Flamingo DeepMind: Обзор

Научно-исследовательский семинар МОП

Field & Tasks

Open-ended tasks:

visual question-answering (prompt: a question which it has to answer);

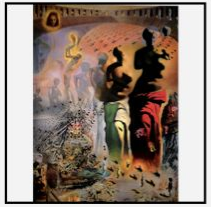
captioning tasks (the ability to describe a scene or an event from image or video)

Close-ended tasks:

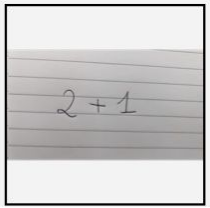
multiple-choice visual question-answering

For tasks lying anywhere on this spectrum, a *single* Flamingo model can achieve a new state of the art with few-shot learning, simply by prompting the model with task-specific examples

Cherries

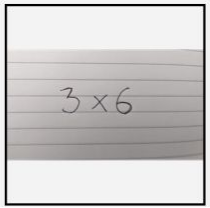
Input Prompt						Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is	**a flamingo. They are found in the Caribbean and South America.**
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.		What is the name of the city where this was painted? Answer:	**Arles.**
	Output: "Underground"		Output: "Congress"		Output:	**"Soulomes"**

Cherries



2+1=3

5+6=11



→ **3x6=18**

pandas: 3

dogs: 2

→ **giraffes: 4**

What happens to the man after hitting the ball? Answer:

→ **he falls down.**

This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?

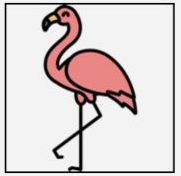
It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.



What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.

Related Works: Image Language

- (ImageBERT: 2020) Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data – require fine-tuning on new tasks
- (TCL 2022) Vision-Language Pre-Training - Triple Contrastive Learning – contrastive learning, our paper does not use it
- **Key idea:** freeze the pretrained LM weights (large model) to prevent catastrophic forgetting
  - A New AI Trend: **Chinchilla (70B)** Greatly Outperforms GPT-3 (175B) and Gopher (280B)
- **Web-scale vision and language training datasets**
  - CM3: A CAUSAL MASKED MULTIMODAL MODEL OF THE INTERNET – primarily evaluates on language-only benchmarks in a zero-shot or fine-tuned setup
  - Our paper emphasize few-shot learning and vision tasks

Basement

Pretrain the vision encoder:

a contrastive objective on our datasets of image and text pairs, using the two-term contrastive loss

- Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021

Similar to **Perceiver** and **DETR**, we learn a predefined number of latent input queries which are fed to a Transformer and cross-attend to the visual features.

- Perceiver: General perception with iterative attention. In *ICML*, 2021

Language Model

- Chinchilla (70B): Training compute-optimal large language models. arXiv:2203.15556, 2022

