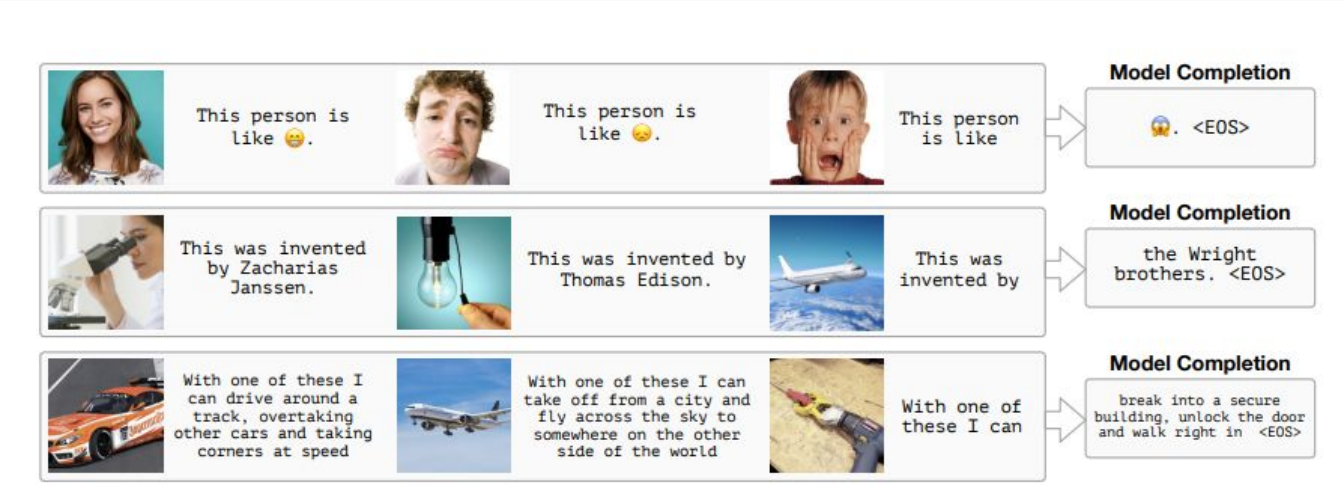
Multimodal Few-Shot Learning with Frozen Language Models

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, Felix Hill (DeepMind)

Presented by Md Mohaiminul Islam, Yan-Bo Lin, Akshay Paruchuri

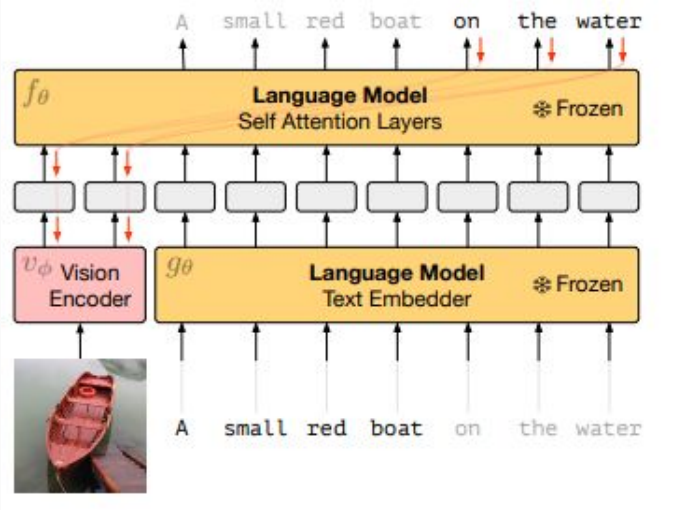
Introduction

- Auto-regressive language models have the capacity to learn a new language task after being given only a few examples
  - Authors target transfer of few-shot learning capability to a multi-modal setting, particularly vision and language
  - A vision encoder is trained to represent each image as a sequence of continuous embeddings, and then a pre-trained, frozen language model with the visual prefix (aforementioned sequence) provides a caption



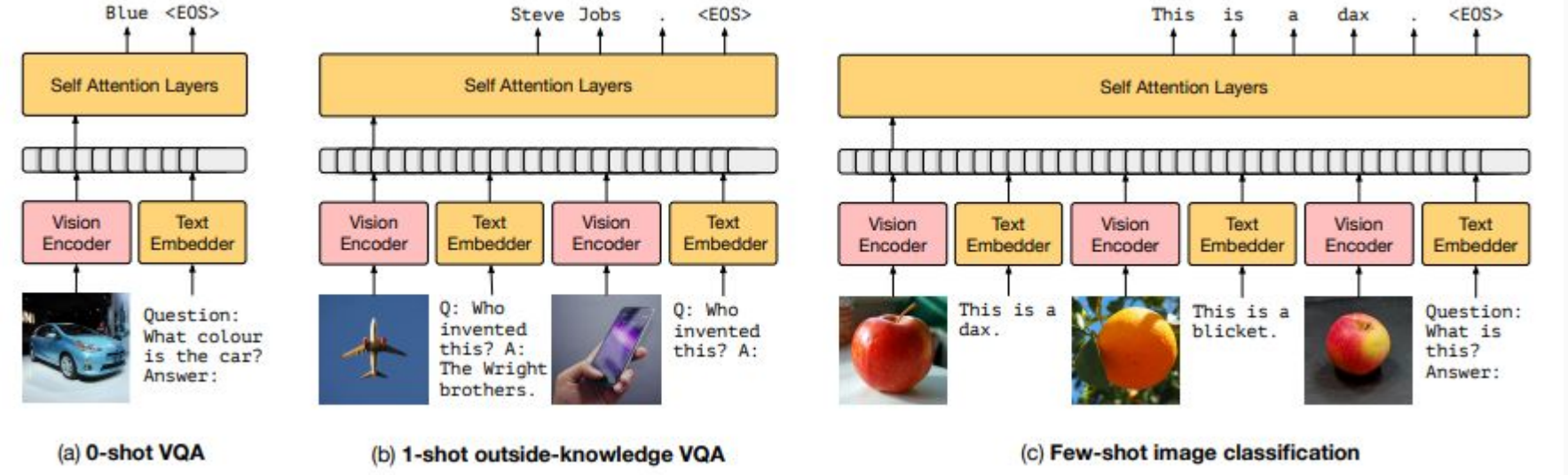
Introduction

- *Frozen*, a method for giving a pre-trained language model access to visual information in a way that extends its few-shot learning capabilities to a multimodal setting, and without updating the weights of the language model
- *Frozen* is capable of rapid task adaptation, encyclopedic knowledge, and fast concept binding.



