Microsoft

Microsoft Research

Summit 2021

How Much Can GPT-3 Benefit Few-Shot Visual Reasoning?

Zhe Gan
Principal Researcher

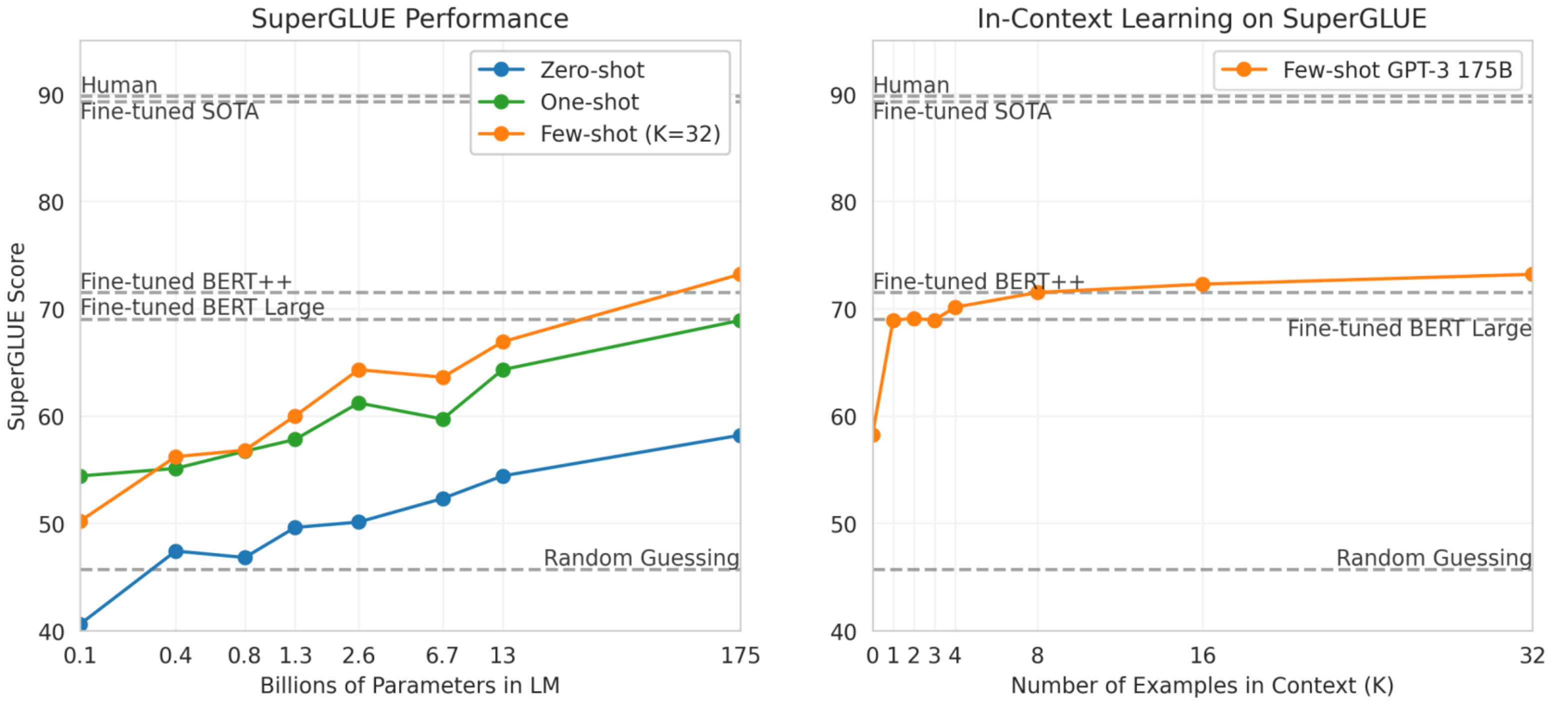
Language Model Pre-training

- Large-scale language model pre-training has become a central training paradigm for NLP
- Parameter-counts are frequently measured in billions (e.g., GPT-3) rather than millions (e.g., BERT)

Model	Company	Param. Count
GPT	OpenAI	110M
BERT-Large	Google	340M
GPT-2	OpenAI	1.5B
MegatronLM	NVIDIA	8.3B
T-NLG	Microsoft	17B
GPT-3	OpenAI	175B
Switch-C	Google	1.6T

Language Models are Few-Shot Learners

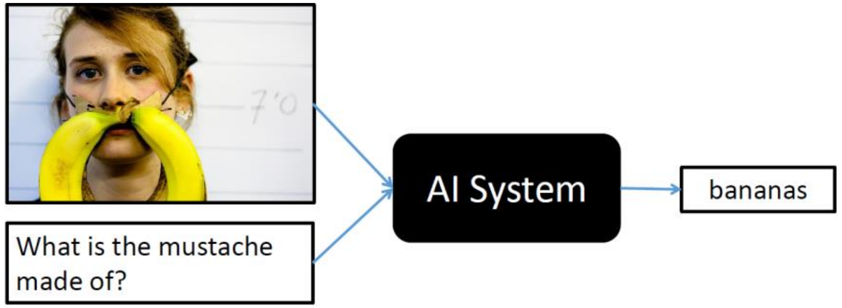
- By providing only a few in-context examples, GPT-3 with 175B parameters has demonstrated strong few-shot performance



[1] Brown, Tom, et al. "Language Models are Few-Shot Learners." In NeurIPS 2020.

