

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



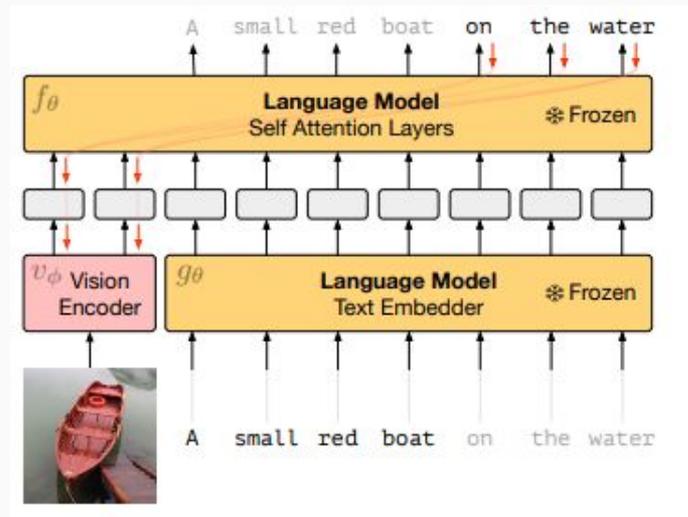
Introduction

- 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



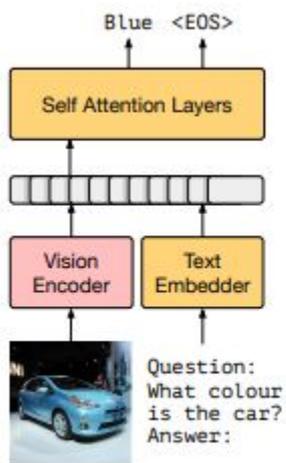
Introduction

- *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.

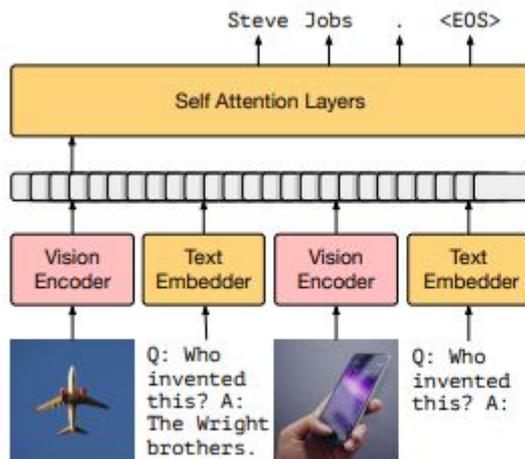


Introduction

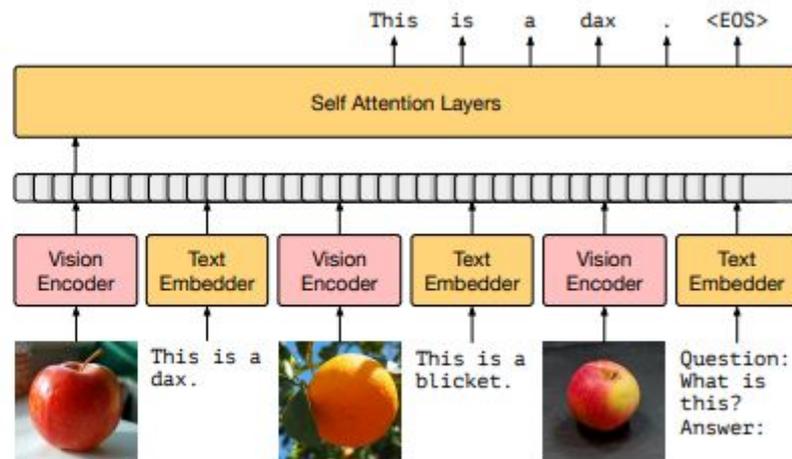
- Rapid task adaptation - rapid adaptation to new tasks, typically based on a few examples
- Encyclopedic knowledge - fast access to general knowledge that may be captured in text
- Fast concept binding - fast binding of visual and linguistic elements