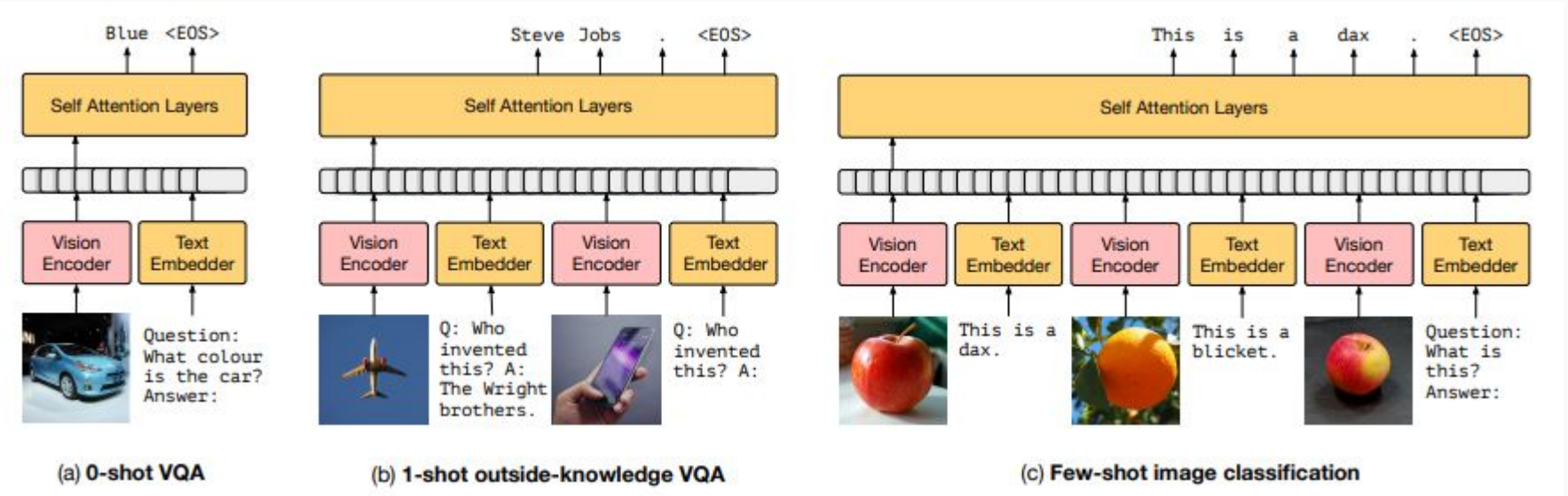
Introduction

- Rapid task adaptation - rapid adaptation to new tasks, typically based on a few examples
- Encyclopedic knowledge - fast access to general knowledge that may be captured in text
- Fast concept binding - fast binding of visual and linguistic elements



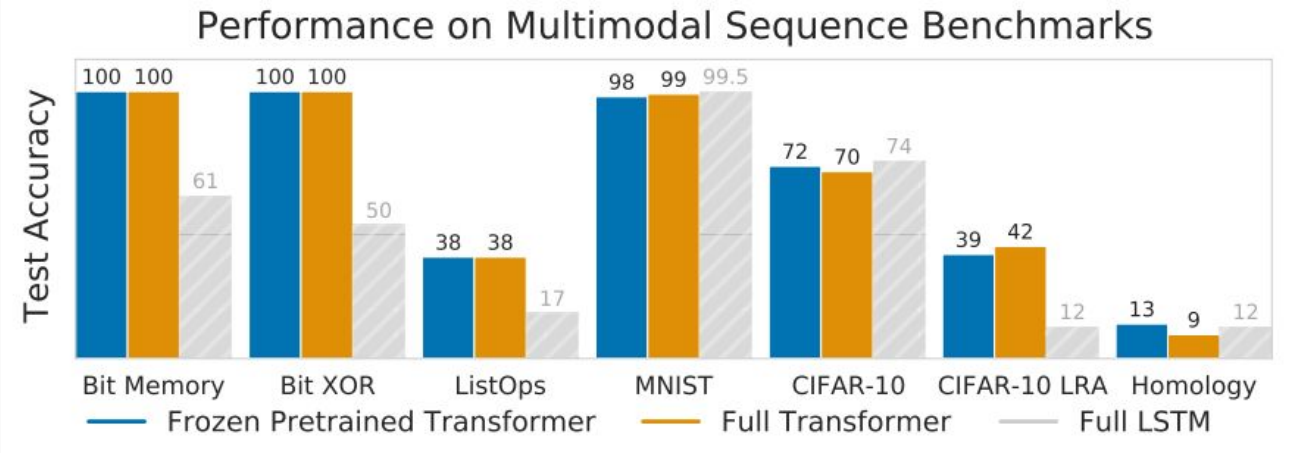
Introduction

- Contributions
  - *Frozen*
  - Models such as *Frozen* can transfer their capacity for rapid task adaptation, encyclopedic knowledge, and fast concept binding from a language-only to a multi-modal setting
  - Quantify aforementioned capabilities on a range of existing and new benchmarks
- Inference-Time interface for *Frozen*