Can GPT-3 also Benefit Visual Reasoning Tasks?

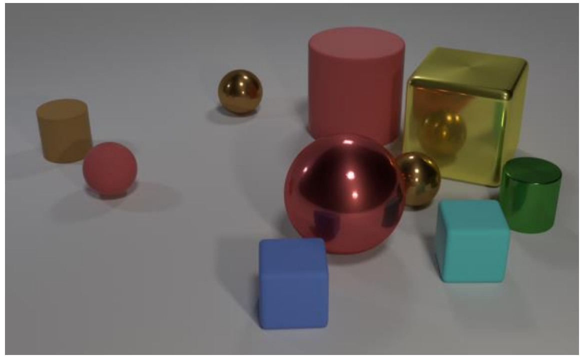
GQA



VQA

VCR

Referring Expressions

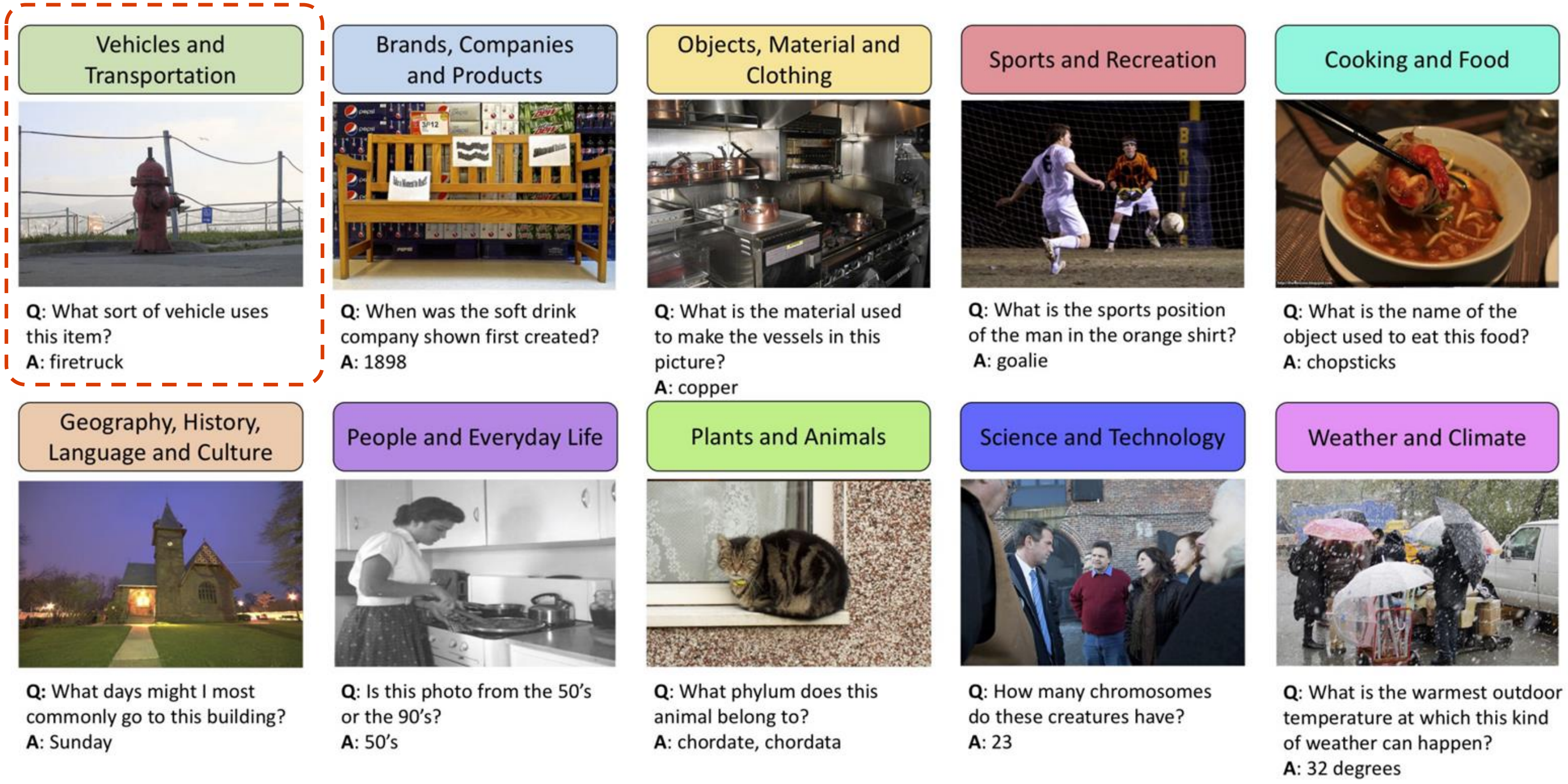


CLEVR

NLVR2

Knowledge-Based VQA

- *OK-VQA*: A VQA benchmark requiring external knowledge not present in the image to correctly answer the question