(a) 0-shot VQA



(b) 1-shot outside-knowledge VQA



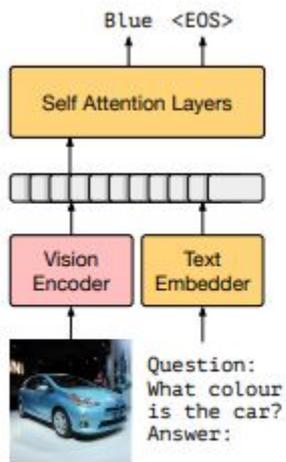
(c) Few-shot image classification

Introduction

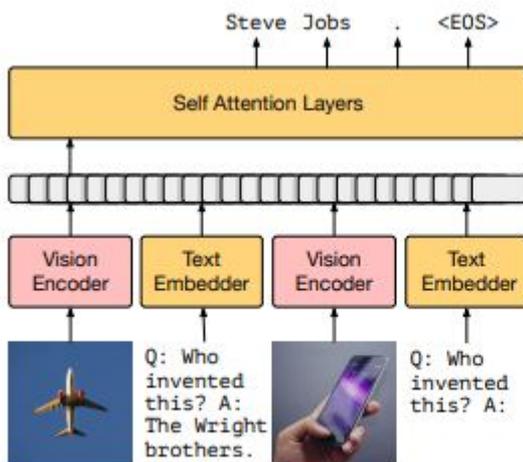
- Contributions

- *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

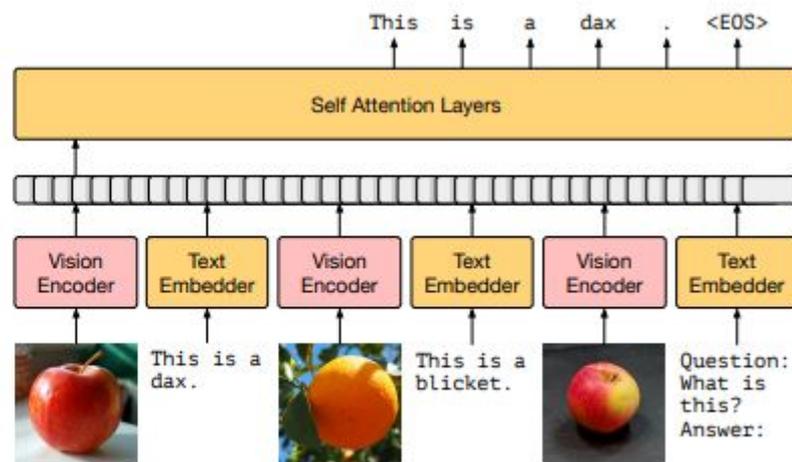
- Inference-Time interface for *Frozen*



(a) 0-shot VQA



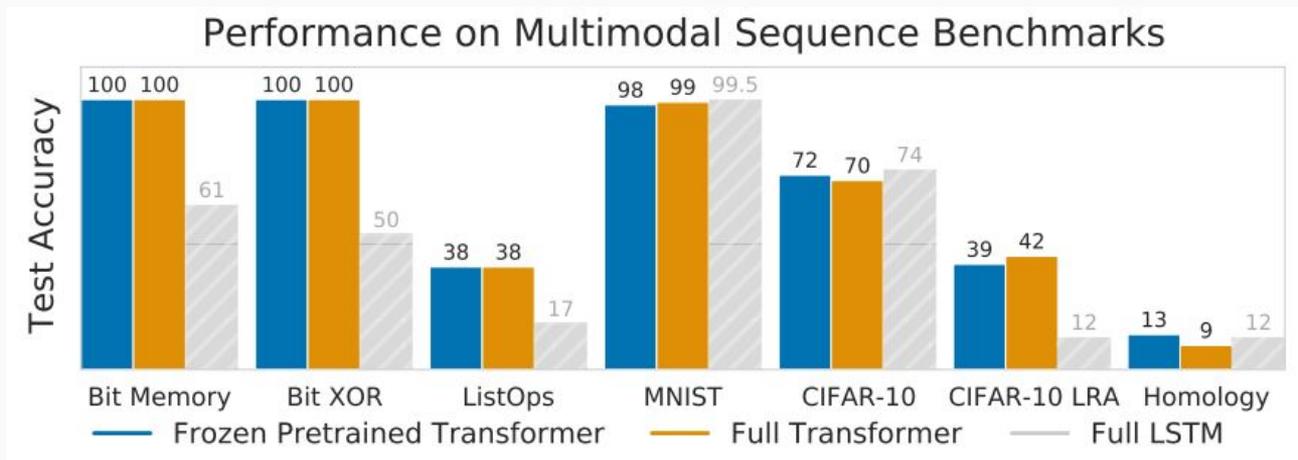
(b) 1-shot outside-knowledge VQA



(c) Few-shot image classification

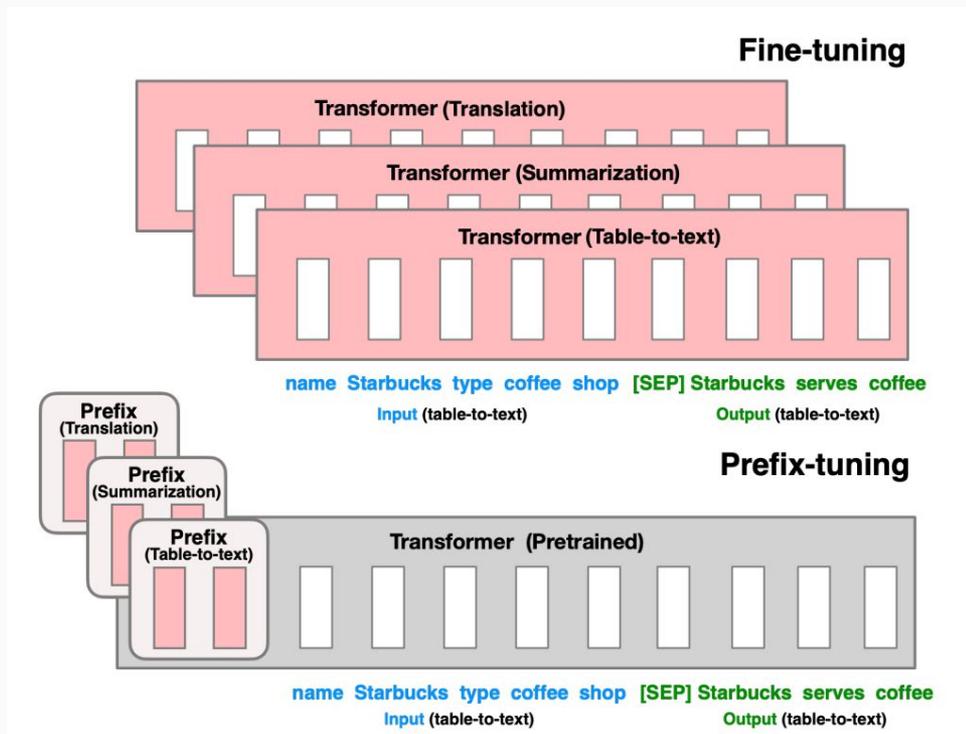
Related Work

- 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.



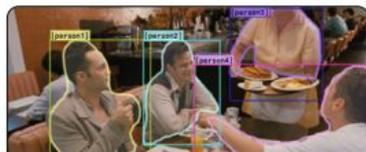
Related Work

- 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



Is there something to cut the vegetables with?

VQA



Why is [person471] pointing at [person143]?

- a) He is telling [person471] that [person143] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person143].
- d) He is giving [person143] directions.

VCR Q→A

Rationale: a) is correct because...

- a) [person143] has the pancakes in front of him.
- b) [person471] is taking everyone's order and asked for clarification.
- c) [person471] is looking at the pancakes both she and [person143] are smiling slightly.
- d) [person471] is delivering food to the table, and she might not know whose order is whose.