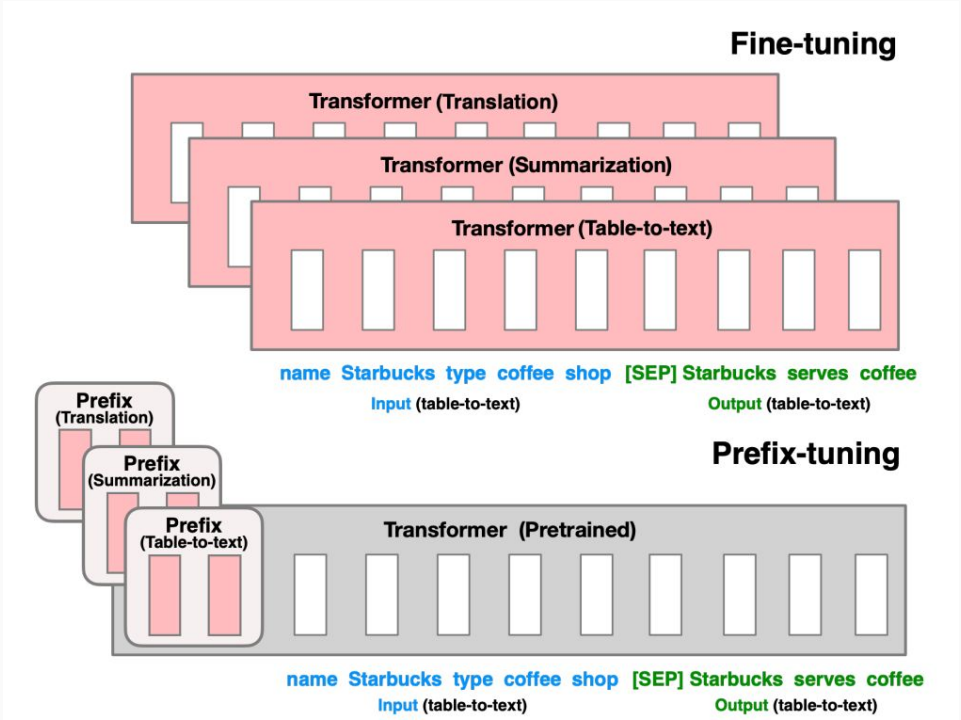
Related Work

- Pretrained Transformers As Universal Computation Engines, Kevin Lu et al. 2021 - Knowledge encoded in transformer language models can be a valuable prior for tasks involving reasoning and memory across discrete sequences
  - Promising results on diverse classification tasks, numerical computation, image classification, and protein fold prediction



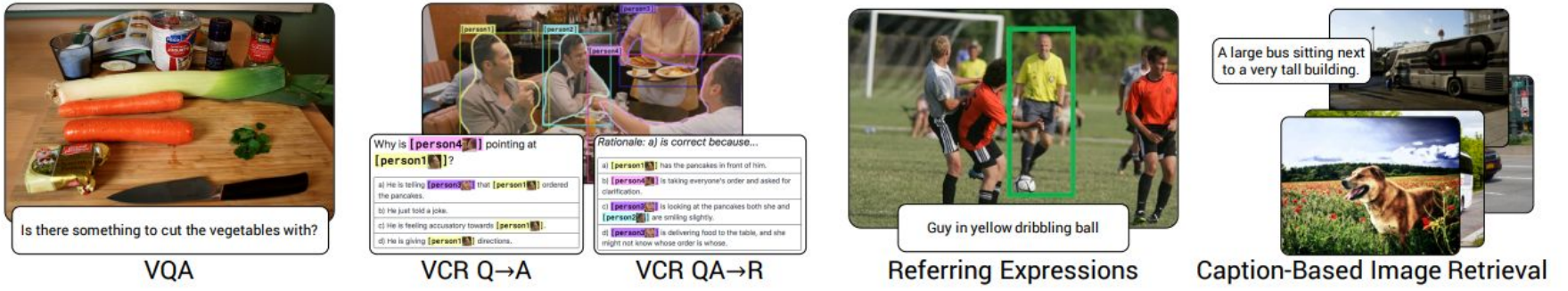
Related Work

- Effectiveness of prefix tuning was an important motivation, for example as shown in Prefix-Tuning: Optimizing Continuous Prompts for Generation, Li et al. 2021.



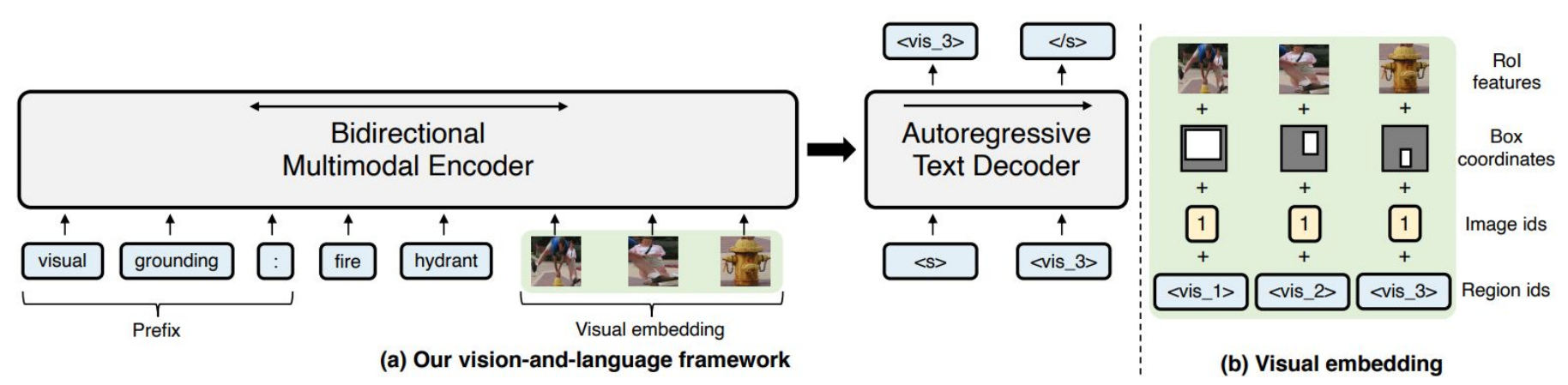
Related Work

- ViLBERT paper by Lu et al. 2019 - Models are first trained with aligned data on task-agnostic cross-modal objectives and then fine-tuned to specific tasks → can yield SOTA performance, but specialized to specific tasks