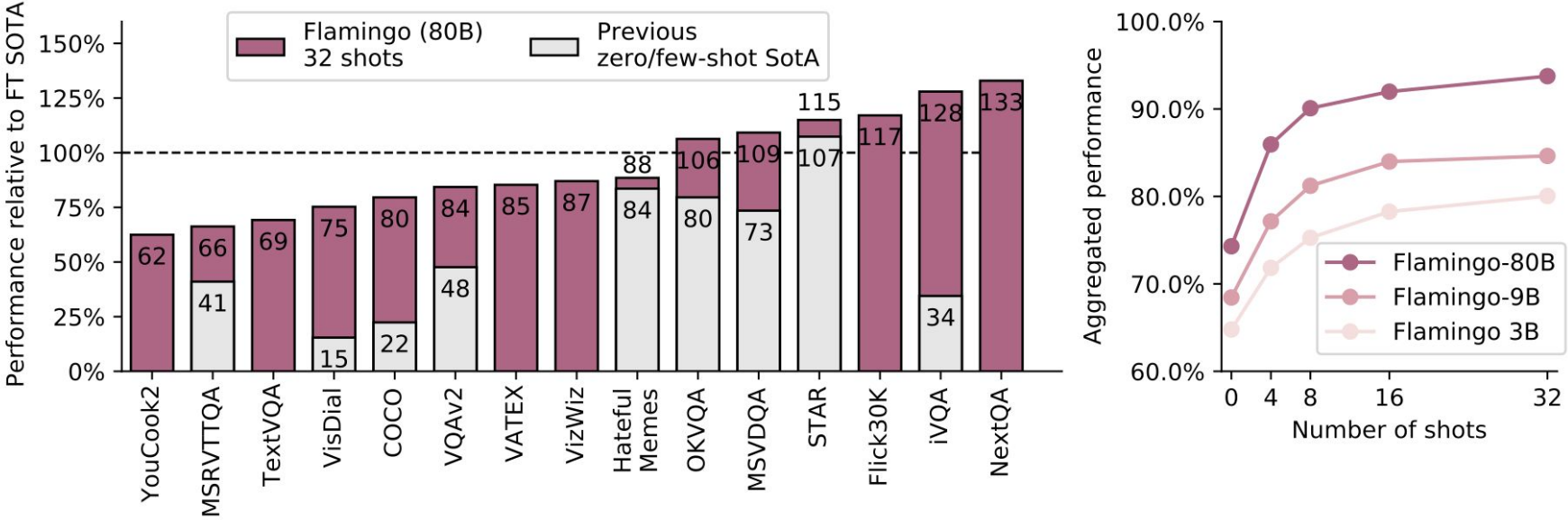


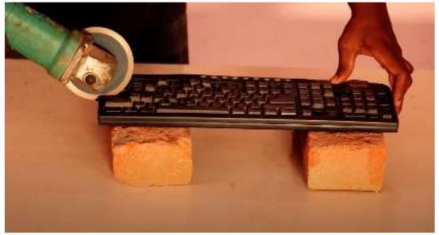
Figure 2: **Flamingo results overview.** *Left*: Our largest model, dubbed *Flamingo*, outperforms state-of-the-art fine-tuned models on 6 of the 16 tasks we consider with no fine-tuning. For the 9 tasks with published few-shot results, *Flamingo* sets the new few-shot state of the art. *Note:* We omit RareAct, our 16th benchmark, as it is a zero-shot benchmark with no available fine-tuned results to compare to. *Right*: Flamingo performance improves with model size and number of shots.

Zero-Shot Cross-Modal Retrieval on COCO 2014

Rank	Model	Image-to-text R@1 ↑	Image-to-text R@5	Image-to-text R@10	Text-to-image R@1	Text-to-image R@5	Text-to-image R@10	Paper	Code	Result	Year	Tags
1	**TCL**	71.4	90.8	95.4	53.5	79.0	87.1	Vision-Language Pre-Training with Triple Contrastive Learning			2022	
2	**ALBEF**	68.7	89.5	94.7	50.1	76.4	84.5	Align before Fuse: Vision and Language Representation Learning with Momentum Distillation			2021	
3	**CoCa**	66.3	86.2	91.8	51.2	74.2	82.0	CoCa: Contrastive Captioners are Image-Text Foundation Models			2022	
4	**Flamingo**	65.9	87.3	92.9	48.0	73.3	82.1	Flamingo: a Visual Language Model for Few-Shot Learning			2022	