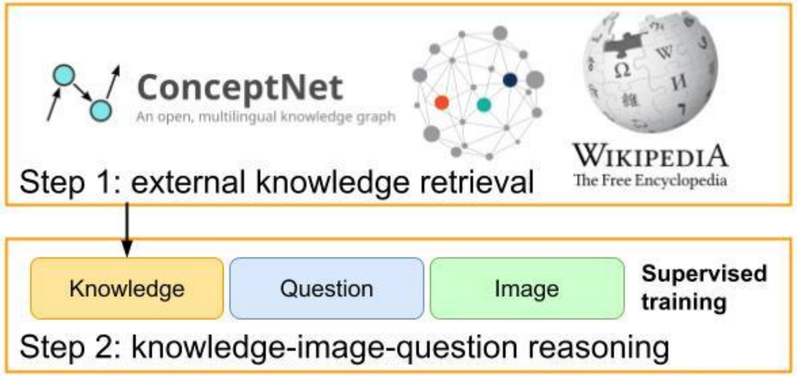


[1] Marino, Kenneth, et al. "Ok-vqa: A visual question answering benchmark requiring external knowledge." In CVPR 2019.

Previous Methods vs. Ours

- Previous methods:
  - Separate two steps: knowledge retrieval and reasoning
  - Using *explicit* and *structured* KBs
  - The retrieved knowledge might be noisy and irrelevant to the question
  - The re-embedded knowledge features during reasoning might deviate from their original meanings in the knowledge source

(a) Previous: separate knowledge retrieval and reasoning

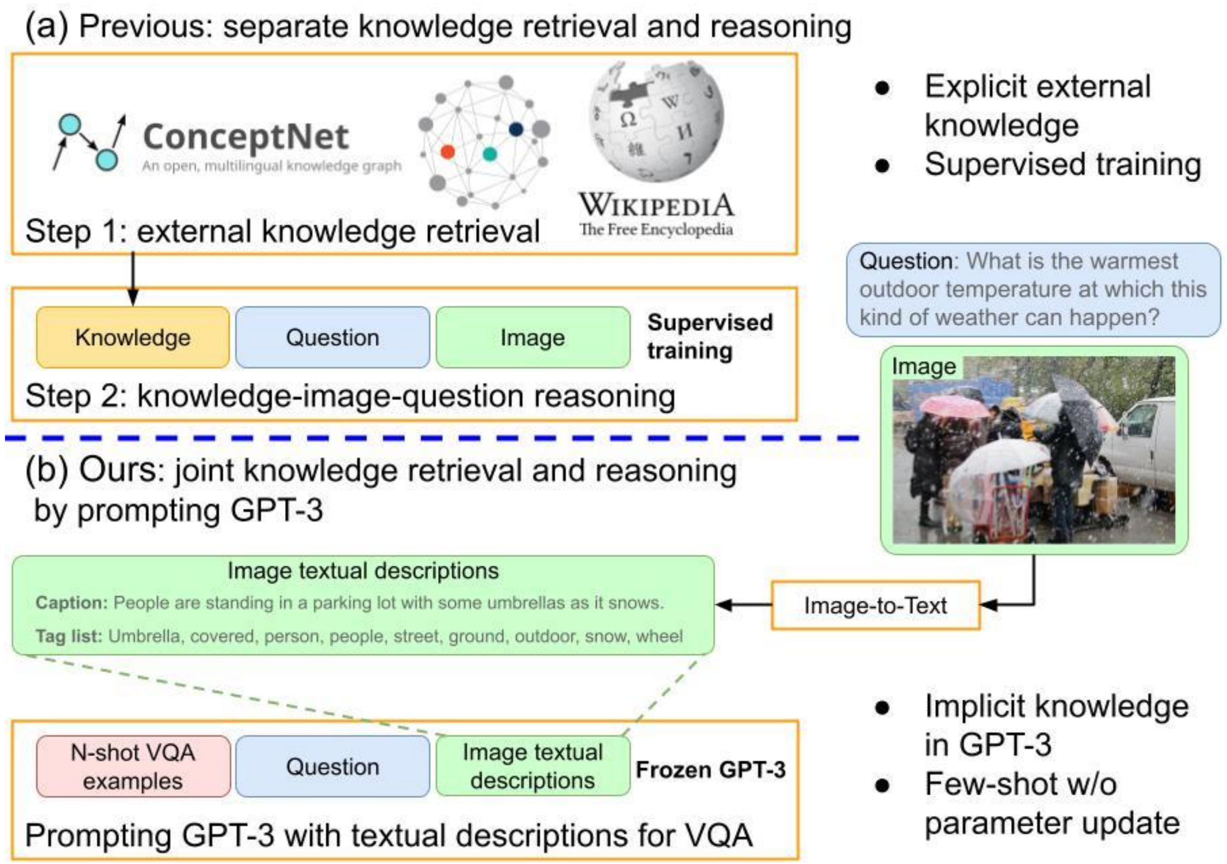


- Explicit external knowledge
- Supervised training

Question: What is the warmest outdoor temperature at which this kind of weather can happen?

