{\text{large}}$	T5	740M	✓	Faster R-CNN	64M	✗	47.7	-	16.5	29.3
VLKD$_{\text{ViT-B/16}}$	BART	407M	✓	ViT-B/16	87M	✓	38.6	39.7	10.5	-
VLKD$_{\text{ViT-L/14}}$	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	-
Pretrained models conjoined by natural language and zero training.										
PICa	GPT-3	175B	✗	VinVL-Caption	259M	✓	-	-	17.7	-
PnP-VQA$_{\text{base}}$	UnifiedQAv2	223M	✗	BLIP-Caption	446M	✓	54.3	55.2	23.0	34.6
PnP-VQA$_{\text{large}}$	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>

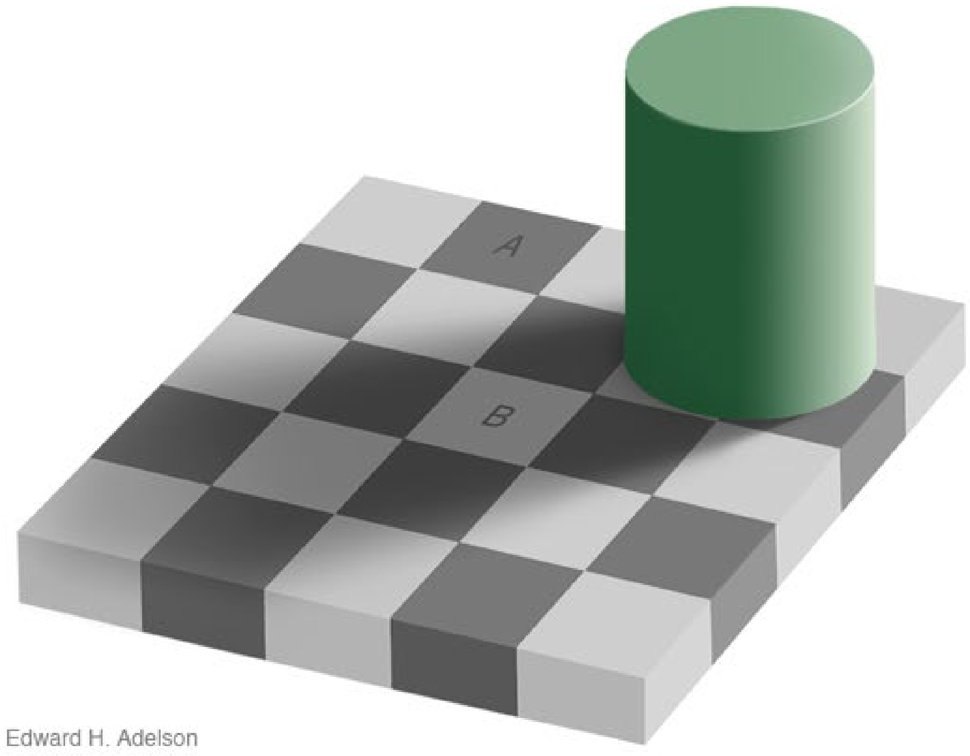
Table 2: Comparison with state-of-the-art models on zero-shot VQA. Flamingo (Alayrac et al., 2022) inserts additional parameters into the language model and perform training using billion-scale vision-language data. The best accuracy is bolded and the second best is underlined.

