New Frontiers in Multimodal Grounding

Jack Hessel
AI2

A bit about me

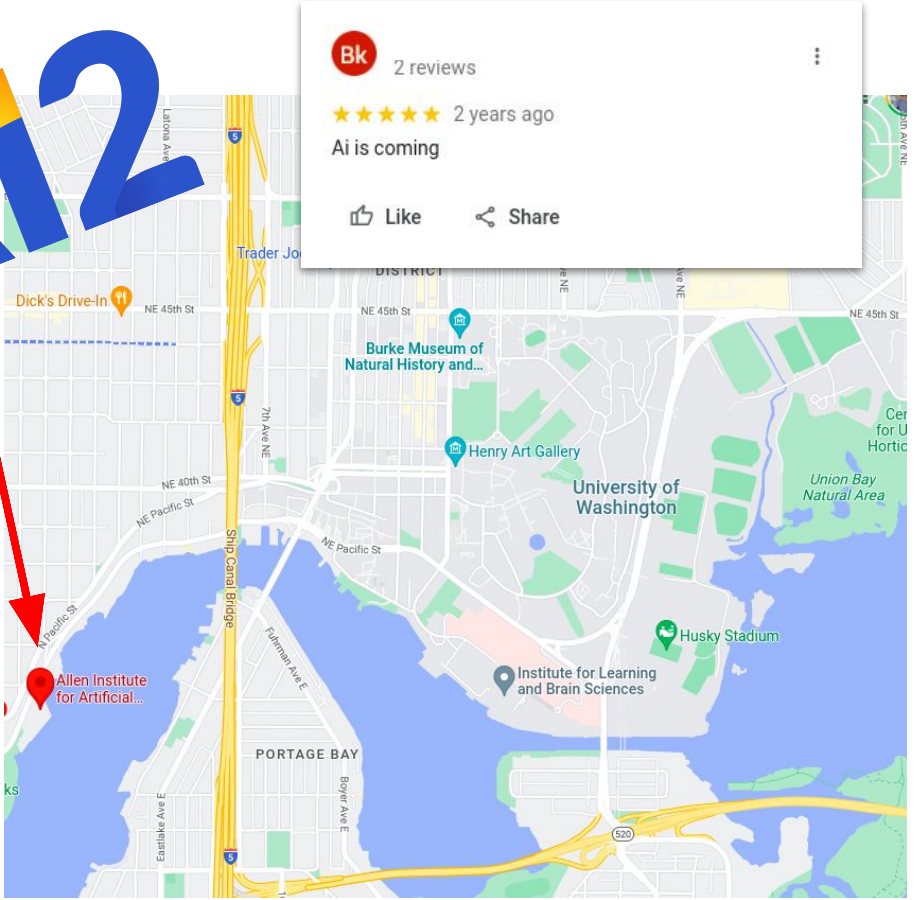
- Research Scientist at AI2
    - I write NLP papers for a living!
- Mostly focus on: multimodal models, datasets, etc.
- The last time I was in a classroom was mid-2020 😱 when I was defending my PhD
    - no one was allowed to physically attend... so this is the first time I've been in a classroom with others since Jan 2020 (?)!

(Me, without a mask)

A bit about AI2

- **AI2** = "Allen Institute for Artificial Intelligence"
- **Founded** by Microsoft co-founder Paul Allen
- **Mission:** *"to contribute to humanity through high-impact AI research and engineering."*
- **Mosaic,** my team, is lead by Prof. Yejin Choi from CSE. Our goal: commonsense reasoning!

MOSAIC
COMMONSENSE



New Frontiers in Multimodal Grounding

Jack Hessel
AI2

What is multimodal grounding?

What is multimodal grounding?

A collection of tasks requiring connection between more than one modality.

What is multimodal grounding?

A collection of tasks requiring connection between more than one modality.

Alt-text Generation

Chrome's new AI feature solves one of the web's eternal problems

To help blind and low-vision users, Google is using machine learning to generate descriptions for millions of images.

[Wu et al. 2017;
Sharma et al. 2019]

What is multimodal grounding?

A collection of tasks requiring connection between more than one modality.

<u>Alt-text Generation</u>

[Wu et al. 2017;
Sharma et al. 2019]

<u>Human-Robot Interaction</u>

"Here are the yellow ones"

[Matuszek et al. 2012]

What is multimodal grounding?

A collection of tasks requiring connection between more than one modality.

Alt-text Generation

[Wu et al. 2017;
Sharma et al. 2019]

Human-Robot Interaction

"Here are the yellow ones"

[Matuszek et al. 2012]

Web Video Parsing

[Kim et al. 2014]

Why study multimodal grounding?

Cross-modal reasoning is easy for humans, hard for computers

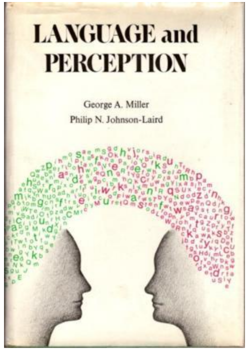
[Zellers et al. 2019]

Why study multimodal grounding?

Cross-modal reasoning is important beyond AI

Cognitive psychology work
since at least the 1970s.

[Miller and Johnson-Laird 1976]



"Symbol Grounding Problem"

[Harnad 1990]

"How are those symbols
(e.g., the words in our heads)
connected to the things they refer to?"

What is multimodal grounding?

A collection of tasks requiring connection between more than one modality.

What is multimodal grounding?

A collection of tasks requiring **connection** between more than one modality.

Does my multimodal model learn cross-modal interactions?

It’s harder to tell than you might think!

Jack Hessel and Lillian Lee
EMNLP 2020

Setting:

t How many cats are there?

v 

$$f(t, v)$$

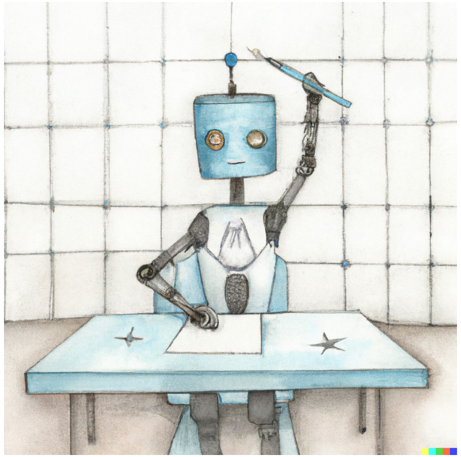
Some prediction

Setting:

t

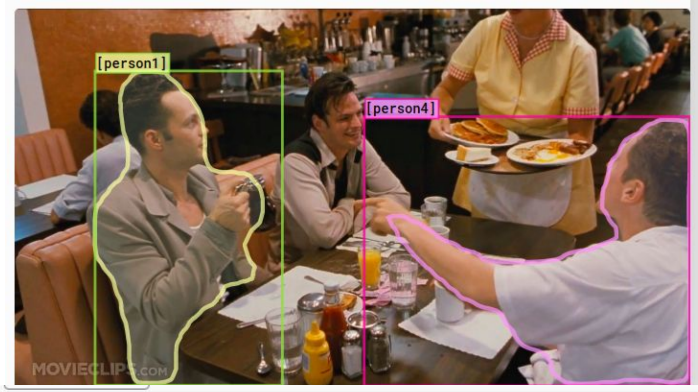
How
many
cats
are
there?

v



Some prediction

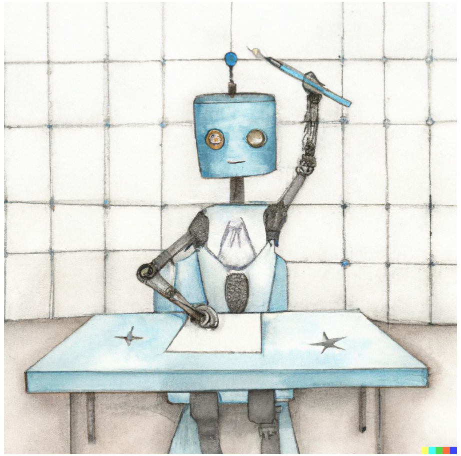
v



Why is [person4] pointing at [person1]?

t

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.