Previous Methods vs. Ours

- Previous methods:
  - Separate two steps: knowledge retrieval and reasoning
- Our method:
  - **PICa**: **P**rompting GPT-3 via the use of **I**mage **Ca**ptions
  - Treating GPT-3 as an *implicit* and *unstructured* KB
  - 4 shots outperform supervised SOTA

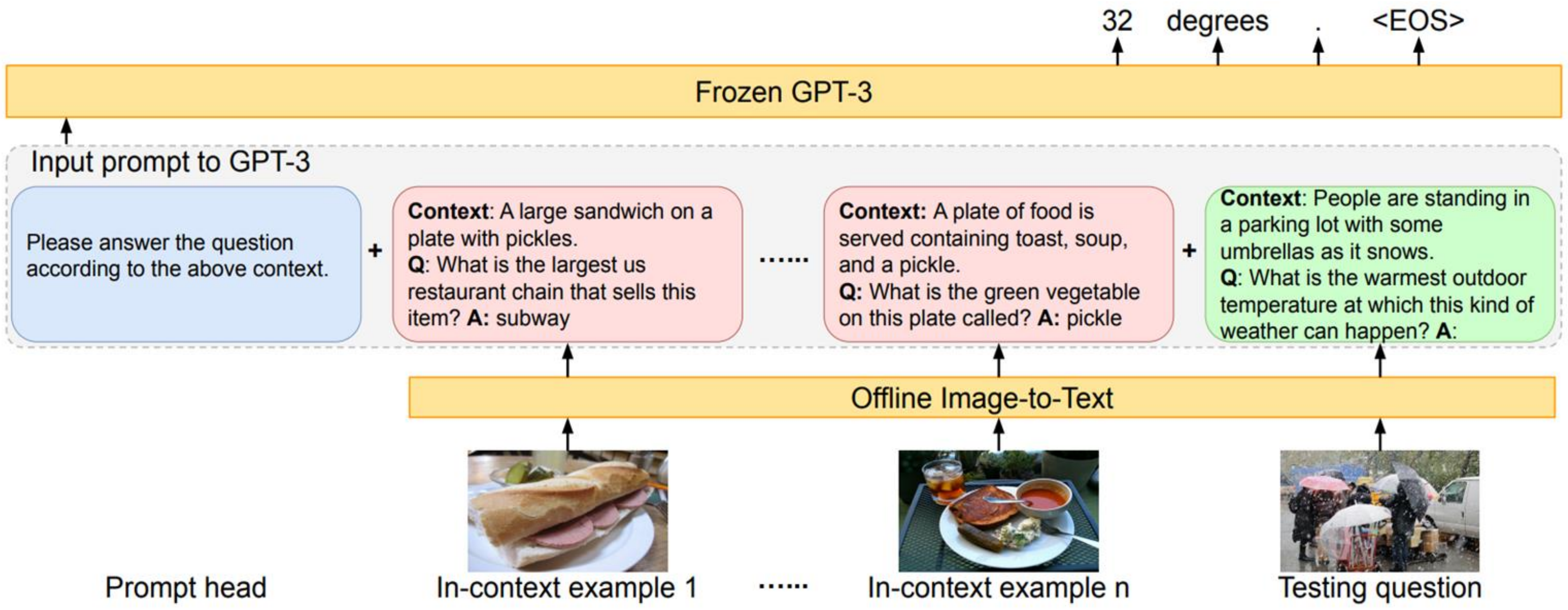


- Explicit external knowledge
- Supervised training

- Implicit knowledge in GPT-3
- Few-shot w/o parameter update

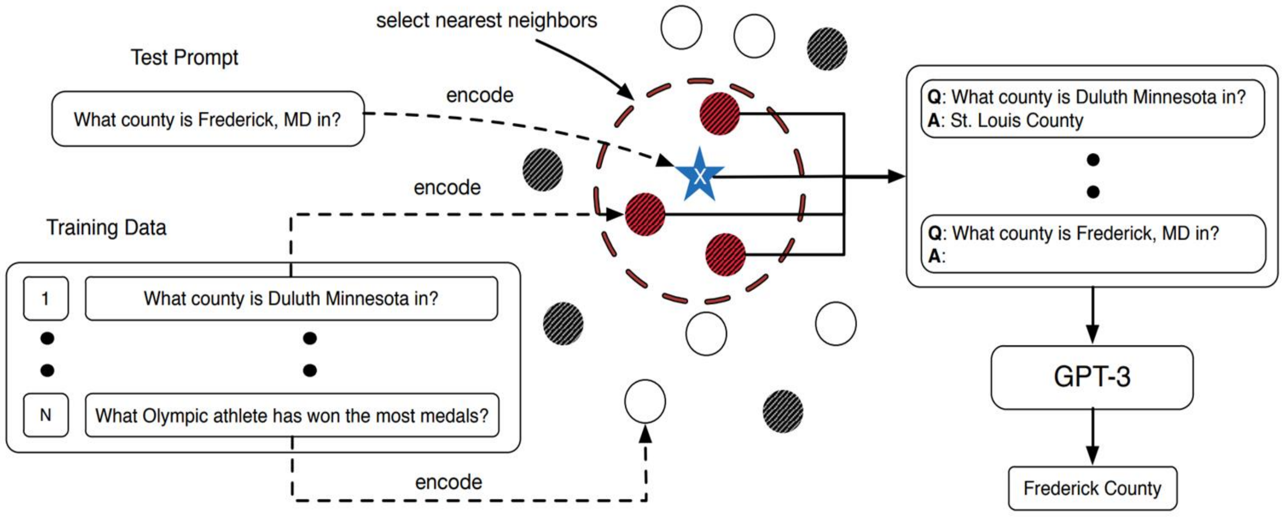
How Do We Prompt GPT-3?