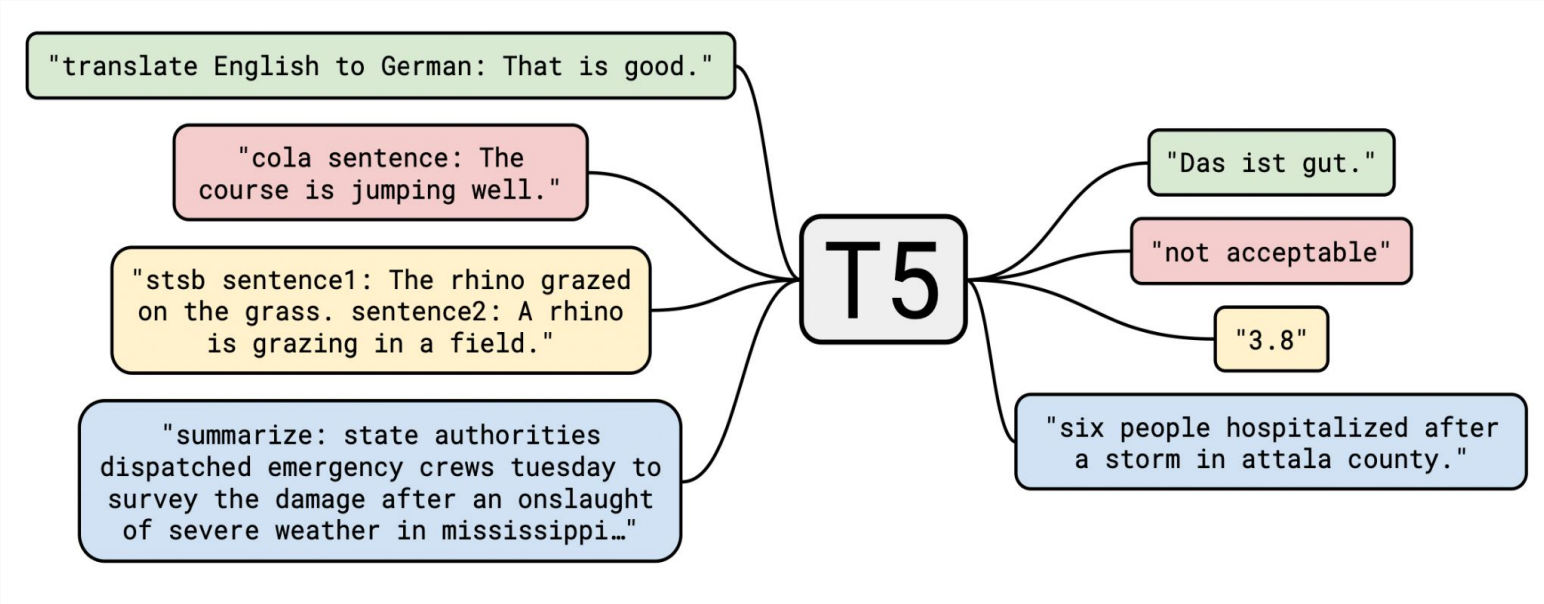


Related Work

- Pre-trained, frozen visual models or models where all weights are updated with training data for a specific task also were referenced, for example Unifying Vision-and-Language Tasks via Text Generation, Cho et al. 2021