VCR QA→R



Guy in yellow dribbling ball

Referring Expressions

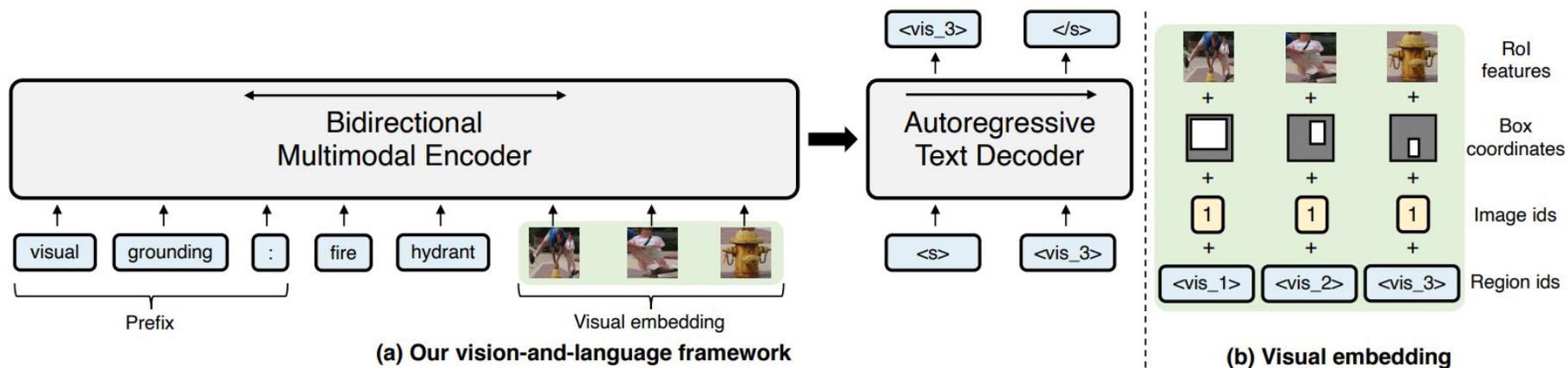


A large bus sitting next to a very tall building.