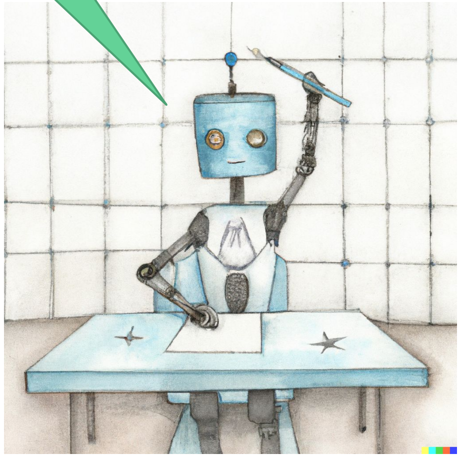
v

Why is [person4] pointing at [person1]?

t

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

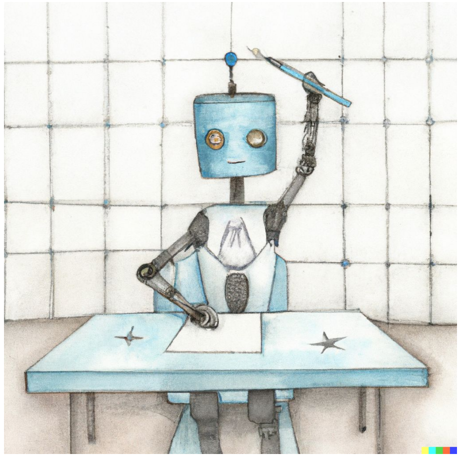
I think it's "A"!!!



t

1. What would happen if [bird1] turned and bit [person1] ?

a) [person1] would probably cat call [person2] .	36.3%
b) Cult members like [person1] would try to capture them.	0.7%
c) [person1] could get injured by the animal.	4.2%
d) [person1] would stop smiling and probably yell.	58.7%

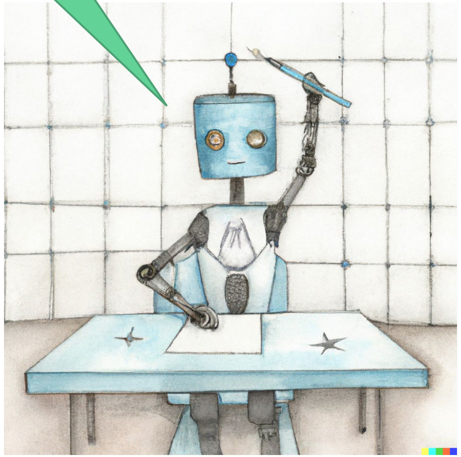


I think it's "D"!!!

t

1. What would happen if [bird1] turned and bit [person1] ?

a) [person1] would probably cat call [person2] .	36.3%
b) Cult members like [person1] would try to capture them.	0.7%
c) [person1] could get injured by the animal.	4.2%
d) [person1] would stop smiling and probably yell.	58.7%

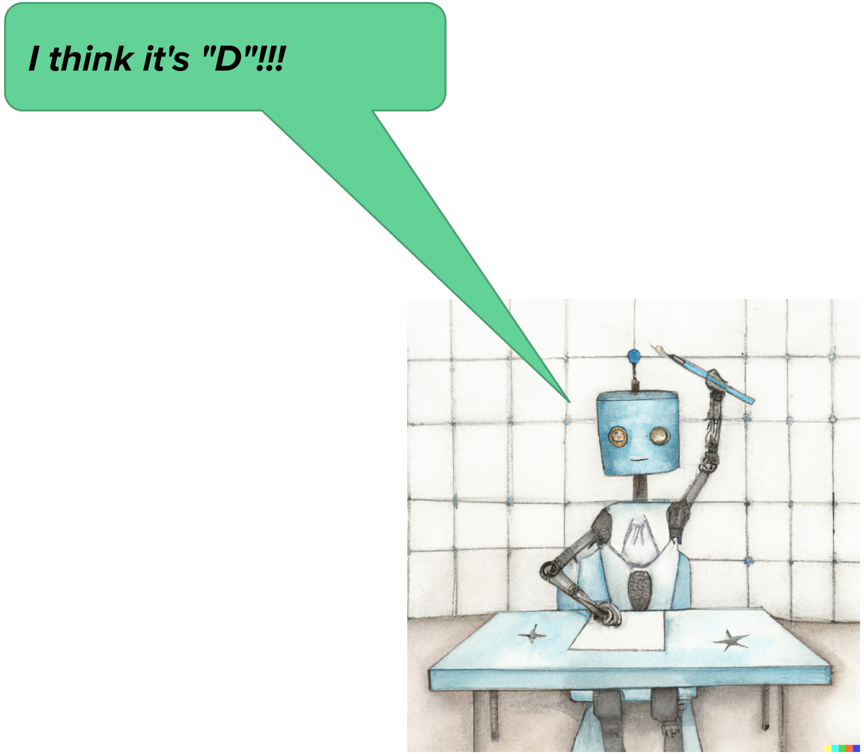


v

t

1. What would happen if [bird1] turned and bit [person1] ?

- a) [person1] would probably cat call [person2] . 36.3%
- b) Cult members like [person1] would try to capture them. 0.7%
- c) [person1] could get injured by the animal. 4.2%
- d) [person1] would stop smiling and probably yell. 58.7%



An important question:

Does a given image-text task

<u>require learning cross-modal connections?</u>

An important question:

Does a given image-text task
<u>require learning cross-modal connections?</u>

Making the V in VQA Matter:
Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal[*†] Tejas Khot[*†] Douglas Summers-Stay[‡] Dhruv Batra[§] Devi Parikh[§]

<u>Strategy:</u>
To defeat models that ignore the image, rebalance the dataset!

A design strategy seen in:
- NLVR2 (Suhr et al., 2019)
- GQA (Hudson and Manning, 2019)
... and more!

1. What would happen if [bird1] turned and bit [person1] ?

a) Cult members like [person1] would try to capture them.	**0.7%**
b) [person1] could get injured by the animal.	**4.2%**
c) [person1] would stop smiling and probably yell.	**58.7%**

1. What would happen if [bird1] turned and bit [person1] ?

a) Cult members like [person1] would try to capture them.	**0.7%**
b) [person1] could get injured by the animal.	**4.2%**
c) [person1] would stop smiling and probably yell.	**58.7%**

1. What would happen if [bird1] turned and bit [person1] ?

a) Cult members like [person1] would try to capture them. **0.7%**

b) [person1] could get injured by the animal. **4.2%**

c) [person1] would stop smiling and probably yell. **58.7%**

1. What would happen if [bird1] turned and bit [person1] ?

a) Cult members like [person1] would try to capture them.	0.7%
b) [person1] could get injured by the animal.	4.2%
c) [person1] would stop smiling and probably yell.	58.7%

But not all tasks can be re-balanced...