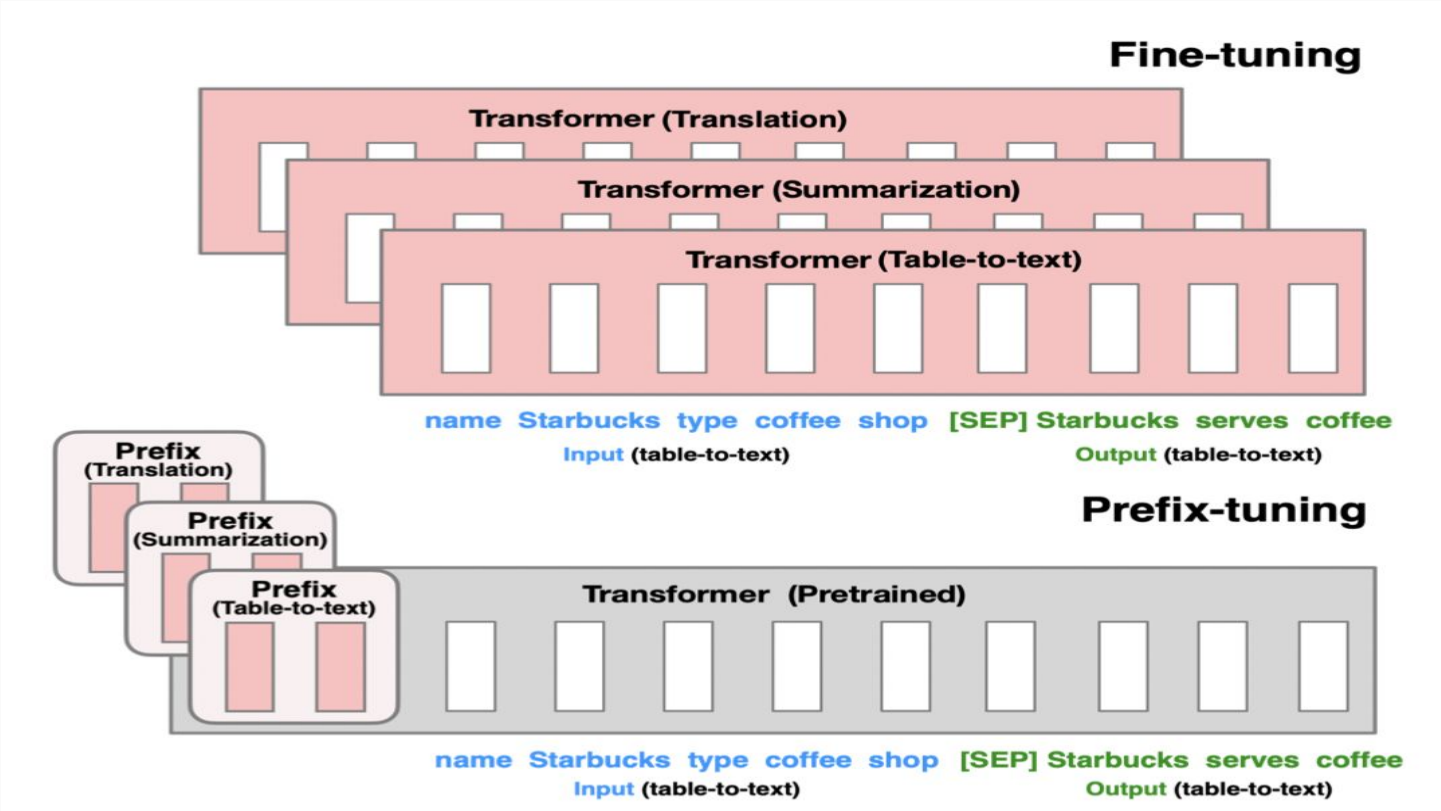


Background: Prompt/ Prefix



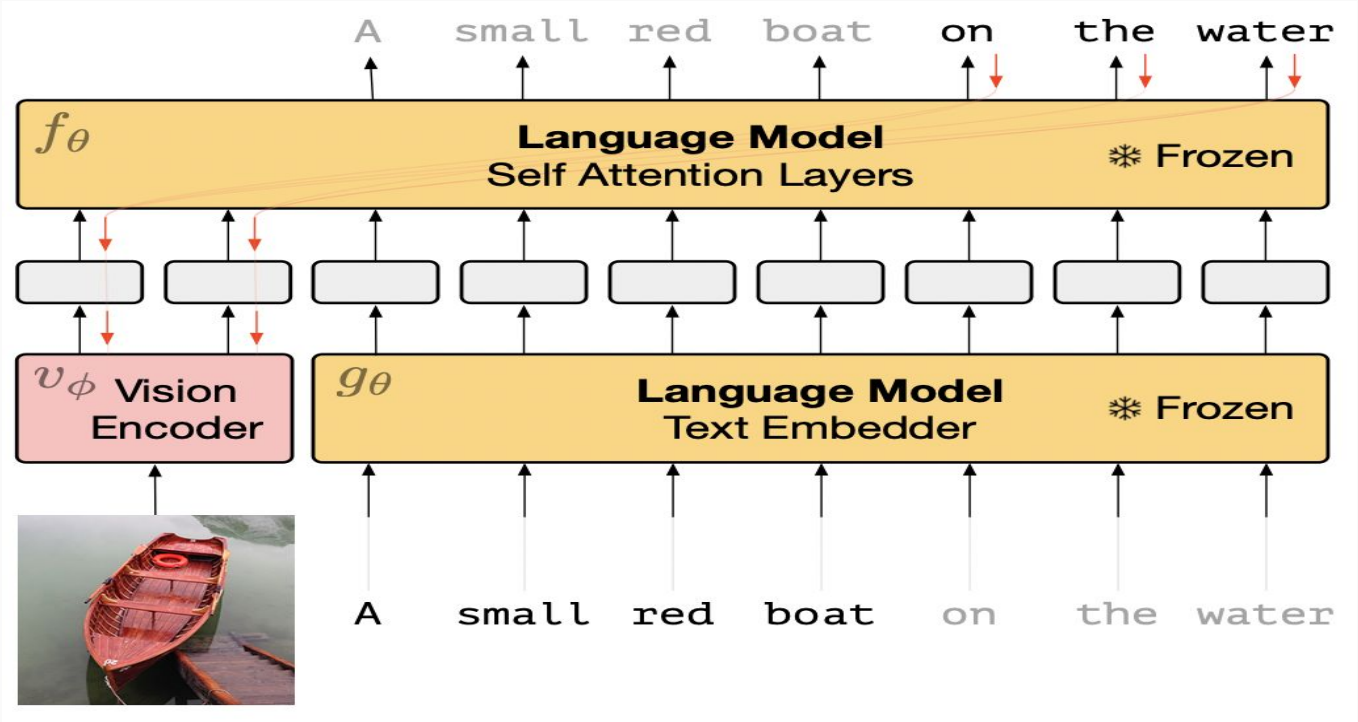
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Background: Prefix Tuning



Prefix-Tuning: Optimizing Continuous Prompts for Generation

The Frozen Method



Architecture

Pre-trained Autoregressive Language Models: GPT-3 pretrained on C4 dataset

$$\log p_\theta(\mathbf{y}) = \sum_l \log p_\theta(y_l | y_1, y_2, ..., y_{l-1}) = \sum_l f_\theta(t_1, t_2, ..., t_{l-1})_{y_l}$$