Few-shot: #1 Action Recognition on RareAct

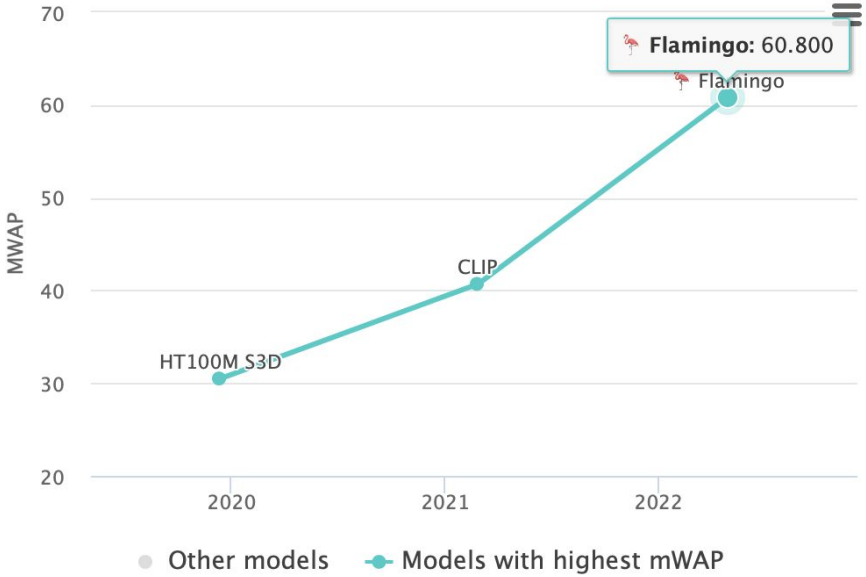
Blend Phone



Cut Keyboard

Unplug Oven

Microwave Shoes



Strengths

- Useful tricks are proposed for bridging pre-trained vision models and language models, including vision feature resampling, inserting cross-modal attention into language model

- The paper is well written in general and easy to understand

- A new dataset is also mentioned in the paper to pre-train the Flamingo model

Weaknesses: Ethics

- Highly **publicized** by the authors' organization **before/during the review period**

- The dataset, the trained model are **proprietary**. Community unable to verify/reproduce the specific results presented in the paper.
  - some of the experiments have been performed on evaluation benchmarks (VQAv2, VizWiz, STAR, VisDial, TextVQA, etc.) where **test sets are hidden** and they obtain their numbers by submitting to the official leaderboard;
  - additional experiments on a public dataset may be included for the camera ready version of the paper

- "CM3 follows the paper's similar approach" - REVIEWS

  It can do image generation in addition to text generation, not mentioned few-shot ability

The Authors: Research scientists DeepMind, London

Jean-Baptiste Alayrac

Ecole polytechnique
Telecom ParisTech, 2015
Masters in CV (ENS Cachan)

focuses on structured learning from video natural language
Perceiver IO
Thinking Fast and Slow: Efficient text-to-visual retrieval with transformers, 2021
Text-Video Embeddings

Jeff Donahue

UT Austin, 2011(BSc)
UC Berkeley, 2017 (PhD)

End-to-end adversarial text-to-speech, 2020
Generative adversarial networks
Rich feature hierarchies in Object Detection + R Girshick, 2014

Pauline Luc

CentraleSupelec, 2015 (BSc)
Masters in CV (ENS Cachan)
PhD Facebook AI, Paris

Towards learning universal audio representations, 2022
Predicting Deeper into the Future of Semantic Segmentation, 2017
Semantic segmentation using adversarial networks, 2016

Antoine Miech

PhD WILLOW Inria Paris DI ENS
(Ecole Normale Supérieure,
with Ivan Laptev)

Just Ask: Learning to Answer Questions from Millions of Narrated Videos, 2021
Thinking Fast and Slow: Efficient Text-to-Visual Retrieval with Transformers, 2021
video understanding and weakly-supervised machine learning

Advisors: Ivan Laptev & Josef Sivic

INRIA - Willow Project
Département d'Informatique de l'Ecole Normale Supérieure

Senior researchers at INRIA Paris, Ivan is team leader of the WILLOW project-team

- area chair for CVPR, ICCV and ECCV
- program chair of ICCV'23

! Machines Can See summits (2017-2021), Russia (head of scientific board at VisionLabs)

Josef Sivic is also Senior researcher and principal investigator

Intelligent Machine Perception project
Czech Institute of Informatics, Robotics, and Cybernetics
Czech Technical University in Prague

Zero-Shot Video Question Answering via Frozen Bidirectional Language Models (2022), A. Yang, A. Miech, J. Sivic, I. Laptev and C. Schmid;

TubeDETR: Spatio-Temporal Video Grounding with Transformers (2022), A. Yang, A. Miech, J. Sivic, I. Laptev and C. Schmid;

History Aware Multimodal Transformer for Vision-and-Language Navigation (2021), S. Chen, P.-L. Guhur, C. Schmid and I. Laptev;