Proposing work	Task (structure)	Abbv.	# image+text
Kruk et al. (2019)	Instagram		
	↳ intent (7-way clf)	I-INT	1299
	↳ semiotic (7-way clf)	I-SEM	1299
	↳ contextual (7-way clf)	I-CTX	1299
Vempala and Preoţiuc-Pietro (2019)	Twitter visual-ness (4-way clf)	T-VIS	4471
Hessel et al. (2017)	Reddit popularity (Pairwise-ranking)	R-POP	88K
Borth et al. (2013)	Twitter sentiment (binary clf)	T-ST1	603
Niu et al. (2016)	Twitter sentiment (binary clf)	T-ST2	4511

[Kruk and Lubin et al. 2019]

The grass is always greener

[Hessel et al. 2017]

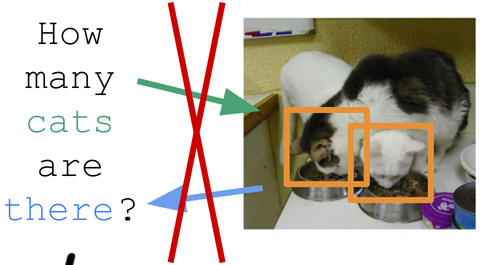
Awesome!

(b) Image adds to the tweet meaning & Text is not represented in image

[Vempala + Preoțiuc-Pietro 2019]

What does it mean to learn cross-modal connections?

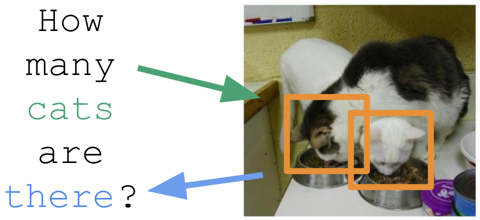
Multimodally additive model



t v

$$f(t, v) = f_t(t) + f_v(v)$$

Multimodally interactive model



t v

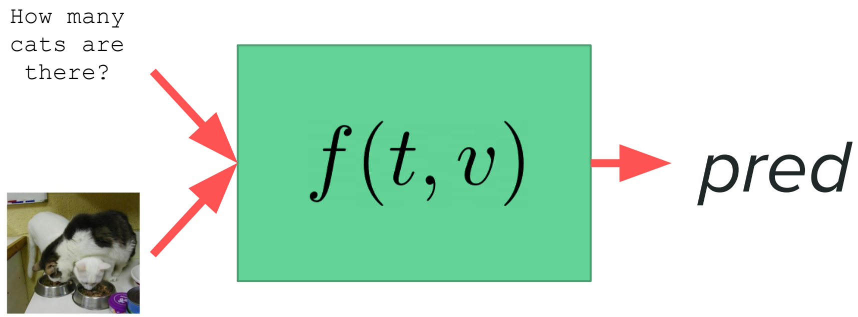
$$f(t, v) = f_{int}(t, v)$$

[Friedman 2001; Friedman et al. 2008; Hooker 2004]

Prototypical model comparisons

(numbers only for illustration, they aren't real)

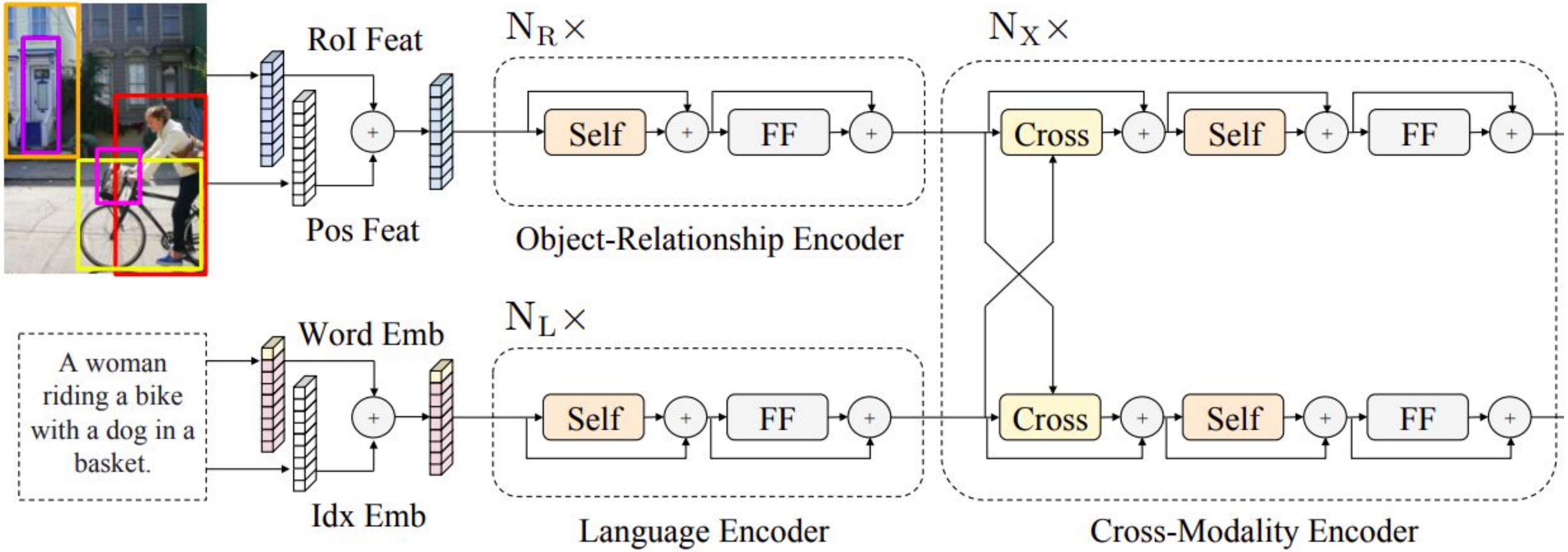
Method	Accuracy
Text Only	55
Image Only	57
Text+Image Ensemble	60
Our Fancy Method	**62**



"Because our fancy method outperforms the image text ensemble, our model is utilizing interesting cross-modal interactions/attention/etc. to produce more accurate predictions"

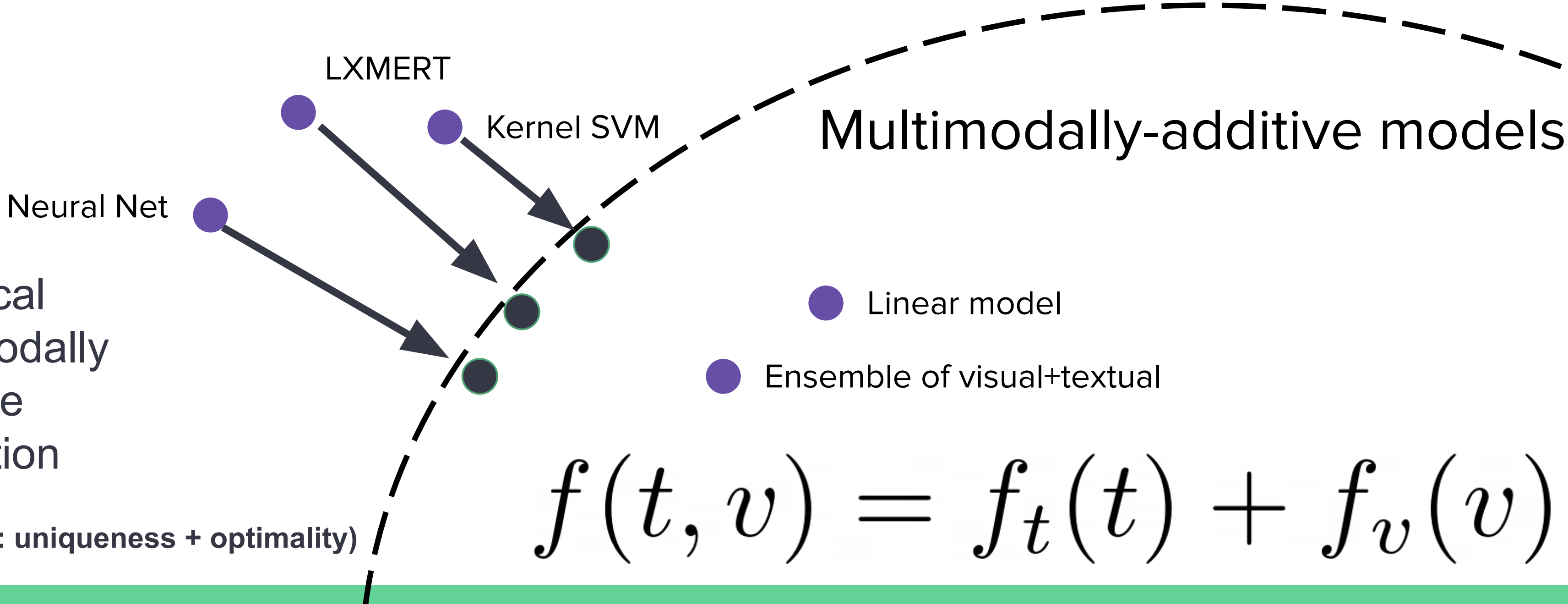
Our finding: this argument can be unreliable!