Vision Encoder: NF-ResNet-50. Final output vector after global average pooling

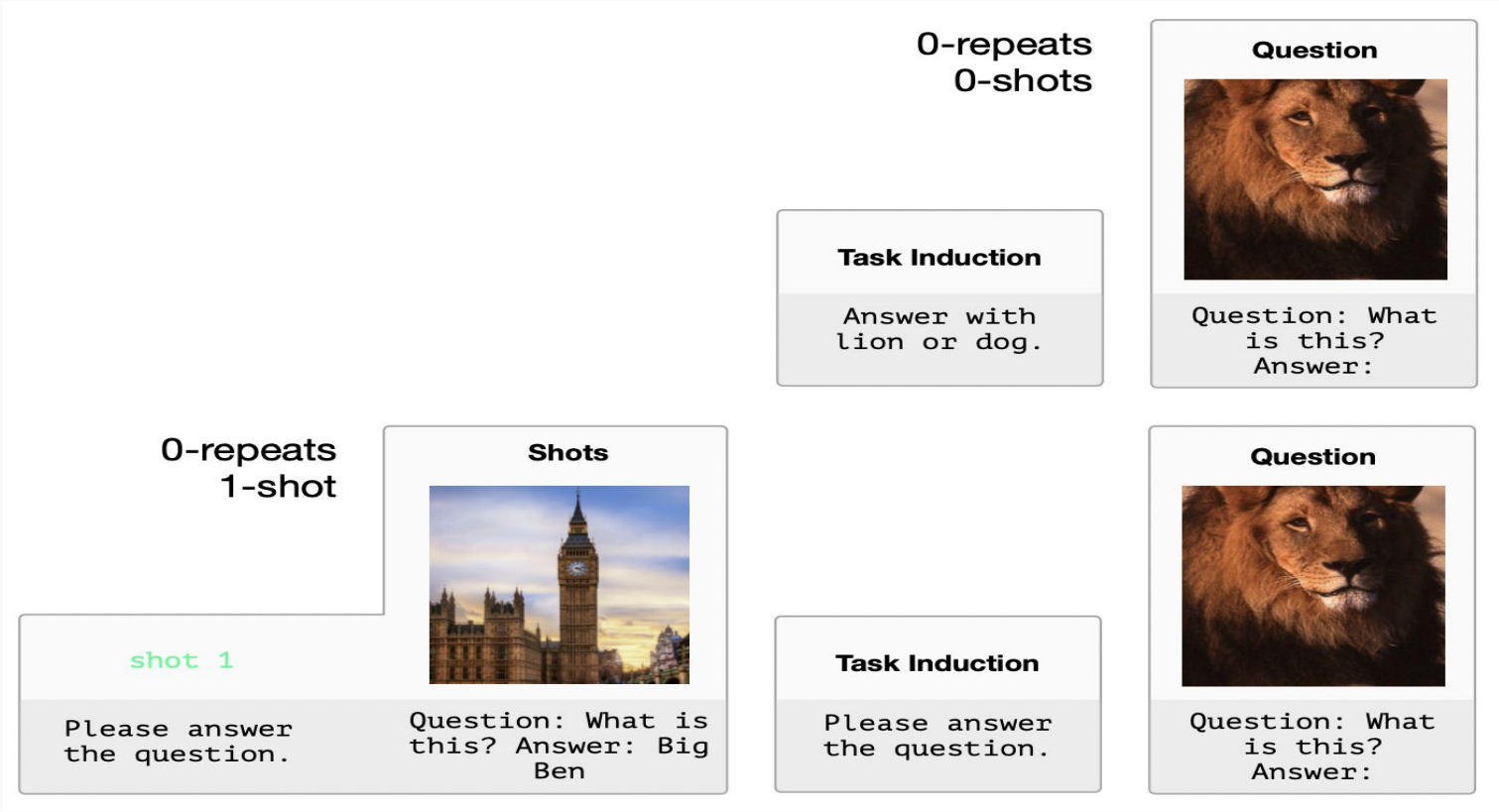
Visual Prefix: Linearly map vision encoder output to n*D channels

Training

only the parameters φ of the vision encoder using paired image-caption data from the Conceptual Captions dataset

$$\log p_{\theta,\phi}(\mathbf{y}|x) = \sum_{l} \log p_{\theta,\phi}(y_l|\mathbf{x}, y_1, y_2, \ldots, y_{l-1})$$
$$= \sum_{l} f_\theta(i_1, i_2, \ldots, i_n, t_1, t_2, \ldots, t_{l-1})_{y_l}$$

Task induction, Shots, Repeats



Task induction, Shots, Repeats

1-repeats
1-shot

repeat 0 **Shots** repeat 1

Please answer the question.

Question: What is this? Answer: Big Ben

Question: What is this? Answer: Big Ben

Task Induction

Please answer the question.

Question

Question: What is this? Answer:

0-repeats
2-shots

Shots

shot 1

Please answer the question.

Question: What is this? Answer: Big Ben

Shots

shot 2

Please answer the question.

Question: Type of animal? Answer: dog

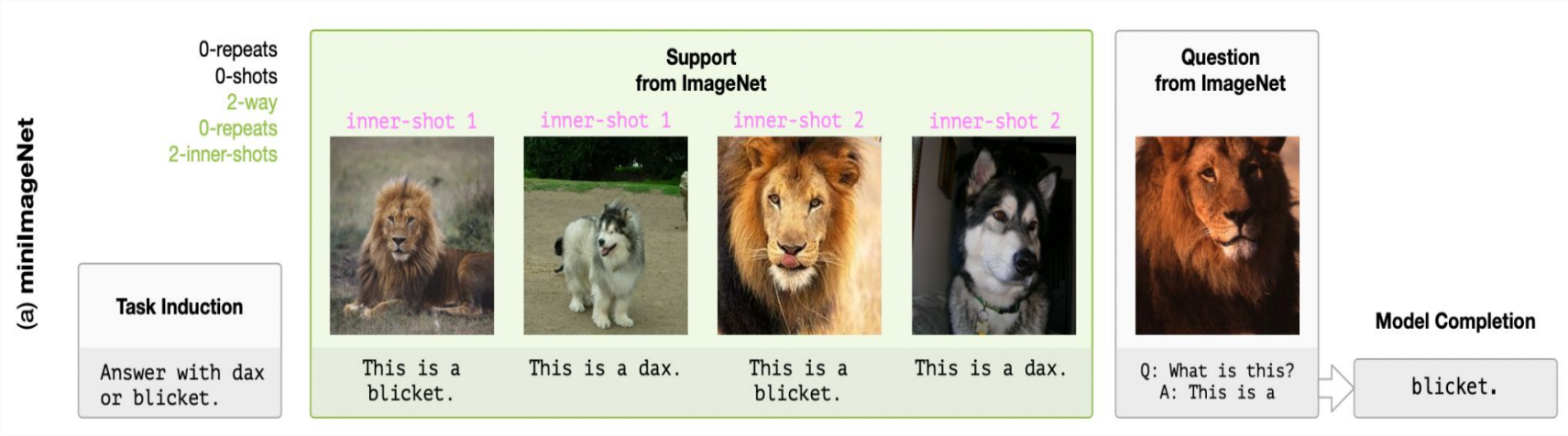
Task Induction

Please answer the question.