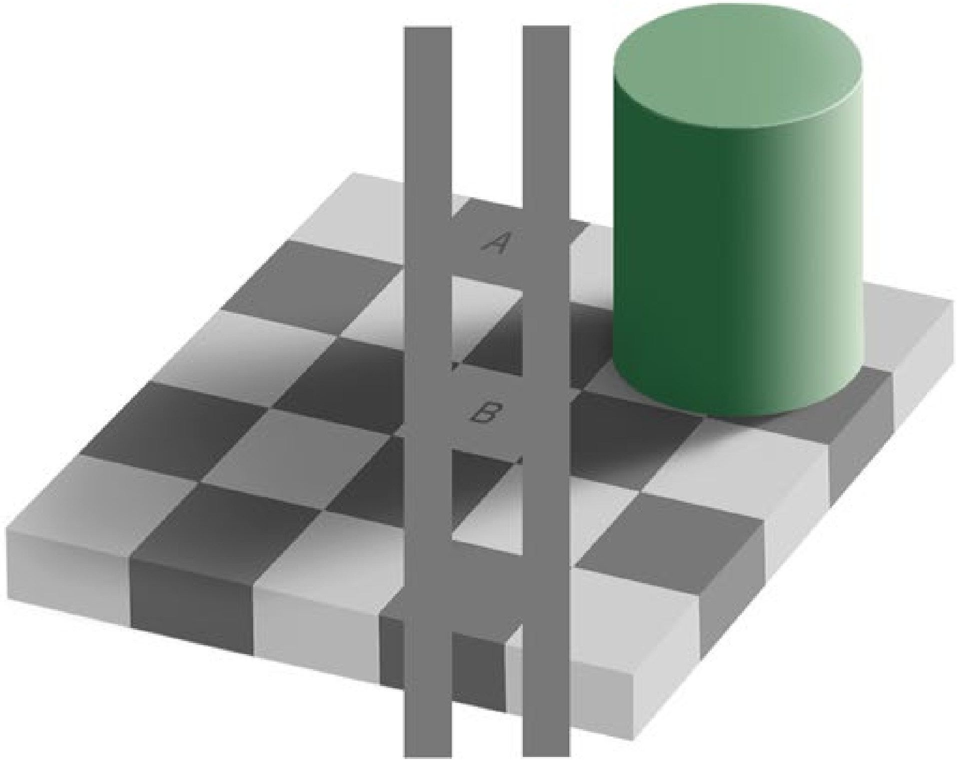


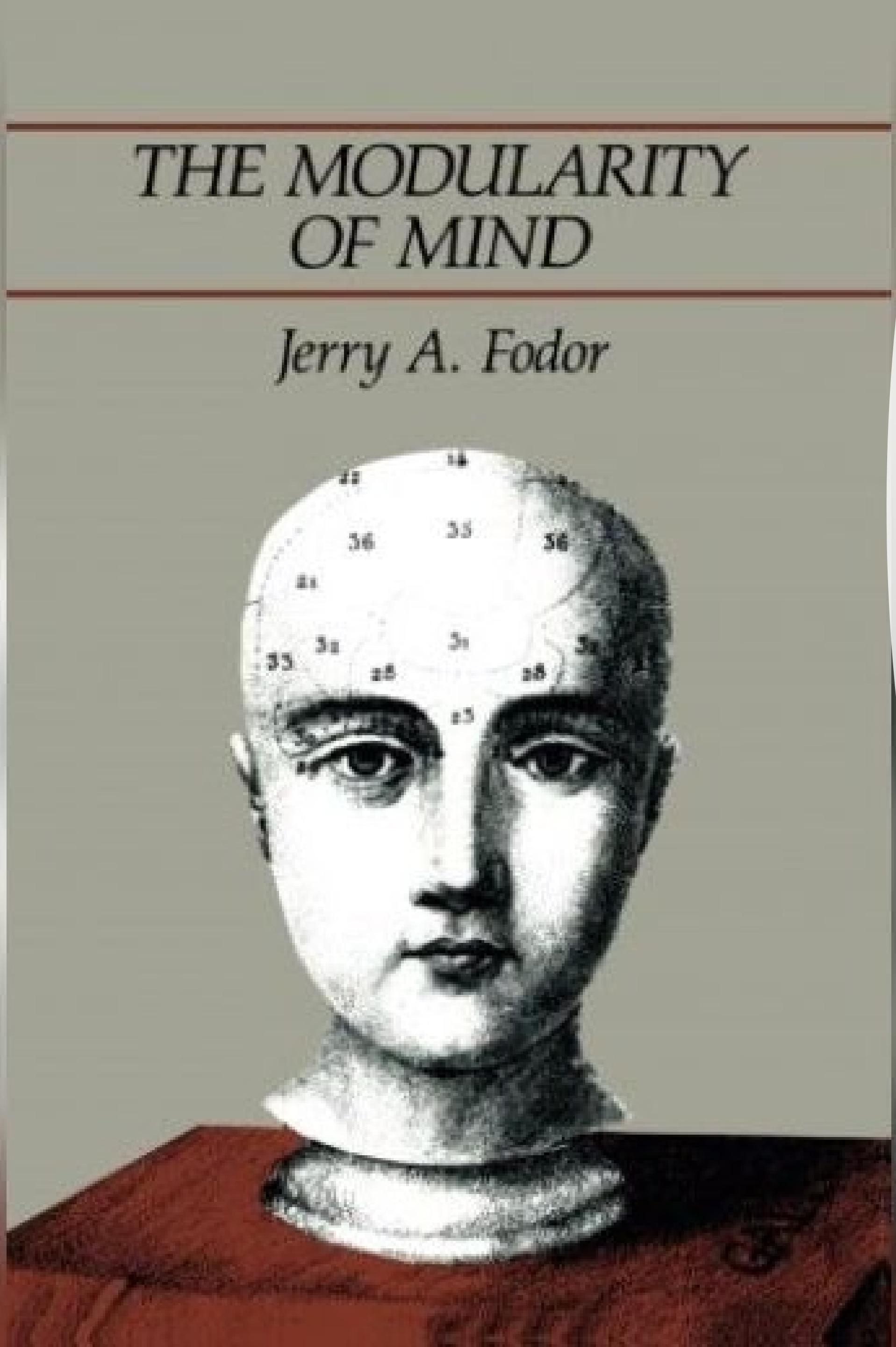
Modular System Design?

- Modularity in the human mind.
- End-to-end training is the go-to option for machine learning

Perceptive Modules are Encapsulated







Modular System Design?

- Modularity in the human mind.
- End-to-end training is the go-to option for machine learning
- Maybe modularity only makes sense when the modules scale up.

From QA Models to Generic Models?

Paper

- Need to demonstrate the QA task to generic models
- We generate synthetic question / answers from the question-guided captions and include them in the context.

Question: The girl behind the man likely is of what relation to him?

GT Answer: daughter

Captions 1: a man is riding the back of a little girl on a motorcycle
Captions 2: an image of bearded man and a girl on a motorcycle riding on the motorcycle
Captions 3: man and child sitting on a motorcycle on the street

Synthetic Question 1: who is holding on to the bearded man on the back of the motorcycle?
Answer: A girl
Synthetic Question 2: what is the size of the girl riding on the motorcycle?
Answer: little
Question: The girl behind the man likely is of what relation to him?
Predicted Answer: daughter

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, Steven CH Hoi. From Images to Textual Prompts: Zero-shot VQA with Frozen Large Language Models. CVPR 2023

Synthetic Question-answer Pairs Generation

Paper

- We extract answers from the generated captions: nouns, verbs, adjectives, and numbers.
- To generate questions from answers, we finetune a T5-Large network.
- Or, we may use templates based on Parts-of-Speech.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, Steven CH Hoi. From Images to Textual Prompts: Zero-shot VQA with Frozen Large Language Models. CVPR 2023

From QA Models to Generic Models?