It can be difficult to tell what multimodally interactive models learn…



[LXMERT: Tan and Bansal, 2019]

Simplifying models with function projection



Empirical
Multimodally
Additive
Projection

(We prove: uniqueness + optimality)

$$f(t,v) = f_t(t) + f_v(v)$$

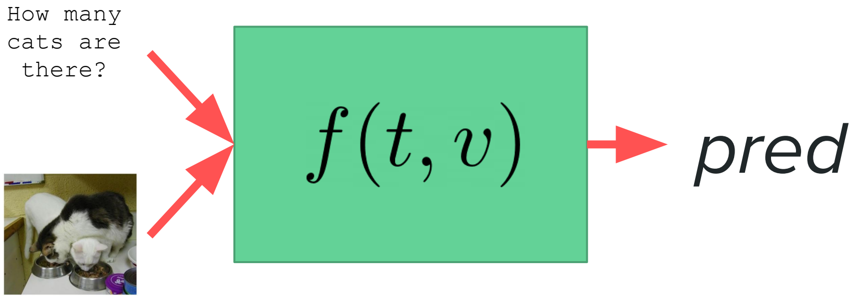
EMAP
in 20 lines of Python

```python
'''
Example implementation of EMAP
'''
import numpy as np
import collections


def emap(idx2logits):
    '''Example implementation of EMAP (more efficient ones exist)

    inputs:
      idx2logits: This nested dictionary maps from image/text indices
        function evals, i.e., idx2logits[i][j] = f(t_i, v_j)

    returns:
      projected_preds: a numpy array where projected_preds[i]
        corresponds to \hat f(t_i, v_i).
    '''
    all_logits = []
    for k, v in idx2logits.items():
        all_logits.extend(v.values())
    all_logits = np.vstack(all_logits)
    logits_mean = np.mean(all_logits, axis=0)

    reversed_idx2logits = collections.defaultdict(dict)
    for i in range(len(idx2logits)):
        for j in range(len(idx2logits[i])):
            reversed_idx2logits[j][i] = idx2logits[i][j]

    projected_preds = []
    for idx in range(len(idx2logits)):
        pred = np.mean(np.vstack(list(idx2logits[idx].values())), axis=0)
        pred += np.mean(np.vstack(list(reversed_idx2logits[idx].values())), axis=0)
        pred -= logits_mean
        projected_preds.append(pred)

    projected_preds = np.vstack(projected_preds)
    return projected_preds
```

Prototypical model comparisons

(numbers only for illustration, they aren't real)

Method	Accuracy
Text Only	55
Image Only	57
Text+Image Ensemble	60
Our Fancy Method	**62**
↳ + EMAP	???



First test: EMAP for Balanced image+text tasks

	LXMERT	+EMAP	Const.
VQA2	70.3		23.4
GQA	60.3		18.1

[Hudson and Manning, 2019]

[Goyal and Khot et al., 2017]

First test: EMAP for Balanced image+text tasks

	LXMERT	+EMAP	Const.
VQA2	70.3	40.5	23.4
GQA	60.3	41.0	18.1

GQA

VQA Visual Question Answering

[Hudson and Manning, 2019]

[Goyal and Khot et al., 2017]

Next test: EMAP for Unbalanced image+text tasks

Proposing work	Task (structure)	Abbv.	# image+text
Kruk et al. (2019)	Instagram		
	↳ intent (7-way clf)	I-INT	1299
	↳ semiotic (7-way clf)	I-SEM	1299
	↳ contextual (7-way clf)	I-CTX	1299
Vempala and Preoţiuc-Pietro (2019)	Twitter visual-ness (4-way clf)	T-VIS	4471
Hessel et al. (2017)	Reddit popularity (Pairwise-ranking)	R-POP	88K
Borth et al. (2013)	Twitter sentiment (binary clf)	T-ST1	603
Niu et al. (2016)	Twitter sentiment (binary clf)	T-ST2	4511

[Kruk and Lubin et al. 2019]

[Hessel et al. 2017]

Awesome!

(b) Image adds to the tweet meaning & Text is not represented in image

[Vempala + Preoțiuc-Pietro 2019]

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5
Linear Model (A)	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5
Linear Model (A)	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
Our Best Interactive (I)	**91.3**	**74.4**	**81.5**	**53.4**	**64.2***	**75.5**	**80.9**

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5
Linear Model (A)	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
Our Best Interactive (I)	**91.3**	**74.4**	**81.5**	**53.4**	**64.2***	**75.5**	**80.9**
↳ + EMAP (A)	**91.1**	**74.2**	**81.3**	51.0	**64.1***	**75.9**	**80.7**

Takeaway:

*report the **E**mpirical **M**ultimodally-**A**dditive **P**rojection performance!*

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5
Linear Model (A)	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
Our Best Interactive (I)	**91.3**	**74.4**	**81.5**	**53.4**	**64.2***	**75.5**	**80.9**
↳ + EMAP (A)	**91.1**	**74.2**	**81.3**	51.0	**64.1***	**75.9**	**80.7**

New Frontiers in Multimodal Grounding

Jack Hessel
AI2

How do we build powerful multimodal models...

...capable of modeling cross-modal interactions?

How do we build powerful multimodal models…

…capable of modeling cross-modal interactions?

These days: train the *biggest* possible model you can afford
on the *most* data you can grab from the web!

How do we build powerful multimodal models...

...capable of modeling cross-modal interactions?

These days: train the *biggest* possible model you can afford on the *most* data you can grab from the web!

Web data is
"the best ally we have"
--- Halevy, Norvig, and Pereira, 2009

State-of-the-art circa early 2022

(but it's harder and harder to keep up with new web-trained, large models!!)

State-of-the-art circa early 2022

(but it's harder and harder to keep up with new web-trained, large models!!)

Image+Text Tasks

[Goyal et al. 2017; Suhr et al. 2018; Hudson and Manning, 2019; Young et al. 2014]

Video+Text Tasks

[Zhukov et al. 2019; Zhou et al. 2018]

Audio+Text Tasks

[DCASE2022; Panayotov et al. 2015]

State-of-the-art circa early 2022

(but it's harder and harder to keep up with new web-trained, large models!!)