Question

Question: What is this? Answer:

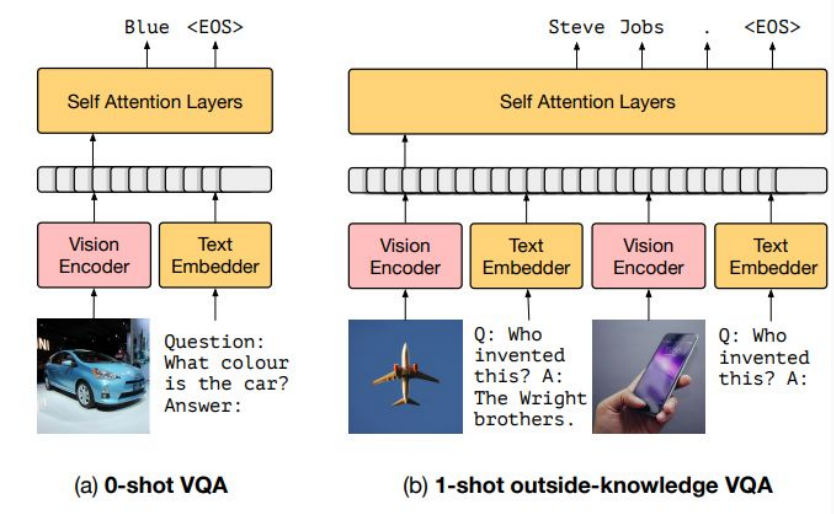
Inner shots

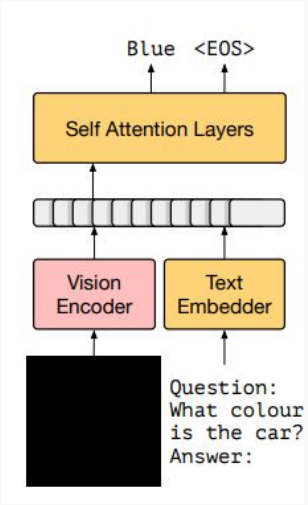

(a) miniImageNet
0-repeats
0-shots
2-way
0-repeats
2-inner-shots
Task Induction
Answer with dax or blicket.
Support from ImageNet
inner-shot 1
inner-shot 1
inner-shot 2
inner-shot 2
This is a blicket.
This is a dax.
This is a blicket.
This is a dax.
Question from ImageNet
Q: What is this? A: This is a
Model Completion
blicket.

Experiment

Zero-shot transfer from captioning to VQA:

- To test rapid adaptation, the model is evaluated on VQA in zero-shot setting.





Experiment

Zero-shot transfer from captioning to VQA:

- Train on Conceptual Captions, transfer to VQAv2.
- Conceptual Captions consist of 3 Million image-caption pairs.
- τ indicates usage of VQAv2.

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

Experiment

Zero-shot transfer from captioning to VQA:

- Train on Conceptual Captions, transfer to VQAv2.
- Conceptual Captions consist of 3 Million image-caption pairs.

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

Experiment

Zero-shot transfer from captioning to VQA:

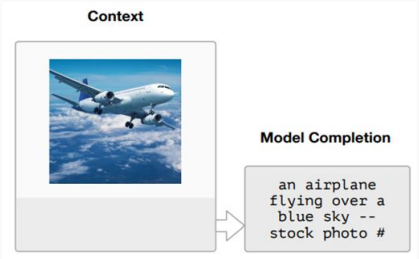
- Train on Conceptual Captions, transfer to VQAv2.
- Conceptual Captions consist of 3 Million image-caption pairs.

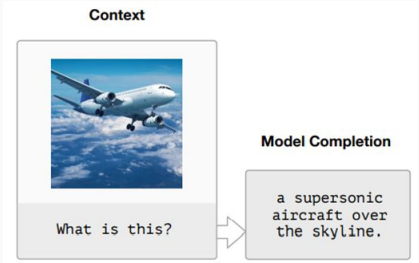
n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

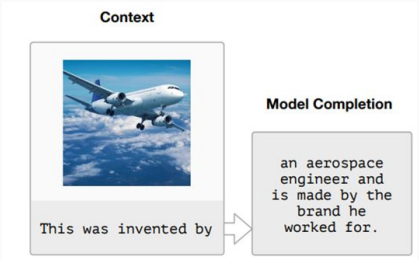
Experiment

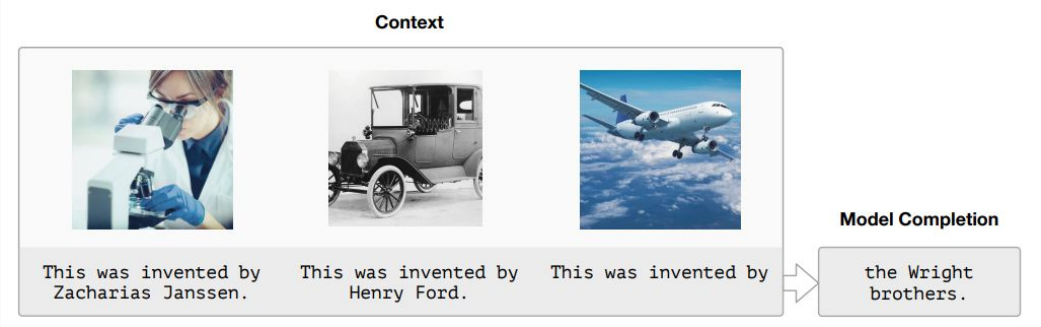
Encyclopedic Knowledge

- transfer of factual knowledge (all knowledge of named entities comes from language model pretraining).