Paper

Models	VQA v2		OK-VQA
	Val	Test	Test
Frozen-7B	29.5		
Flamingo-80B		56.3	**50.6**
PnP-VQA-11B	**63.3**	**64.8**	35.9
Img2Prompt-175B	60.6	61.9	45.6

Table 3. Zero-shot VQA performance with different LLMs.

Methods	VQAv2 val	OK-VQA test
PICa $_{\text{GPT-3 175B}}$	-	17.7
Frozen $_{\text{7B}}$	29.5	5.9
Ours $_{\text{GPT-Neo 2.7B}}$	50.1	31.5
Ours $_{\text{BLOOM 7.1B}}$	52.4	32.4
Ours $_{\text{GPT-J 6B}}$	56.4	37.4
Ours $_{\text{OPT 6.7B}}$	57.6	38.2
Ours $_{\text{OPT 175B}}$	60.6	45.6

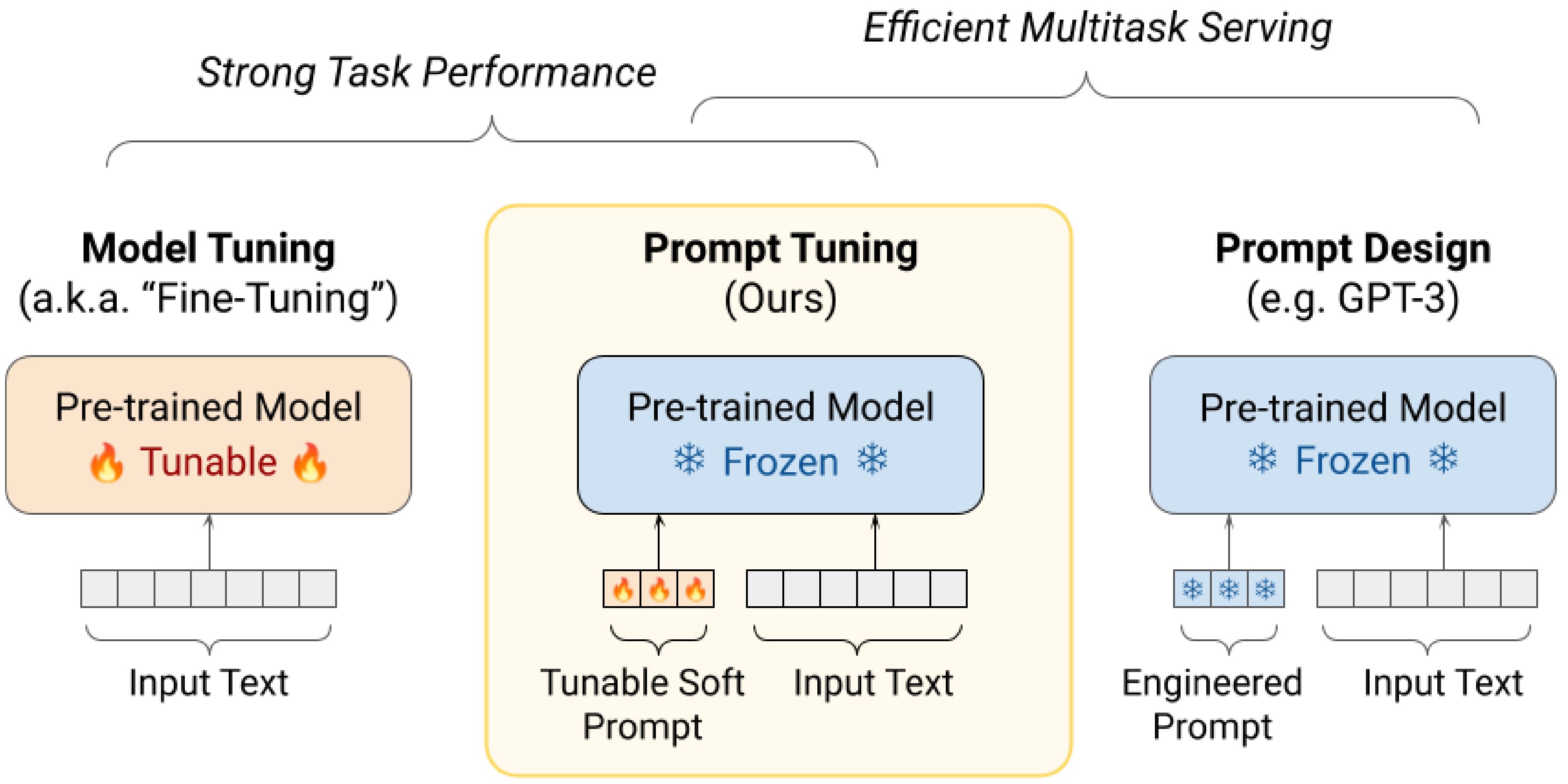
Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, Steven CH Hoi. From Images to Textual Prompts: Zero-shot VQA with Frozen Large Language Models. CVPR 2023

How to simplify deployment of large models?

Prompt tuning is friendly to deployment.