Image+Text Tasks

NLVR2 GQA VQA

[Goyal et al. 2017; Suhr et al. 2018; Hudson and Manning, 2019; Young et al. 2014]

Video+Text Tasks

Crosstask

YouCook2

[Zhukov et al. 2019; Zhou et al. 2018]

Audio+Text Tasks

LibriSpeech ASR corpus

Detection and Classification of Acoustic Scenes and Events

[DCASE2022; Panayotov et al. 2015]

3M Webly Supervised Image-Caption Pairs

Conceptual Captions

[Sharma et al. 2018]

100M Web Video Clips + ASR

HowTo100M

[Miech et al. 2019]

1000 hours of untranscribed speech

Wav2vec 2.0

Meta AI

[Baevski et al. 2020]

Biggest model I've seen recently (text only)

PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery[*] Sharan Narang[*] Jacob Devlin[*]
Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham
Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi
Sasha Tsvyashchenko Joshua Maynez Abhishek Rao[†] Parker Barnes Yi Tay
Noam Shazeer[‡] Vinodkumar Prabhakaran Emily Reif Nan Du Ben Hutchinson
Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari
Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev
Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus
Denny Zhou Daphne Ippolito David Luan[‡] Hyeontaek Lim Barret Zoph
Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick
Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz
Erica Moreira Rewon Child Oleksandr Polozov[†] Katherine Lee Zongwei Zhou
Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta[†] Jason Wei
Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

Google Research

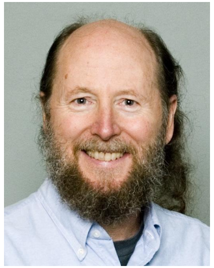
580B parameters!

(6x the number of stars in milky way!)

trained *just* to predict the next word given the ______

<aside>

http://www.incompleteideas.net/IncIdeas/BitterLesson.html



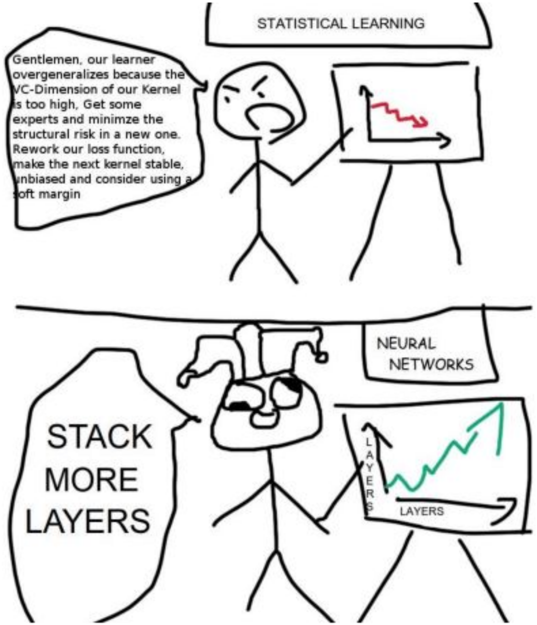
The Bitter Lesson

Rich Sutton

March 13, 2019

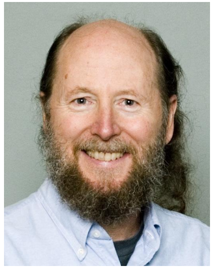
"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."

https://xkcd.com/1838/



StAcK MoRe LaYeRs

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

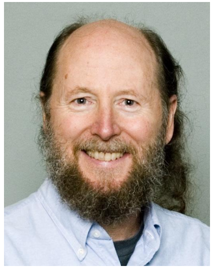


The Bitter Lesson

Rich Sutton

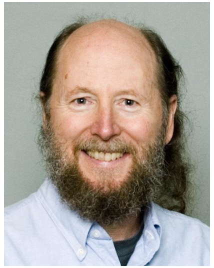
March 13, 2019

"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."



"... general methods that leverage computation are ultimately the most effective, and by a large margin."

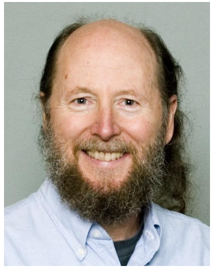
Adding sweetener 🍬 to the bitter ☕ lesson:



"... general methods that leverage computation are ultimately the most effective, and by a large margin."

Adding sweetener 🍬 to the bitter ☕ lesson:

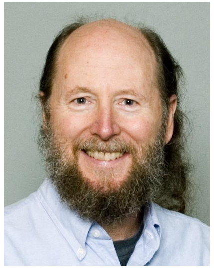
- Even outside of AI, it's very normal for methods to become quickly outdated.



"... general methods that leverage computation are ultimately the most effective, and by a large margin."

Adding sweetener 🍬 to the bitter ☕ lesson:

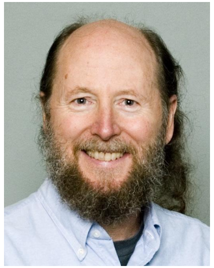
- Even outside of AI, it's very normal for methods to become quickly outdated.
- Our tools work better than ever before.



"... general methods that leverage computation are ultimately the most effective, and by a large margin."

Adding sweetener 🍬 to the bitter ☕ lesson:

- Even outside of AI, it's very normal for methods to become quickly outdated.
- Our tools work better than ever before.
- "Most effective" ➡ who gets to define this? how do you define this?



"... general methods that leverage computation are ultimately the most effective, and by a large margin."

Adding sweetener 🍬 to the bitter ☕ lesson:

- Even outside of AI, it's very normal for methods to become quickly outdated.
- Our tools work better than ever before.
- "Most effective" ➜ who gets to define this? how do you define this?
- What an "AI researcher" is is in flux --- opportunities to shape the field abound!

</aside>

MERLOT:

Multimodal Neural Script Knowledge Models

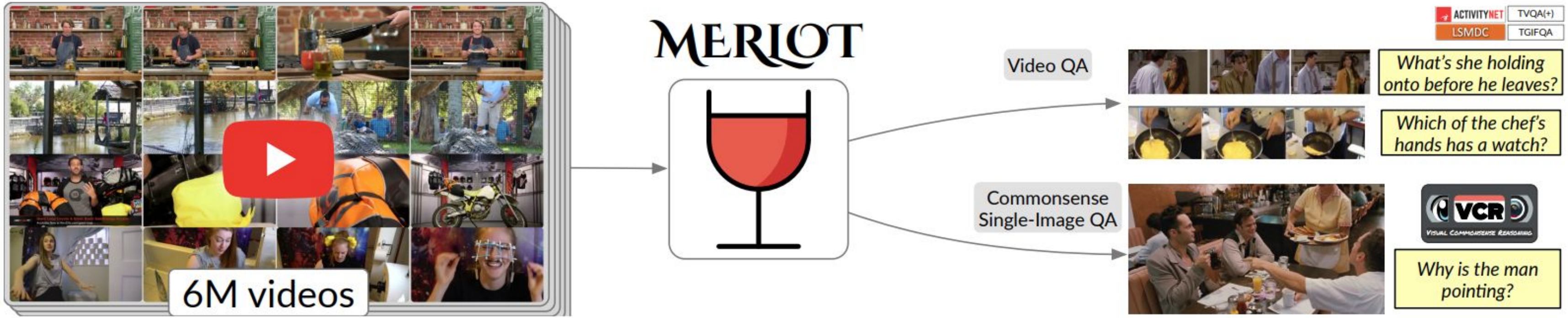
Rowan Zellers*, Ximing Lu*, Jack Hessel*, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi
NeurIPS 2021