- Convert images into textual descriptions (captions, tags)
- Produce the answer in an open-ended text generation manner



How to Enhance the Performance?

- In-context example selection
  - “Better” in-context examples based on question and image similarity



[1] Liu, Jiacheng, et al. "What Makes Good In-Context Examples for GPT-3?“, 2021.

When was this type of transportation invented?

Testing question

How is this dish cooked?

What is the name of a popular skateboarding trick?

What is the scientific name of those animals?

Random

When did this type of transportation originate?

When was this mode of transportation invented?

How common is this form of transportation?

Question

Who invented this mode of transportation?

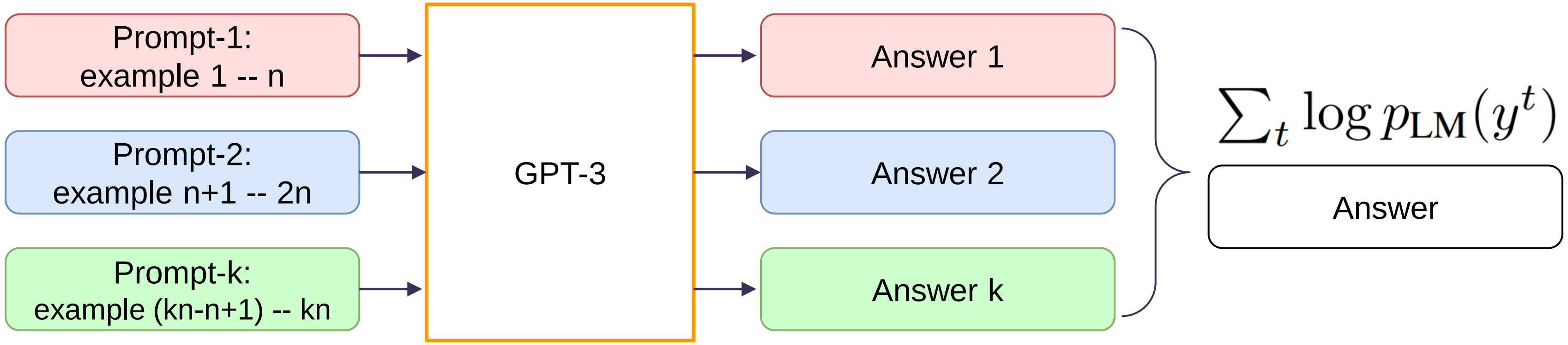
The driver of this type of vehicle is called a what?

Is this train regulated or unregulated?

Question+Image

How to Enhance the Performance?

- In-context example selection
  - "Better" in-context examples based on question and image similarity
- Multi-query ensemble
  - Merge predictions from multiple queries with different examples



PICa Outperforms Supervised SOTA by +8.6 Points

- *PICa-Base*: w/o in-context selection and multi-query ensemble
- *PICa-Base* (43.3) already surpasses SOTA (39.4)
- *PICa-Full* further boosts performance (48.0)
- Both captions and tags are useful for GPT-3 prompting

Method	Image Repr.	Knowledge Resources	Few-shot	Accuracy
MUTAN+AN (Ben-Younes et al. 2017)	Feature Emb.	Wikipedia	✗	27.8
Mucko (Zhu et al. 2020)	Feature Emb.	Dense Captions	✗	29.2
ConceptBert (Garderes et al. 2020)	Feature Emb.	ConceptNet	✗	33.7
ViLBERT (Lu et al. 2019)	Feature Emb.	None	✗	35.2
KRISP (Marino et al. 2021)	Feature Emb.	Wikipedia + ConceptNet	✗	38.9
MAVEx (Wu et al. 2021)	Feature Emb.	Wikipedia + ConceptNet + Google Images	✗	39.4
Frozen (Tsimpoukelli et al. 2021)	Feature Emb.	Language Model (7B)	✓	12.6
PICa-Base	Caption	GPT-3 (175B)	✓	42.0
PICa-Base	Caption+Tags	GPT-3 (175B)	✓	43.3
PICa-Full	Caption	GPT-3 (175B)	✓	46.9
PICa-Full	Caption+Tags	GPT-3 (175B)	✓	**48.0**

How Many Shots are Enough?

- 4 shots outperform supervised state-of-the-art (39.4)
- More shots lead to better performance
- PICa outperforms Frozen by a significant margin

	Method	Image Repr.	n=0	n=1	n=4	n=8	n=16	Example engineering
(a)	Frozen (Tsimpoukelli et al. 2021)	Feature Emb.	5.9	9.7	12.6	-	-	✗
(b)	`PICa-Base`	Caption	17.5	32.4	37.6	39.6	42.0	✗
(c)	`PICa-Base`	Caption+Tags	16.4	34.0	39.7	41.8	43.3	✗
(d)	`PICa-Full`	Caption	17.7	40.3	44.8	46.1	46.9	✓
(e)	`PICa-Full`	Caption+Tags	17.1	40.8	45.4	46.8	**48.0**	✓

Why GPT-3 are So Powerful on OK-VQA?

- Encyclopedia and commonsense knowledge

(a) When was this type of transportation invented?

Context: A blue and yellow train traveling down train tracks.
Answer: 1804
GT Answer: ['1804', '1804', '1804', '1804', '1800s']
Acc.: 1.0

(b) When was this type of two wheeled vehicle invented?
Context: A row of motorcycles parked next to each other.
Answer: 1885
GT Answer: ['1885', '1885', '30's', '1845', '1915']

Acc.: 1.0

(c) Where can you get these?