Prompt Tuning

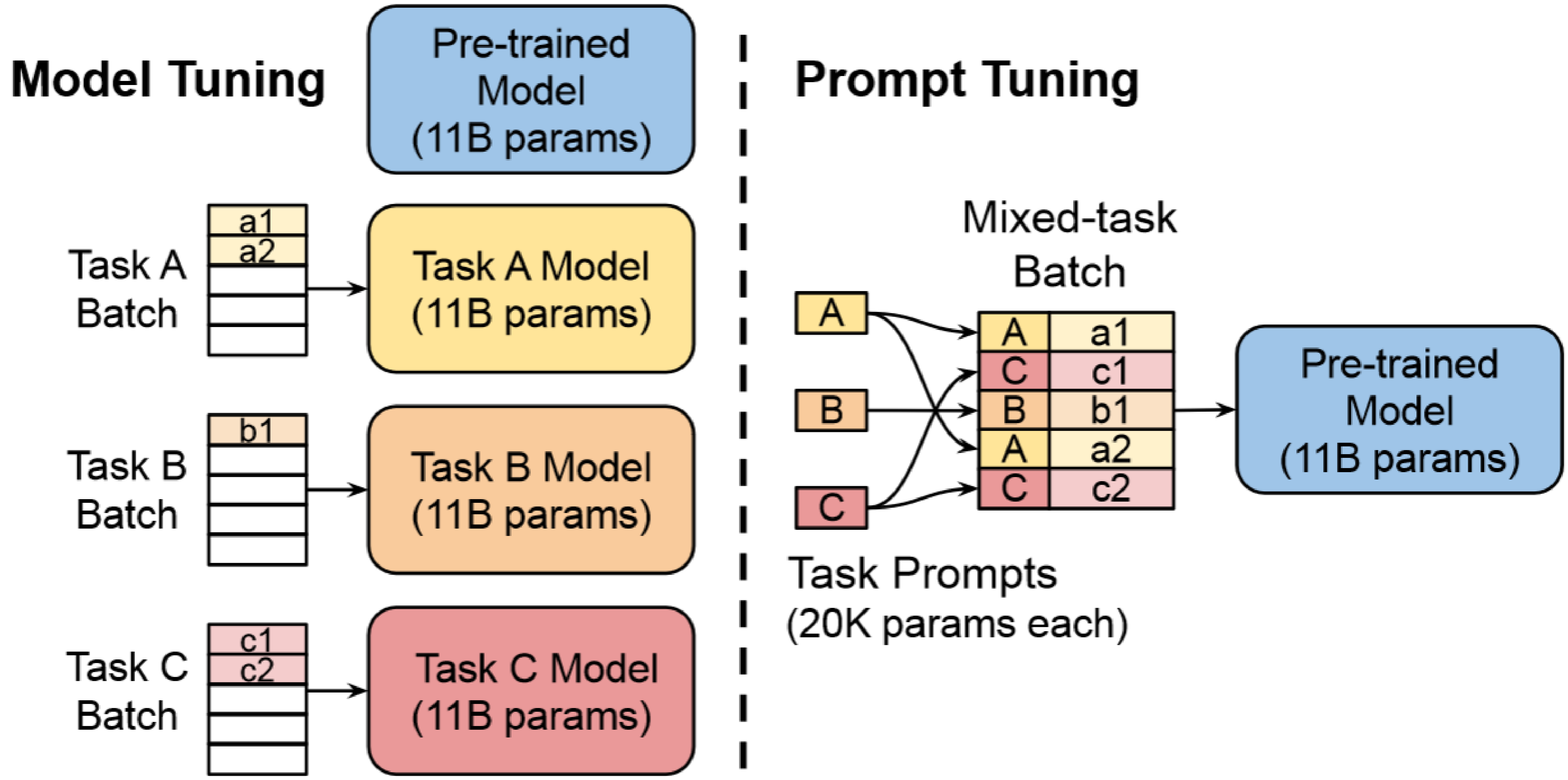
Brian Lester, Rami Al-Rfou, Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. EMNLP 2021



Typically about 100 words, each having about 1024 dimensions.

Prompt Tuning

Brian Lester, Rami Al-Rfou, Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. EMNLP 2021



- 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.

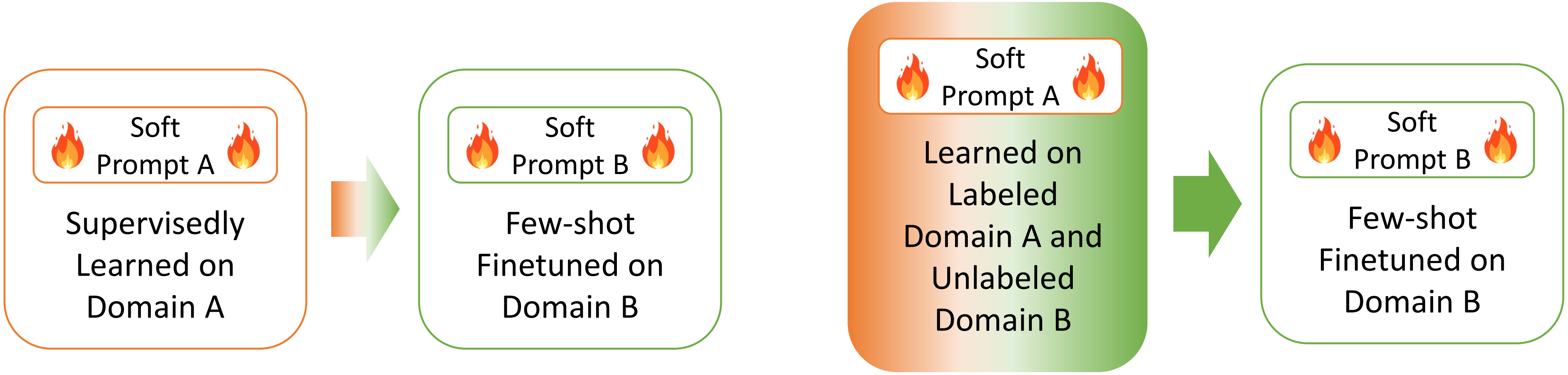
How can we improve the sample efficiency of prompt tuning?