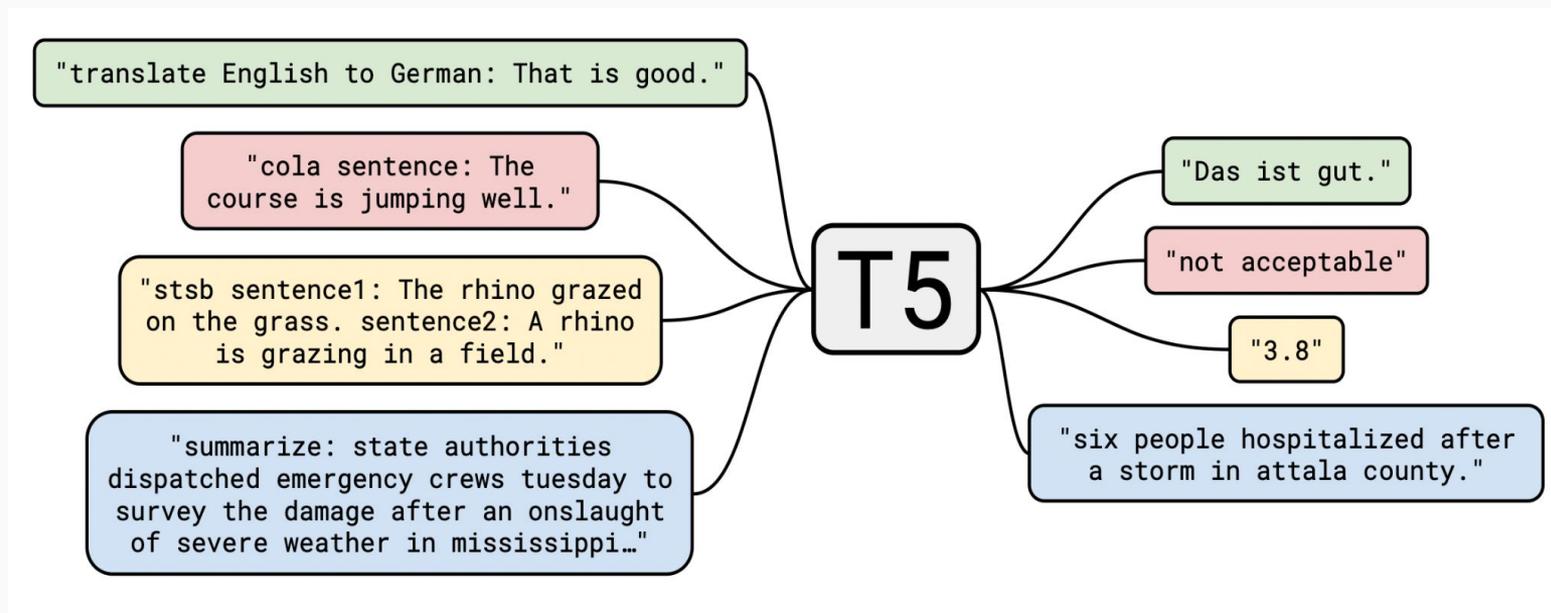
Caption-Based Image Retrieval

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

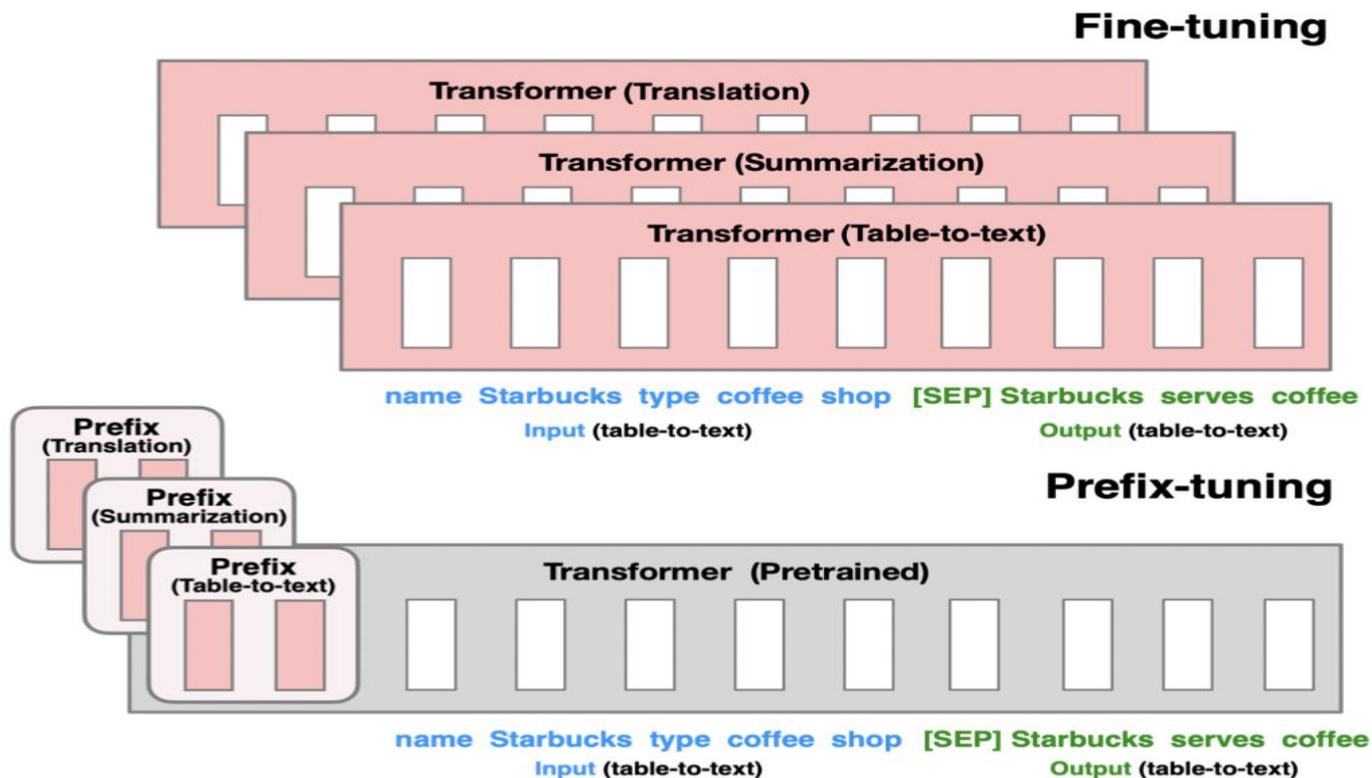


Background: Prompt/ Prefix



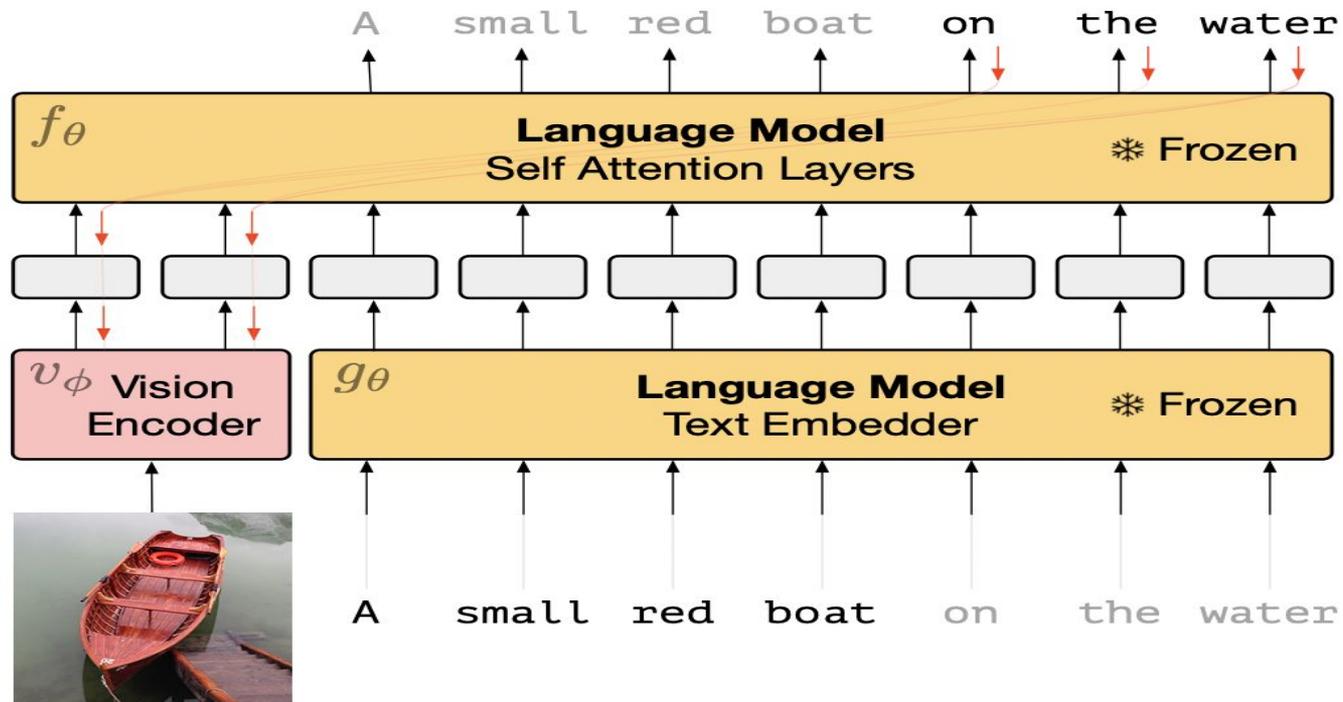
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Background: Prefix Tuning



Prefix-Tuning: Optimizing Continuous Prompts for Generation

The Frozen Method



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

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

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

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

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

$$\begin{aligned}\log p_{\theta, \phi}(\mathbf{y} | \mathbf{x}) &= \sum_l \log p_{\theta, \phi}(y_l | \mathbf{x}, y_1, y_2, \dots, y_{l-1}) \\ &= \sum_l f_{\theta}(i_1, i_2, \dots, i_n, t_1, t_2, \dots, t_{l-1})_{y_l}\end{aligned}$$

Task induction, Shots, Repeats

0-repeats
1-shot

shot 1

Please answer
the question.

Shots



Question: What is
this? Answer: Big
Ben

0-repeats
0-shots

Task Induction

Answer with
lion or dog.

Question



Question: What
is this?
Answer:

Question



Question: What
is this?
Answer:

Task Induction

Please answer
the question.

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



This is a
blicket.

inner-shot 1



This is a dax.

inner-shot 2



This is a
blicket.

inner-shot 2



This is a dax.