Context: A shopping cart filled with bananas and other items.
Answer: grocery store
GT Answer: ['grocery', 'grocery', 'supermarket', 'store', 'grocery store']
Acc.: 0.6

(d) Where does this vehicle stop?

Context: A blue and white train traveling down train tracks.
Answer: train station
GT Answer: ['train station', 'train station', 'train station', 'station', 'station']
Acc.: 1.0

(e) What animal do you have to watch out for when doing this sport?
Context: A man holding a surfboard on top of a beach.
Answer: shark
GT Answer: ['shark', 'shark', 'shark', 'wave', 'shark']
Acc.: 1.0

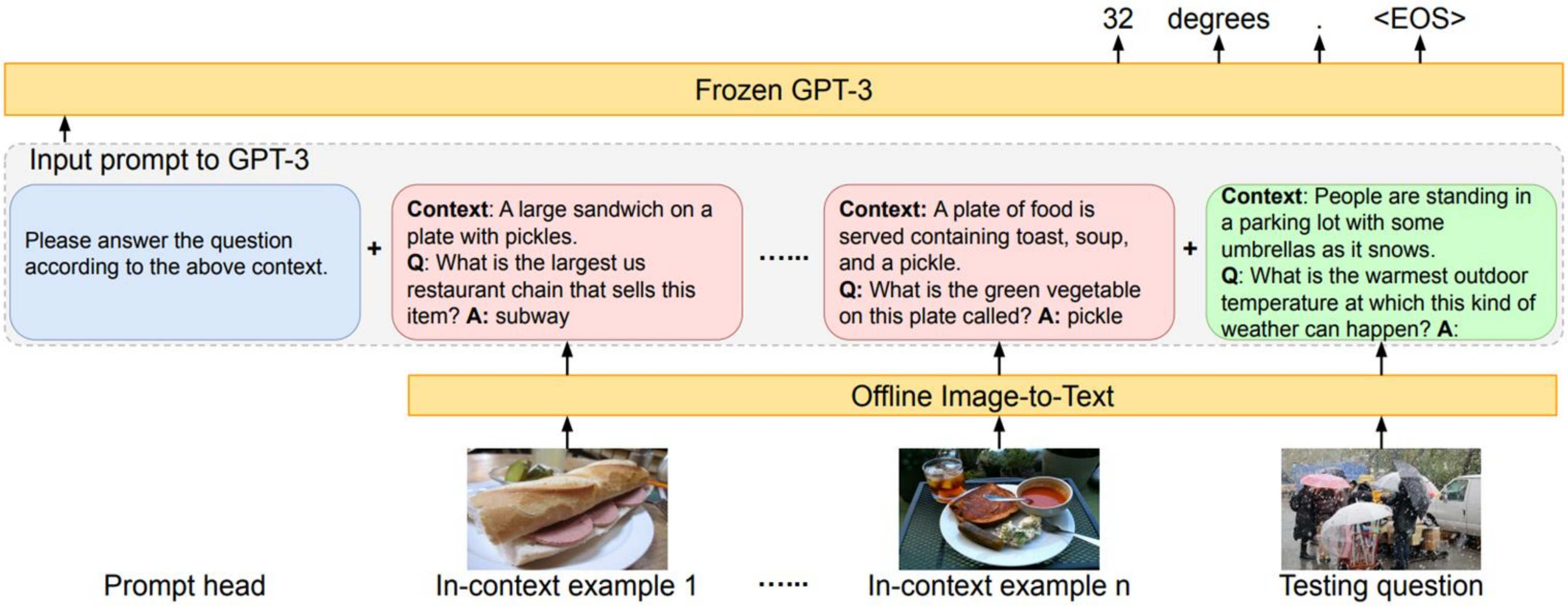
Why GPT-3 Are So Powerful on OK-VQA?

- GPT-3 also generates answer rationales reasonably well

(a) When was this type of transportation invented?	**(b)** When was this type of two wheeled vehicle invented?	**(c)** Where can you get these?	**(d)** Where does this vehicle stop?	**(e)** What animal do you have to watch out for when doing this sport?
Context: A blue and yellow train traveling down train tracks.	**Context**: A row of motorcycles parked next to each other.	**Context**: A shopping cart filled with bananas and other items.	**Context**: A blue and white train traveling down train tracks.	**Context**: A man holding a surfboard on top of a beach.
Answer: 1804	**Answer**: 1885	**Answer**: grocery store	**Answer**: train station	**Answer**: shark
GT Answer: ['1804', '1804', '1804', '1804', '1800s']	**GT Answer**: ['1885', '1885', '30's', '1845', '1915']	**GT Answer**: ['grocery', 'grocery', 'supermarket', 'store', 'grocery store']	**GT Answer**: ['train station', 'train station', 'train station', 'station', 'station']	**GT Answer**: ['shark', 'shark', 'shark', 'wave', 'shark']
Acc.: 1.0	**Acc.**: 1.0	**Acc.**: 0.6	**Acc.**: 1.0	**Acc.**: 1.0
(a) This is because: first steam-powered locomotive was invented in 1804.	**(b) This is because:** first motorcycle was invented in 1885	**(c) This is because:** grocery store is most common place get food	**(d) This is because:** train station is only place where train stops	**(e) This is because:** sharks are dangerous animals