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

Su et al. On Transferability of Prompt Tuning for Natural Language Processing. 2021

Improving the Sample Efficiency of Prompt Tuning with Domain Adaptation

Paper



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

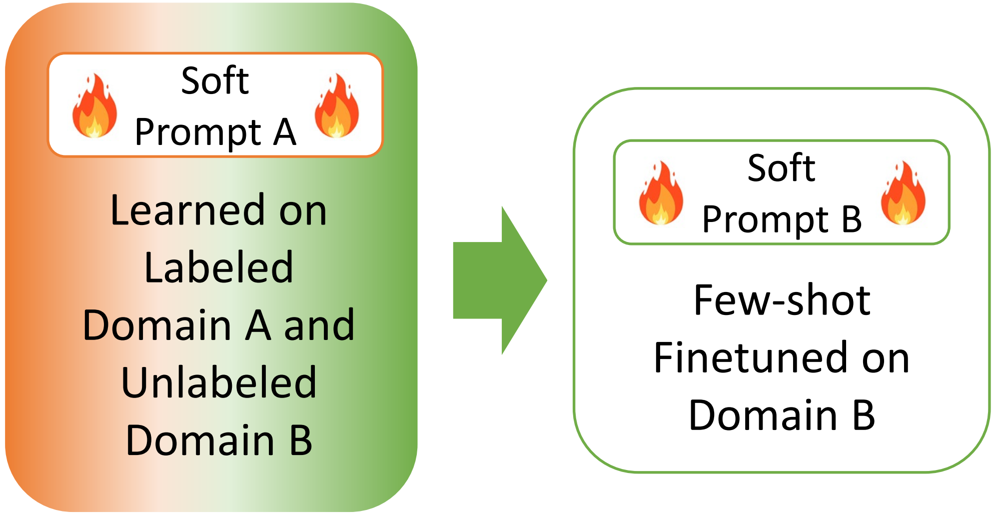
We propose **bOosting Prompt TunIng with doMain Adaptation (OPTIMA)**

Yuxian Gu, Xu Han, Zhiyuan Liu, Minlie Huang. PPT: Pre-trained Prompt Tuning for Few-shot Learning. 2022

OPTIMA: Intuition #1

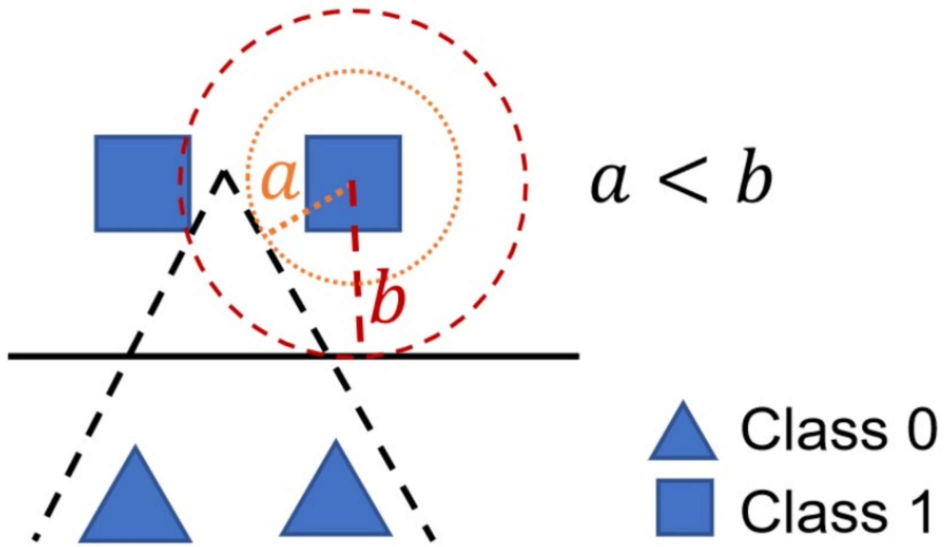
Paper

- The target domain has no labels.
- It is easy to overfit the source domain.
- Therefore, we need a smooth decision boundary



We propose **bOosting Prompt TunIng with doMain Adaptation (OPTIMA)**

Adversarial Training (Madry et al 2018)

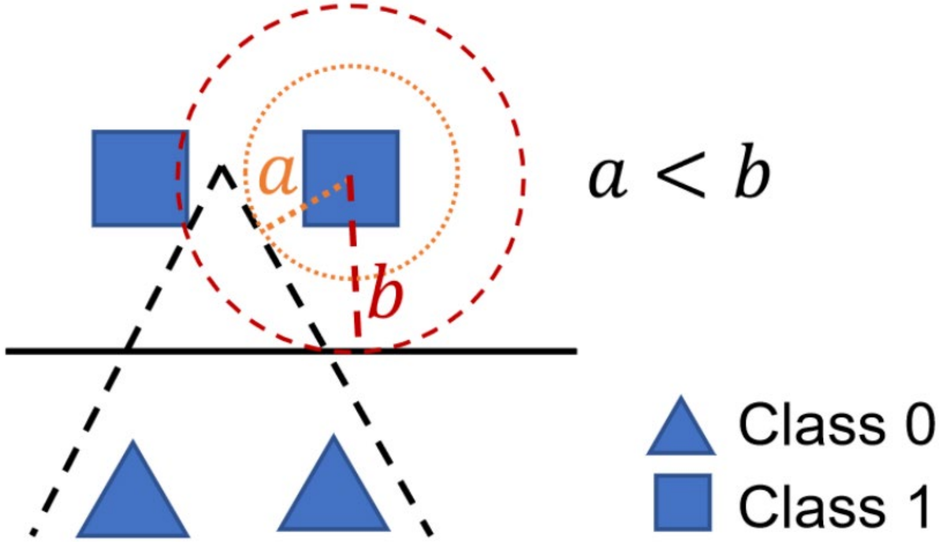


- Dotted decision boundary = non-smooth
- Solid decision boundary = smooth

Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. ICLR, 2018.