**Question
from ImageNet**



Q: What is this?
A: This is a

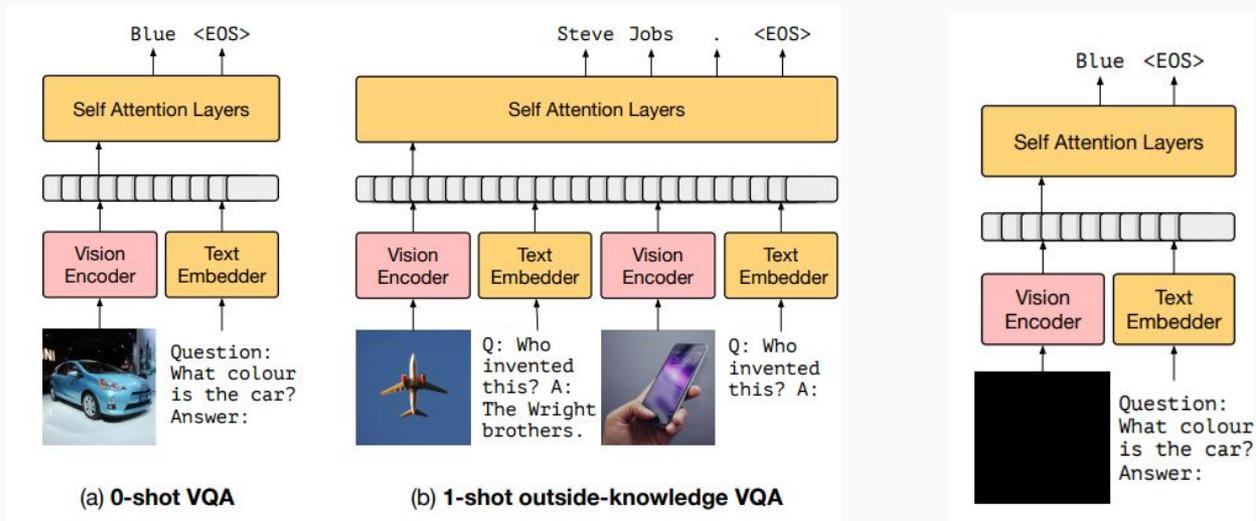
Model Completion

blicket.

Experiment

Zero-shot transfer from captioning to VQA:

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



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
<i>Frozen</i>	29.5	35.7	38.2	\times
<i>Frozen</i> <small>scratch</small>	0.0	0.0	0.0	\times
<i>Frozen</i> <small>finetuned</small>	24.0	28.2	29.2	\times
<i>Frozen</i> <small>train-blind</small>	26.2	33.5	33.3	\times
<i>Frozen</i> <small>VQA</small>	48.4	–	–	✓
<i>Frozen</i> <small>VQA-blind</small>	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

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
<i>Frozen</i>	29.5	35.7	38.2	✗
<i>Frozen</i> <small>scratch</small>	0.0	0.0	0.0	✗
<i>Frozen</i> <small>finetuned</small>	24.0	28.2	29.2	✗
<i>Frozen</i> <small>train-blind</small>	26.2	33.5	33.3	✗
<i>Frozen</i> <small>VQA</small>	48.4	–	–	✓
<i>Frozen</i> <small>VQA-blind</small>	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

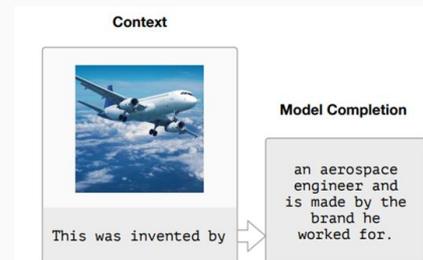
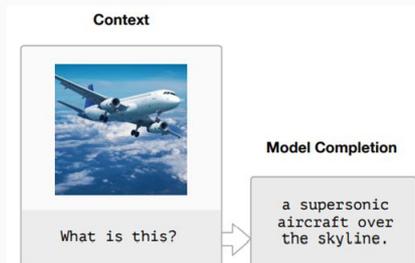
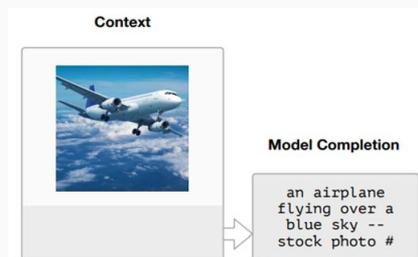
Zero-shot transfer from captioning to VQA:

- 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
<i>Frozen</i>	29.5	35.7	38.2	✗
<i>Frozen</i> <small>scratch</small>	0.0	0.0	0.0	✗
<i>Frozen</i> <small>finetuned</small>	24.0	28.2	29.2	✗
<i>Frozen</i> <small>train-blind</small>	26.2	33.5	33.3	✗
<i>Frozen</i> <small>VQA</small>	48.4	–	–	✓
<i>Frozen</i> <small>VQA-blind</small>	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