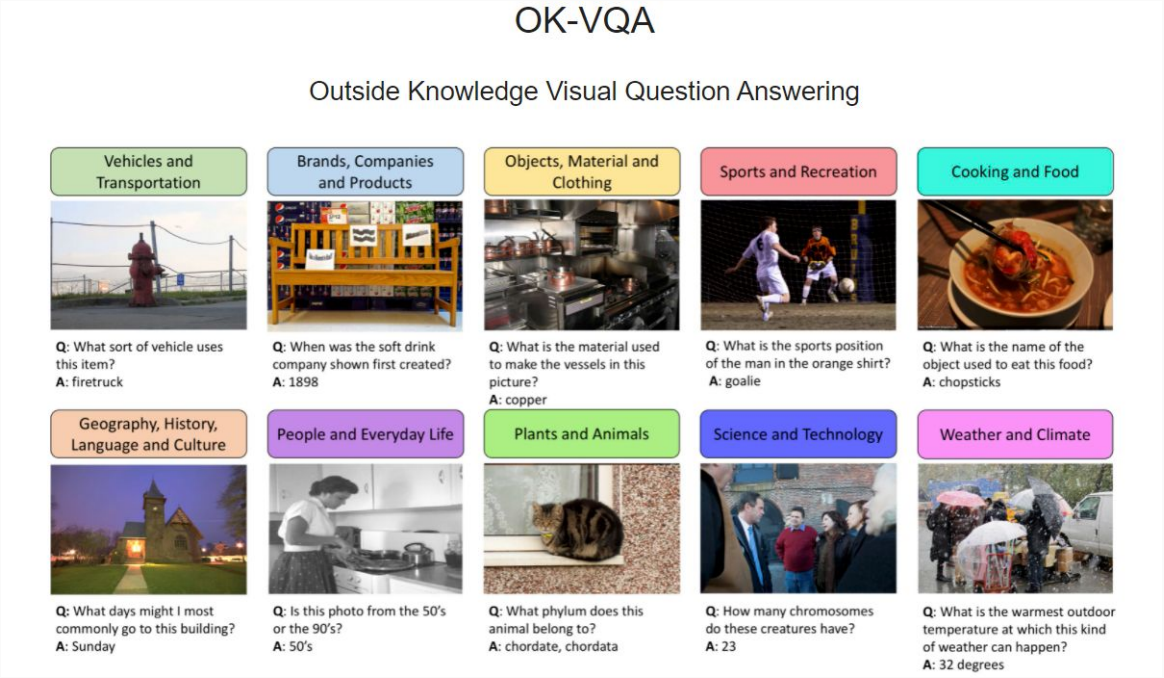




Experiment

Encyclopedic Knowledge

- Train on Conceptual Captions, transfer on OK-VQA.



Experiment

Encyclopedic Knowledge

- Conventional Frozen is in 7B parameters.

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	5.9	9.7	12.6	✗
Frozen 400mLM	4.0	5.9	6.6	✗
Frozen finetuned	4.2	4.1	4.6	✗
Frozen train-blind	3.3	7.2	0.0	✗
Frozen VQA	19.6	–	–	✗
Frozen VQA-blind	12.5	–	–	✗
MAVEx [42]	39.4	–	–	✓

Experiment

Encyclopedic Knowledge

- Finetined model fails again.

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	5.9	9.7	12.6	✗
Frozen 400mLM	4.0	5.9	6.6	✗
Frozen finetuned	4.2	4.1	4.6	✗
Frozen train-blind	3.3	7.2	0.0	✗
Frozen VQA	19.6	–	–	✗
Frozen VQA-blind	12.5	–	–	✗
MAVEx [42]	39.4	–	–	✓

Experiment

Encyclopedic Knowledge

- Train on Conceptual Captions and VQAv2.

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	5.9	9.7	12.6	✗
Frozen 400mLM	4.0	5.9	6.6	✗
Frozen finetuned	4.2	4.1	4.6	✗
Frozen train-blind	3.3	7.2	0.0	✗
Frozen VQA	19.6	–	–	✗
Frozen VQA-blind	12.5	–	–	✗
MAVEx [42]	39.4	–	–	✓

Experiment

Fast Concept Binding

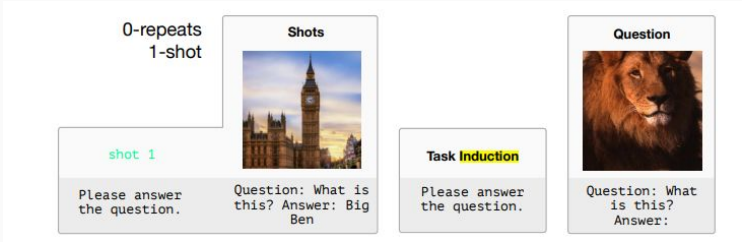
- fast-binding refers to a model's ability to associate a word with a visual category in a few shots and immediately use that word in an appropriate way (?
- Open-Ended miniImageNet: create a nonsense name.

- Real-Name miniImagenet: real-name (e.g., lion -> fruit bat)

Experiment

Fast Concept Binding

- Different variant of few-shot setting.



Experiment

Fast Concept Binding

- Two-way and Five-way few-shot:

Two-way

Task Induction	✗	✓	✓	✓	✓	✓	✓
Inner Shots	1	1	3	5	1	1	1
Repeats	0	0	0	0	1	3	5
Frozen	29.0	53.4	57.9	58.9	51.1	57.7	58.5
***Frozen* (Real-Name)**	1.7	33.7	66	66	63	65	63.7
Frozen $_{\text{test-blind}}$	–	48.5	46.7	45.3	–	–	–
Frozen $_{\text{test-blind}}$ **(Real-Name)**	–	1.0	12.6	33.0	–	–	–
ANIL Baseline [31]	–	73.9	81.7	84.2	–	–	–

Five-way

Task Induction	✗	✓	✓	✓	✓	✓	✓
Inner Shots	1	1	3	5	1	1	1
Repeats	0	0	0	0	1	3	5
Frozen	18.0	20.2	22.3	21.3	21.4	21.6	20.9
***Frozen* (Real-Name)**	0.9	14.5	34.7	33.8	33.8	33.3	32.8
Frozen $_{\text{test-blind}}$	–	18.6	19.9	19.8	–	–	–
Frozen $_{\text{test-blind}}$ **(Real-Name)**	–	4.6	22.6	20.8	–	–	–
ANIL Baseline [31]	–	45.5	57.7	62.6	–	–	–

Experiment

Fast-VQA and Real-Fast-VQA

- Two-way and Five-way few-shot:

Discussion Questions

- Place some discussion Qs here