Adversarial Training

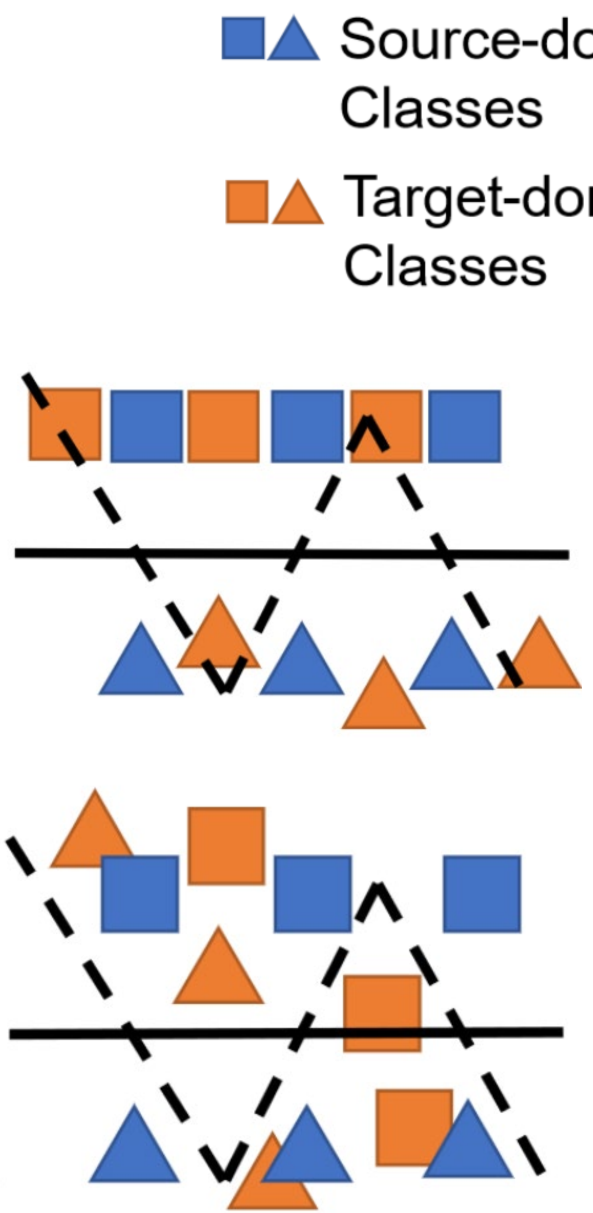
Paper



1. Find a small perturbation $\boldsymbol{\delta}$ to $(\boldsymbol{x}, y)$ that causes the network to predict a wrong label.
2. Train the network to predict y on input $\boldsymbol{x} + \boldsymbol{\delta}$, so the network becomes robust to $\boldsymbol{\delta}$.
3. Result:
    - a smooth decision boundary
    - passing through regions with low data density

OPTIMA: Intuition #2

- We only care about the smoothness of the decision boundary where the target and source domains are similar.
- Thus, we learn a perturbation $\boldsymbol{\delta}$ that conflates $\boldsymbol{x}_{\text{source}} + \boldsymbol{\delta}$ and $\boldsymbol{x}_{\text{target}}$



Smooth > Zigzag

Smooth ≈ Zigzag

OPTIMA: Find Perturbation $\boldsymbol{\delta}$

Paper

$x_{\text{source}} + \delta$ and x_{target} cannot be distinguished by an adversarial discriminator.

$$\boldsymbol{\delta}^* = \underset{\|\boldsymbol{\delta}\| \le \epsilon}{\operatorname{argmax}} \ \log P_{\text{disc}}(y = \text{target} | \boldsymbol{x}_{\text{source}} + \boldsymbol{\delta})$$

$$+ KL\left(f_{\boldsymbol{p}}(y|\boldsymbol{x}_{\text{source}} + \boldsymbol{\delta}) \| f_{\boldsymbol{p}}(y|\boldsymbol{x}_{\text{source}})\right)$$

The perturbation $\boldsymbol{\delta}$ causes maximum change in the model prediction.

OPTIMA: Find Soft Prompt $\boldsymbol{p}$

Paper

- The soft prompt p aims to minimize

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \mathop{\mathbb{E}}_{(\boldsymbol{x}_s, y_s) \in \mathcal{D}_s} \left[\ell_{\text{xe}}(\boldsymbol{x}_s, y_s, \boldsymbol{p}) + \ell_{\text{KL}}(\boldsymbol{\delta}^*, \boldsymbol{p}, \boldsymbol{x}_s) \right]$$

Source-domain cross-entropy loss

Changes in predictions caused by the perturbation $\boldsymbol{\delta}^*$.

$\boldsymbol{x}_s$ and y_s are labeled data from the source domain $\mathcal{D}_s$.