MERLOT Reserve:

Neural Script Knowledge through Sound, Language, and Vision

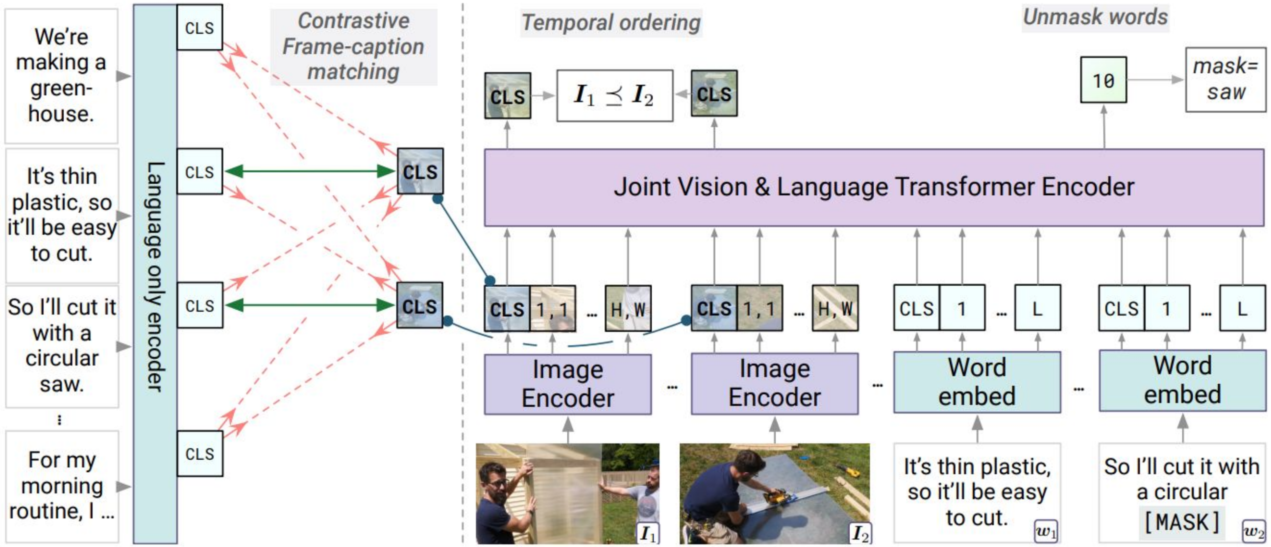
Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, Yejin Choi
CVPR 2022

Key idea:



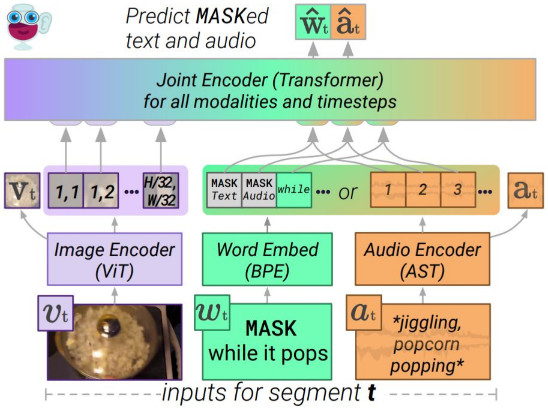
Learning from videos ➔ temporal understanding

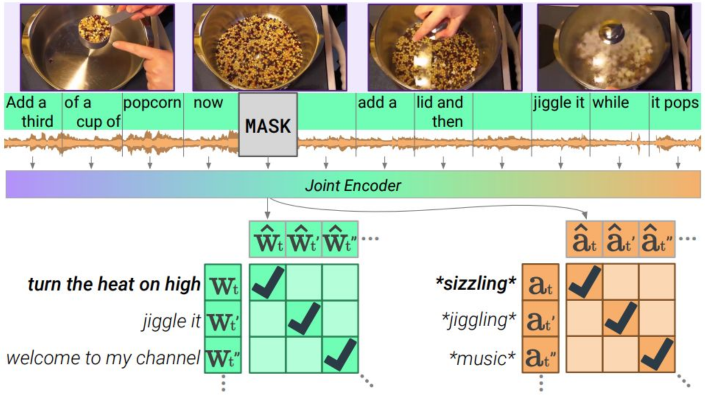
MERLOT



We're making a green-house.
It's thin plastic, so it'll be easy to cut.
So I'll cut it with a circular saw.
For my morning routine, I ...
Language only encoder
CLS
Contrastive Frame-caption matching
Temporal ordering
$I_1 \preceq I_2$
Unmask words
10
mask= saw
Joint Vision & Language Transformer Encoder
CLS 1,1 ... H,W
CLS 1 ... L
Image Encoder
Word embed
I_1
I_2
It's thin plastic, so it'll be easy to cut.
w_1
So I'll cut it with a circular [MASK]
w_2

MERLOT-RESERVE





improvements:

- Added in the audio modality!
- More data! 6M videos ➔ 20M videos!
- More compute! 400M "base" model ➔ 700M "large" model
- More training! for 2 weeks on roughly 512 ***T***PUs :-)

MERLOT + MERLOT Reserve work great!

[Zellers et al. 2019]

	Model	VCR test (acc; %) Q→A	QA→R	Q→AR
Caption/ObjDet-based	ERNIE-ViL-Large [124]	79.2	83.5	66.3
	Villa-Large [39]	78.9	83.8	65.7
	UNITER-Large [21]	77.3	80.8	62.8
	Villa-Base [39]	76.4	79.1	60.6
	ViLBERT [81]	73.3	74.6	54.8
	B2T2 [4]	72.6	75.7	55.0
	VisualBERT [77]	71.6	73.2	52.4
Video-based	MERLOT [128]	80.6	80.4	65.1
	RESERVE-B	79.3	78.7	62.6
	RESERVE-L	**84.0**	**84.9**	**72.0**

Table 2: RESERVE gets **state-of-the-art leaderboard performance on VCR**. We compare it with the largest submitted single models, including image-caption models that utilize heavy manual supervision (e.g. object detections and captions).

But: they don't have /exactly/ the same "magical" generalization "feeling" of the best text-only models out there...

GPT-3 Demo!

(content warning: GPT-3 outputs unfiltered and unrestricted free-text. While it usually doesn't, it can and has output offensive and/or graphic content.)

Hot off the press from DeepMind
(April 28, 2022)

Flamingo:

a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac[*,‡], Jeff Donahue[*], Pauline Luc[*], Antoine Miech[*], Iain Barr[†], Yana Hasson[†], Karel Lenc[†], Arthur Mensch[†], Katie Millican[†], Malcolm Reynolds[†], Roman Ring[†], Eliza Rutherford[†], Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan[*]

Key idea:

instead of training an entirely new generator...

can we just let a large language model "see"?