Key Takeaways

- The first study of using GPT-3 for multimodal tasks
- With 16 in-context examples, GPT-3 surpasses the supervised SOTA by an absolute +8.6 points on the challenging OK-VQA dataset



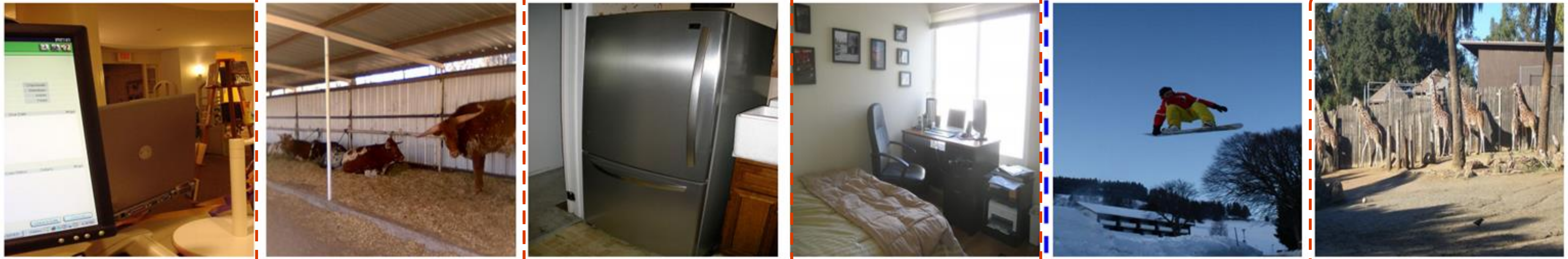
Limitations of Using GPT-3 for Visual Reasoning

- Converting images into captions could lose important visual info.
- We test PICa on VQAv2, with questions on detailed visual contents
- Human performance:
  - 40.8% with question only
  - 57.5% with question and captions
  - 83.3% with question and images

Method	Image Repr.	Few-shot	Acc.
Oscar (Li et al. 2020)	Feature Emb.	✗	73.8
Frozen	Feature Emb.	✓	38.2
PICa-Base	Caption	✓	53.2
PICa-Base	Caption+Tags	✓	54.3
PICa-Full	Caption	✓	55.9
PICa-Full	Caption+Tags	✓	**56.1**
PICa-Full†	GT-Caption-5	✓	59.7

Limitations of Using GPT-3 for Visual Reasoning

- *Success*: knowledge, related descriptions, and reasonable guesses
- *Failures*: missing detailed visual information



(a) What kind of sign is above the doorway?
Context: A computer monitor sitting on top of a desk.

Answer: exit.
GT Answer: ['exit', 'exit', 'exit', 'exit', 'exit', 'exit', 'neon', 'exit', 'exit', 'exit']

Acc.: 1.0

(b) Why is the cow laying down?
Context: A group of cows in a barn with one standing and one laying down.
Answer: tired
GT Answer: ['tired', 'tired', '4', 'resting', 'tired', 'tired', 'resting', 'tired', "it's tired", 'tired']

Acc.: 1.0

(c) What color is the fridge?

Context: A silver refrigerator freezer sitting inside of a kitchen.
Answer: silver
GT Answer: ['gray or silver', 'brushed steel', 'silver', 'steel gray', 'stainless', 'gray', 'silver', 'stainless steel (gray)', 'gray', 'silver']
Acc.: 1.0

(d) What color are the walls?

Context: A bedroom with a desk, chair, computer and pictures on the wall.
Answer: white
GT Answer: ['white', 'off white', 'off white', 'white', 'white', 'beige', 'white', 'white', 'white', 'white']

Acc.: 1.0

(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black
GT Answer: ['red', 'red', 'red', 'orange', 'red', 'red', 'red', 'red', 'red', 'red']