Few-shot Results

Paper

Method	Params	PLM	Source	QQP Acc.	QQP F1	MRPC Acc.	MRPC F1	MNLI 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	59.5 ± 7.8	67.9 ± 12.6	35.6 ± 2.4
PFT	770M		✗	55.1 ± 5.1	57.8 ± 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 Acc.	MRPC → QQP F1	QQP → MRPC Acc.	QQP → MRPC F1	SNLI → MNLI 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		✓	65.0 ± 2.4	64.5 ± 1.5	68.5 ± 2.2	77.6 ± 2.2	75.0 ± 1.0
VAT	102K	T5-Large	✓	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	✗	$\underline{41.6} \pm 3.8$	67.6 ± 6.3		51.2 ± 7.8	
PFT	770M		✗	38.6 ± 5.1	$\underline{71.3} \pm 6.4$		$\underline{57.3} \pm 9.2$	
PPT	410K	T5-XXL	✓	34.7 ± 2.8	54.6 ± 14.0		43.0 ± 14.6	
				MNLI $\rightarrow$ SNLI Acc.	SNLI $\rightarrow$ SICK Acc.	MNLI $\rightarrow$ SICK Acc.	SNLI $\rightarrow$ CB Acc.	MNLI $\rightarrow$ 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		✓	81.5 ± 0.7	69.5 ± 6.8	73.1 ± 4.8	61.6 ± 4.2	66.1 ± 3.3
VAT	102K	T5-Large	✓	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		✓	$\mathbf{82.1}^{*} \pm 0.8$	$\mathbf{73.3} \pm 6.8$	$\mathbf{74.8} \pm 4.4$	$\mathbf{64.8}^{*} \pm 1.1$	$\mathbf{71.2}^{*} \pm 3.1$

Source-domain & Zero-shot Results

Paper

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

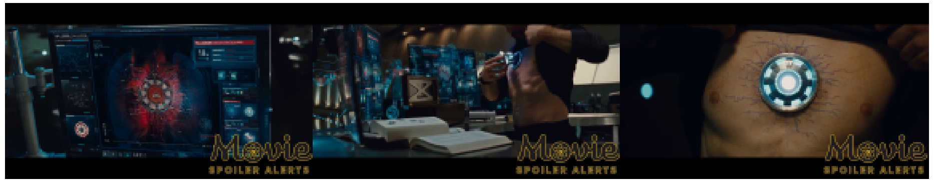
Problems Yet Unsolved?

A new dataset on movie summary understanding.

New Dataset: Synopses of Movie Narratives

Paper

0'44.86 → 0'50.02



The arc reactor however is slowly poisoning him which is causing him to begin to fear death.

0'50.03 → 0'54.02

Stark makes Pepper Potts the CEO of Stark Industries and hires Natalie rushman as his new personal assistant.

- "Watch a movie in 5 minutes" videos
- 869 hours, 683,611 sentences

- Events at the right granularity
- Mental state descriptions
- Semantic gaps between modalities due to storytelling techniques.

Yidan Sun, Qin Chao, Yangfeng Ji, Boyang Li. Synopses of Movie Narratives: a Video-Language Dataset for Story Understanding.

Storytelling Techniques: Symbolism

Paper

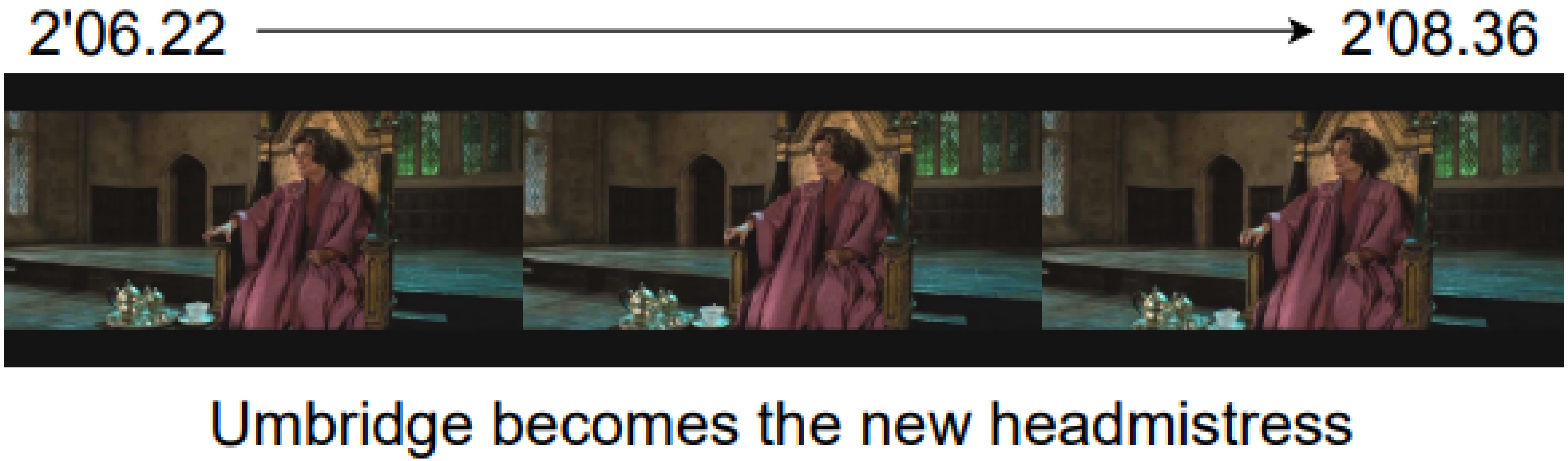


Fig. 4 An example from *Harry Potter and the Order of the Phoenix*. A symbolic object, the chair, is used to represent the event Dolores Umbridge becoming headmistress.

Yidan Sun, Qin Chao, Yangfeng Ji, Boyang Li. Synopses of Movie Narratives: a Video-Language Dataset for Story Understanding.