Modeling details

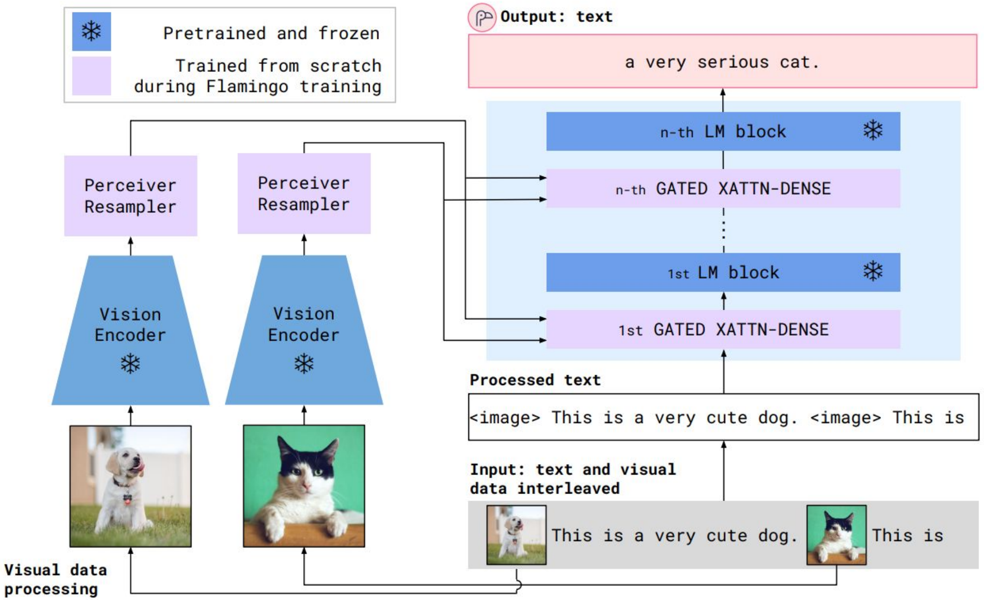


Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

Datasets

Figure 7 | **Training datasets.** Mixture of training datasets of different nature. N corresponds to the number of visual inputs for a single example. For paired image (or video) and text datasets, $N = 1$. T is the number of video frames with $T = 1$ being the special case of images. H, W, C are height, width and color channels.

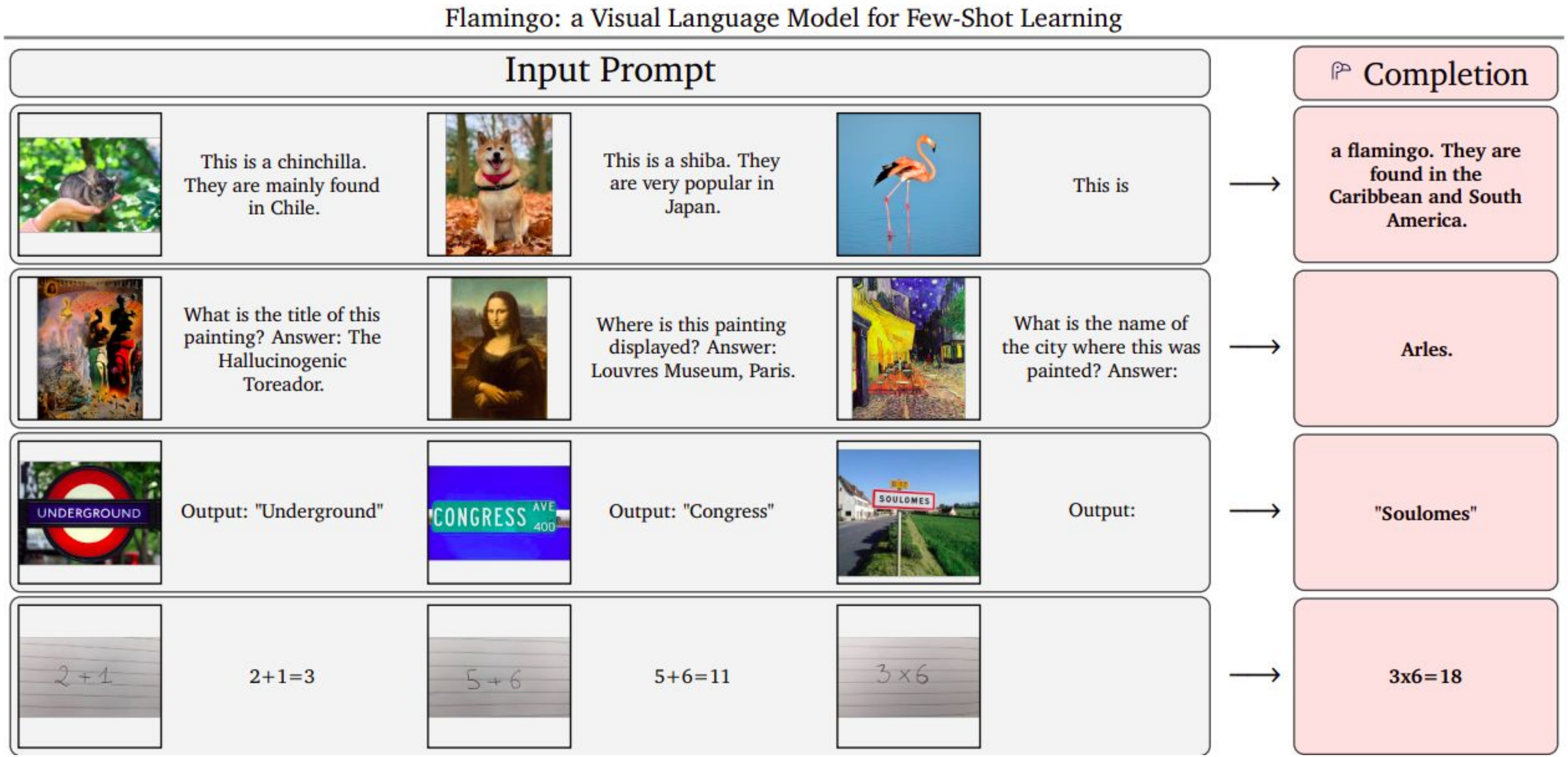
A window into some of the engineering required…

	Requires model sharding	Frozen Language	Frozen Vision	Trainable GATED XATTN-DENSE	Trainable Resampler	Total count
Flamingo-3B	✗	1.4B	435M	1.2B (every)	194M	**3.2B**
Flamingo-9B	✗	7.1B	435M	1.6B (every 4th)	194M	**9.3B**
Flamingo	✓	70B	435M	10B (every 7th)	194M	**80B**