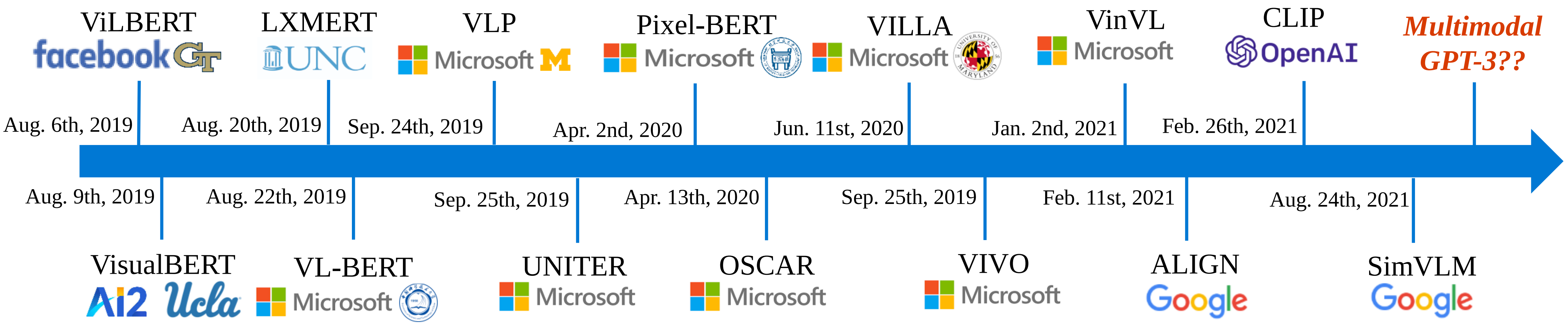
Acc.: 0.0

(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

Acc.: 0.0

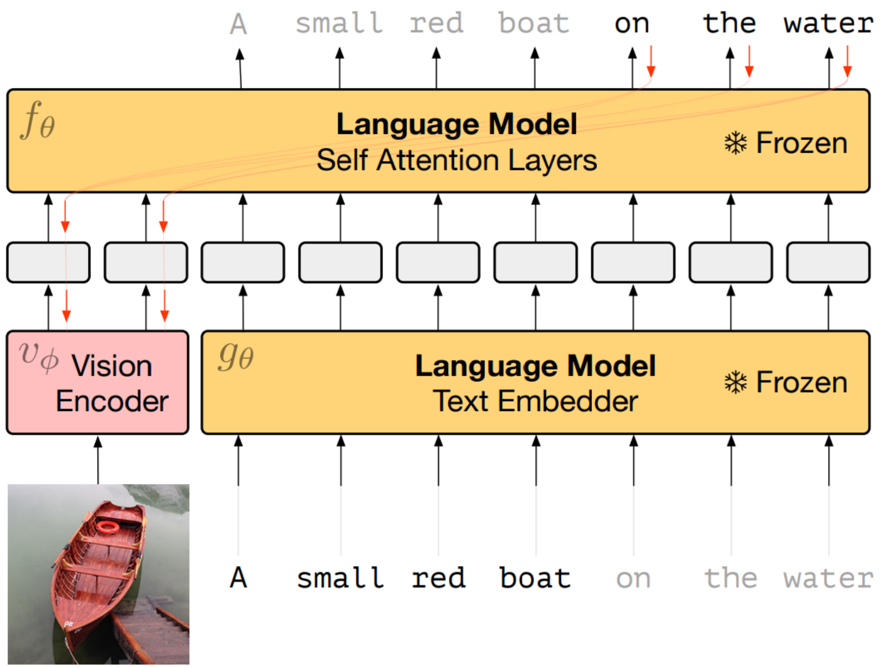
Future Direction

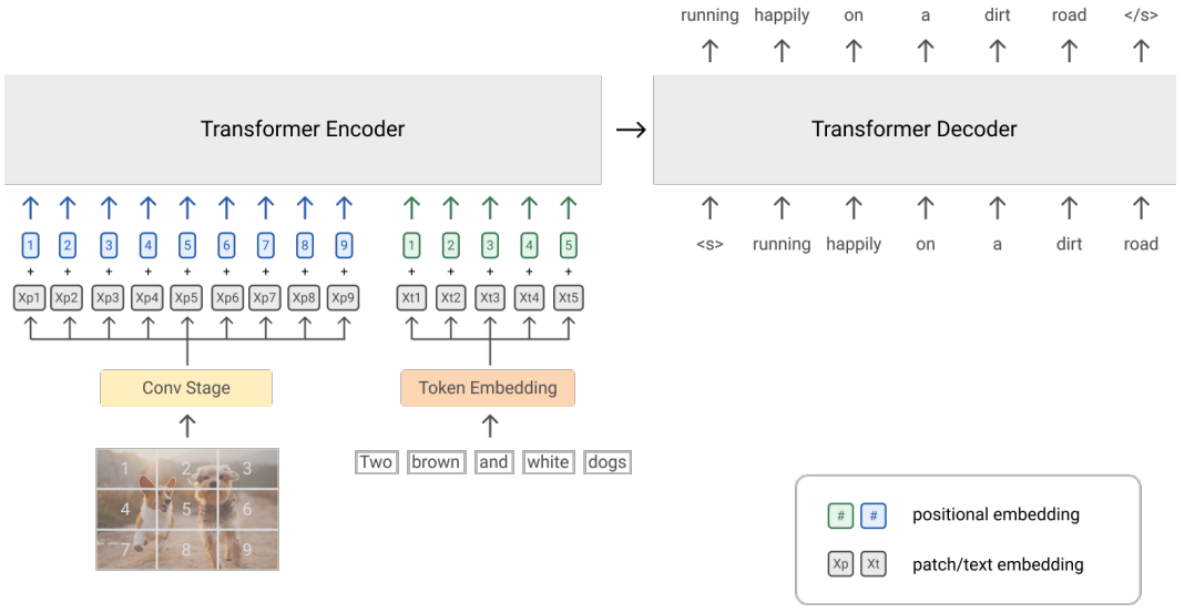
- *Looking back*, Microsoft has been an important player in the vision-language pre-training (VLP) space
- *Looking forward*, when can we have the GPT-3 moment for VLP?



Future Direction

- Multimodal GPT-3:
  - Instead of converting images into captions, *learn a vision encoder* to align with the language embedding space in GPT-3





[1] Tsimpoukelli, Maria, et al. "Multimodal Few-Shot Learning with Frozen Language Models", 2021.
[2] Wang, Zirui, et al. "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision", 2021.

Collaborators

Zhengyuan Yang

Jianfeng Wang

Xiaowei Hu

Yumao Lu

Zicheng Liu

Lijuan Wang

Thank you!