Encyclopedic Knowledge

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



Encyclopedic Knowledge

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

OK-VQA

Outside Knowledge Visual Question Answering

Vehicles and Transportation



Q: What sort of vehicle uses this item?
A: firetruck

Brands, Companies and Products



Q: When was the soft drink company shown first created?
A: 1898

Objects, Material and Clothing



Q: What is the material used to make the vessels in this picture?
A: copper

Sports and Recreation



Q: What is the sports position of the man in the orange shirt?
A: goalie

Cooking and Food



Q: What is the name of the object used to eat this food?
A: chopsticks

Geography, History, Language and Culture



Q: What days might I most commonly go to this building?
A: Sunday

People and Everyday Life



Q: Is this photo from the 50's or the 90's?
A: 50's

Plants and Animals



Q: What phylum does this animal belong to?
A: chordate, chordata

Science and Technology



Q: How many chromosomes do these creatures have?
A: 23

Weather and Climate



Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

Encyclopedic Knowledge

- Conventional Frozen is in 7B parameters.

n-shot Acc.	n=0	n=1	n=4	τ
<i>Frozen</i>	5.9	9.7	12.6	\times
<i>Frozen</i> 400mLM	4.0	5.9	6.6	\times
<i>Frozen</i> finetuned	4.2	4.1	4.6	\times
<i>Frozen</i> train-blind	3.3	7.2	0.0	\times
<i>Frozen</i> VQA	19.6	–	–	\times
<i>Frozen</i> VQA-blind	12.5	–	–	\times
MAVE_x [42]	39.4	–	–	\checkmark

Encyclopedic Knowledge

- Finetuned model fails again.

n-shot Acc.	n=0	n=1	n=4	τ
<i>Frozen</i>	5.9	9.7	12.6	\times
<i>Frozen</i> 400mLM	4.0	5.9	6.6	\times
<i>Frozen</i> finetuned	4.2	4.1	4.6	\times
<i>Frozen</i> train-blind	3.3	7.2	0.0	\times
<i>Frozen</i> VQA	19.6	–	–	\times
<i>Frozen</i> VQA-blind	12.5	–	–	\times
MAVE_x [42]	39.4	–	–	\checkmark

Encyclopedic Knowledge

- Train on Conceptual Captions and VQAv2.

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

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 minilmageNet: create a nonsense name.



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

Fast Concept Binding

- Different variant of few-shot setting.

0-repeats
0-shots

Task Induction
Answer with
Lion or dog.

Question

Question: What
is this?
Answer:

1-repeats
1-shot

repeat 0 Shots repeat 1

Task Induction
Please answer
the question.

Question

Question: What is
this?
Answer:

0-repeats
1-shot

shot 1

Shots


Task Induction
Please answer
the question.

Question

Question: What is
this?
Answer:

0-repeats
2-shots

shot 1

Shots


shot 2

Shots


Task Induction
Please answer
the question.

Question

Question: What
is this?
Answer:

Fast Concept Binding

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

Two-way

Task Induction	X	✓	✓	✓	✓	✓	✓
Inner Shots	1	1	3	5	1	1	1
Repeats	0	0	0	0	1	3	5
<i>Frozen</i>	29.0	53.4	57.9	58.9	51.1	57.7	58.5
<i>Frozen</i> (Real-Name)	1.7	33.7	66	66	63	65	63.7
<i>Frozen</i> test-blind	–	48.5	46.7	45.3	–	–	–
<i>Frozen</i> test-blind (Real-Name)	–	1.0	12.6	33.0	–	–	–
ANIL Baseline [31]	–	73.9	81.7	84.2	–	–	–

Five-way

Task Induction	X	✓	✓	✓	✓	✓	✓
Inner Shots	1	1	3	5	1	1	1
Repeats	0	0	0	0	1	3	5
<i>Frozen</i>	18.0	20.2	22.3	21.3	21.4	21.6	20.9
<i>Frozen</i> (Real-Name)	0.9	14.5	34.7	33.8	33.8	33.3	32.8
<i>Frozen</i> test-blind	–	18.6	19.9	19.8	–	–	–
<i>Frozen</i> test-blind (Real-Name)	–	4.6	22.6	20.8	–	–	–
ANIL Baseline [31]	–	45.5	57.7	62.6	–	–	–

Fast-VQA and Real-Fast-VQA

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

(b) Fast VQA

0-repeats
0-shots
2-way
0-repeats
2-inner-shots

**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 VisualGenome**

	blicket (vase)
	dax (table)

Model Completion

Q: What is the
dax made of? A: wood

Discussion Questions

- Place some discussion Qs here