Storytelling Techniques: Omission of An Obvious Cause or Effect

Paper

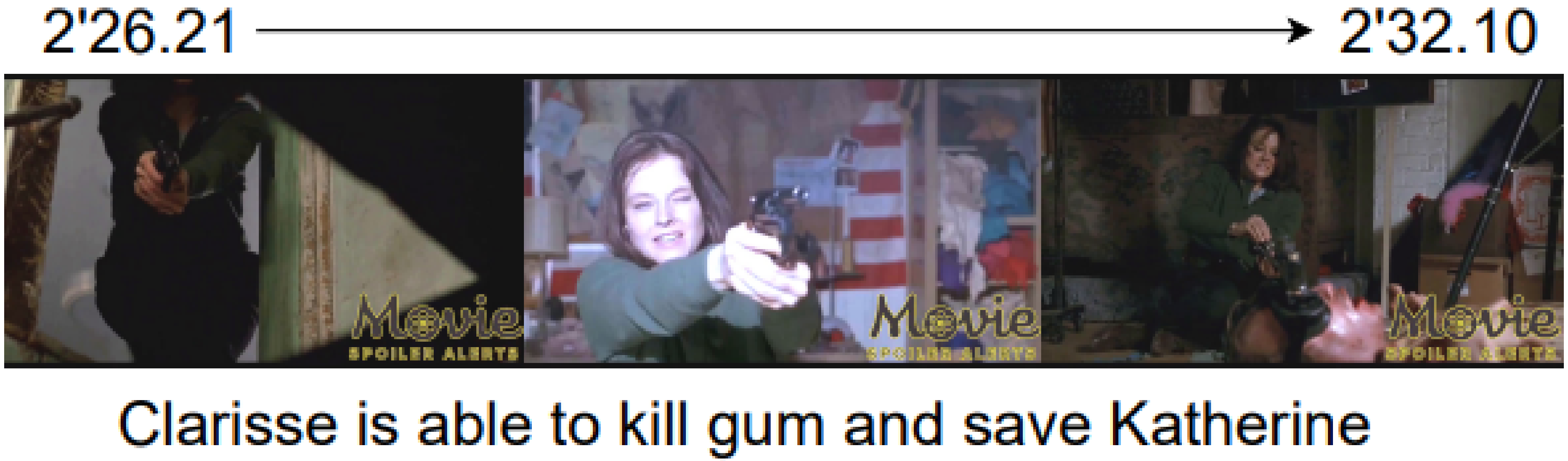


Fig. 3 This example shows three frames from *Silence of the Lambs*. The text (kill) describes the effect of the video (shooting).

Yidan Sun, Qin Chao, Yangfeng Ji, Boyang Li. Synopses of Movie Narratives: a Video-Language Dataset for Story Understanding.

Storytelling Techniques: Long-range Dependency

Paper

1'34.23 → 1'39.43

bilbo, having avoided capture, arranges an escape using empty wine barrels that are sent downstream.

2'14.39 → 2'18.07

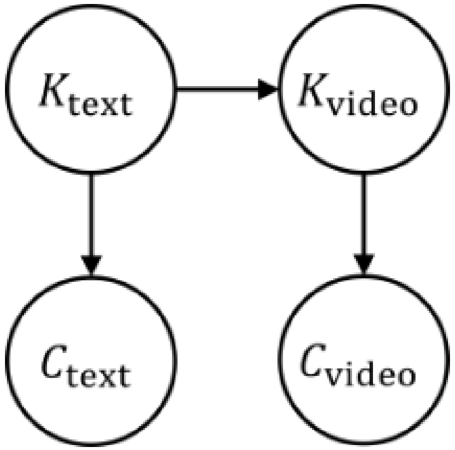
the company is smuggled into asgaroth by a bargeman called bard.

The Cross-modality Semantic Gap: Quantitative Estimates

Paper

Dataset	Estimated Semantic Gap
SyMoN	31.4%
CMD	69.9%
LSMDC	22.9%

Principled Bayesian analysis



Yidan Sun, Qin Chao, Yangfeng Ji, Boyang Li. Synopses of Movie Narratives: a Video-Language Dataset for Story Understanding.

Video-Text Retrieval / Sequence Alignment

- Requires understanding of storytelling techniques.
- Relatively objective measurements

	Clip Acc.	Sent. IoU
Original Split (sub-sentence level)		
UniVL	3.3	1.0
VideoCLIP	4.8	0.6
NeuMATCH-MD (Supervised)	4.0	2.4
UniVL-SyMoN	5.9 ± 0.3	$\mathbf{2.7 \pm 0.2}$
UniVL-SyMoN-memory	$\mathbf{6.5 \pm 0.3}$	2.6 ± 0.2
New Split (sub-sentence level)		
UniVL	7.4	1.0
VideoCLIP	7.6	0.7
UniVL-SyMoN	10.1 ± 0.4	1.9 ± 0.1
UniVL-SyMoN-memory	$\mathbf{13.5 \pm 0.3}$	$\mathbf{2.6 \pm 0.1}$
Original Split (sentence level)		
UniVL	4.6	0.8
VideoCLIP	4.0	1.1
UniVL-SyMoN	7.4 ± 0.1	$\mathbf{3.4 \pm 0.2}$
UniVL-SyMoN-memory	$\mathbf{7.5 \pm 0.4}$	2.1 ± 0.2
New Split (sentence level)		
UniVL	5.7	1.3
VideoCLIP	4.9	1.0
UniVL-SyMoN	7.7 ± 0.2	$\mathbf{3.3 \pm 0.2}$
UniVL-SyMoN-memory	$\mathbf{8.7 \pm 0.3}$	3.2 ± 0.2

Paper

Yidan Sun, Qin Chao, Yangfeng Ji, Boyang Li. Synopses of Movie Narratives: a Video-Language Dataset for Story Understanding.

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

Collaborators

Email: boyang.li@ntu.edu.sg