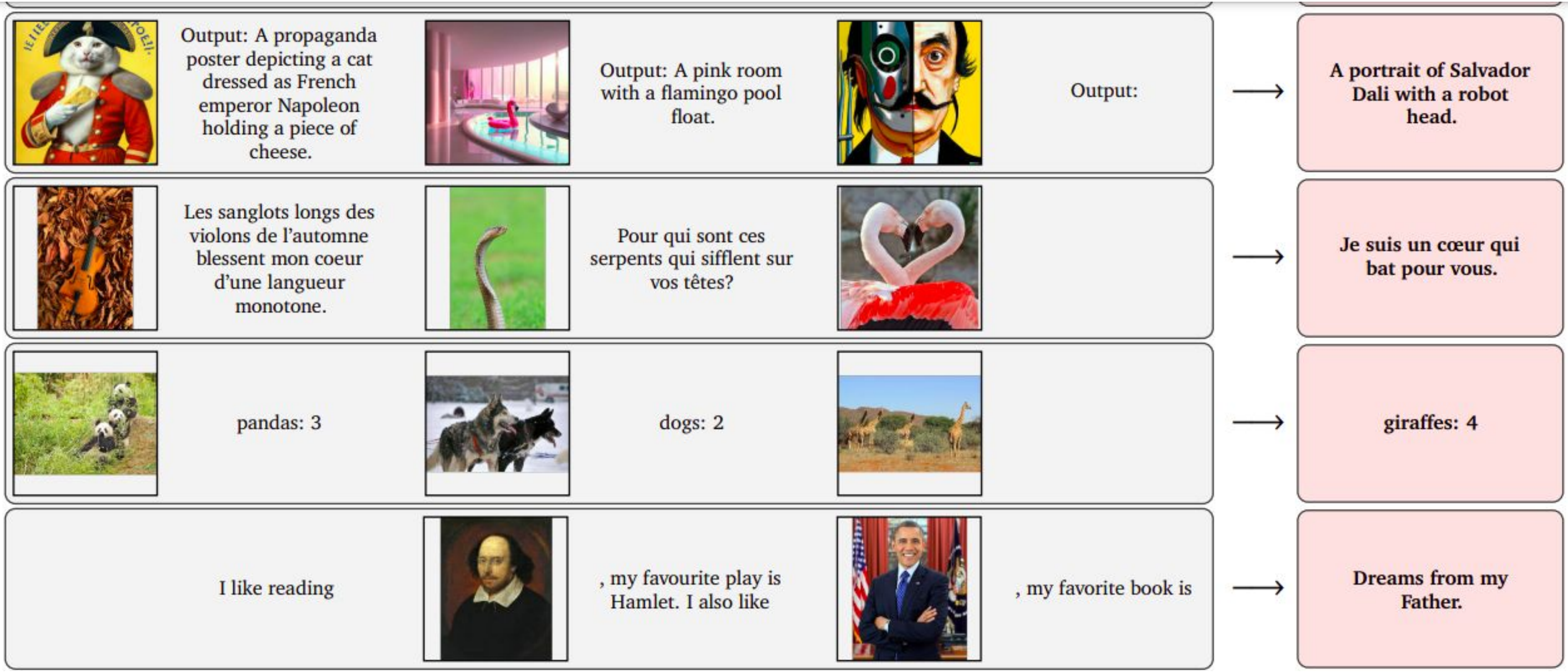
Table 1 | **Parameter counts for Flamingo models.** We focus on increasing the parameter count of the frozen LM and the trainable vision-text GATED XATTN-DENSE modules while maintaining the frozen vision encoder and trainable Resampler to a fixed and small size across the different models. The frequency of the GATED XATTN-DENSE with respect to the original language model blocks is given in parenthesis.

- 80B parameters = 320GB, just for the weights!!
- CPU vs. GPU vs. TPU (trained on 1536 TPU v4 chips for 15 days)
- floating point types + numerical stability + training dynamics…
- learning rates, pre-pretraining, architecture search, tweaks, tweaks, tweaks…

Qualitative results



Qualitative results



Quantitative results

Flamingo: a Visual Language Model for Few-Shot Learning

Method	FT	Shot	OKVQA	VQAv2	COCO	MSVDQA	VATEX	VizWiz	Flick30K	MSRVTTQA	iVQA	YouCook2	STAR	VisDial	TextVQA	NextQA	HatefulMemes	RareAct
Zero/Few shot SOTA	✗	(X)	[39] 43.3 (16)	[124] 38.2 (4)	[134] 32.2 (0)	[64] 35.2 (0)	-	-	-	[64] 19.2 (0)	[145] 12.2 (0)	-	[153] 39.4 (0)	[87] 11.6 (0)	-	-	[94] 66.1 (0)	[94] 40.7 (0)
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	55.6	36.5	30.8	68.6	-
	✗	8	57.5	65.6	108.8	45.5	60.6	44.8	78.2	27.6	44.8	80.7	42.3	56.4	37.3	32.3	**70.0**	-
	✗	16	57.8	66.8	110.5	48.4	62.8	48.4	**78.9**	30.0	45.2	84.2	41.1	**56.8**	37.6	32.9	**70.0**	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	75.4	**31.0**	**45.3**	**86.8**	42.2	OOC	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [39] (10K)	80.2 [150] (444K)	143.3 [134] (500K)	47.9 [32] (27K)	76.3 [165] (500K)	57.2 [70] (20K)	67.4 [162] (30K)	46.8 [57] (130K)	35.4 [145] (6K)	138.7 [142] (10K)	36.7 [138] (46K)	75.2 [87] (123K)	54.7 [147] (20K)	25.2 [139] (38K)	75.4 [60] (9K)	-

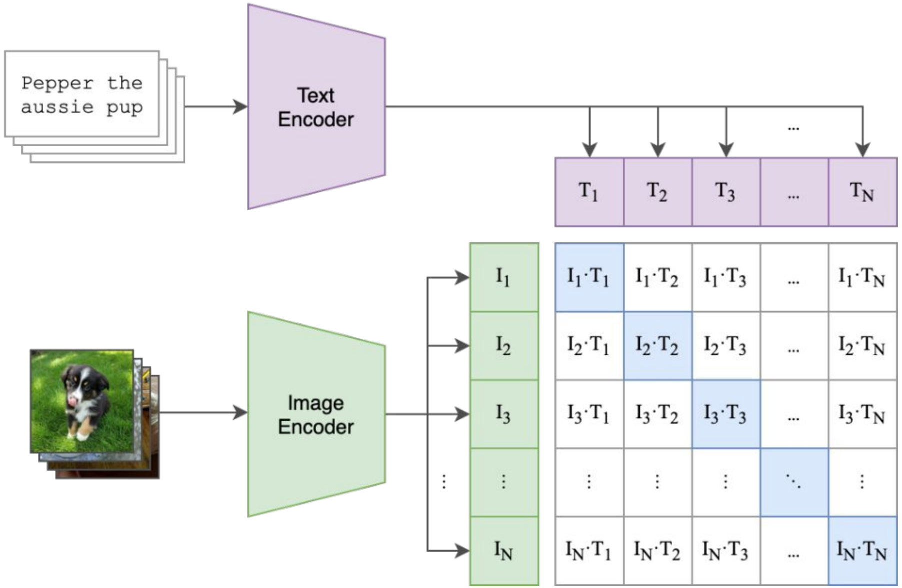
Flamingo Demo!

Sadly, there isn't a public one

other cool multimodal models

CLIP

400M Image+Caption pairs from the web

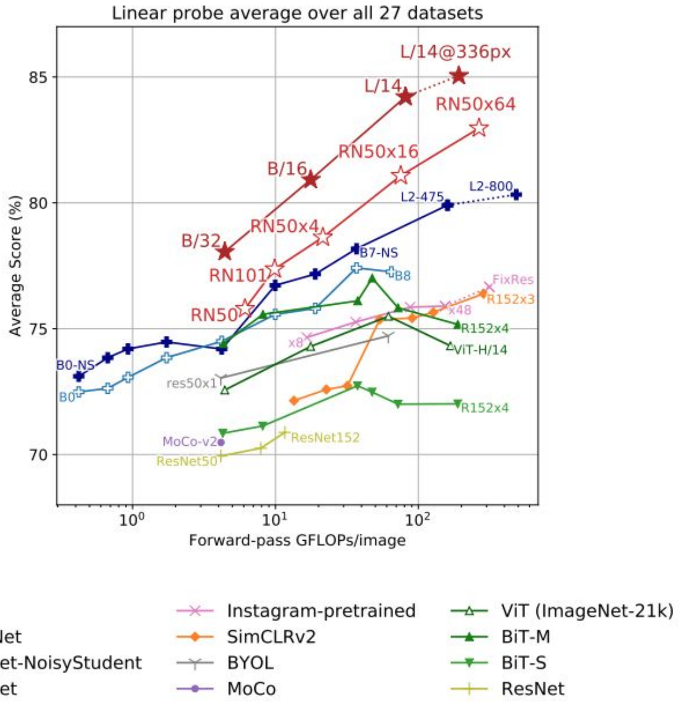


600M-ish parameters

[Radford et al. 2021 https://github.com/openai/CLIP]

CLIP

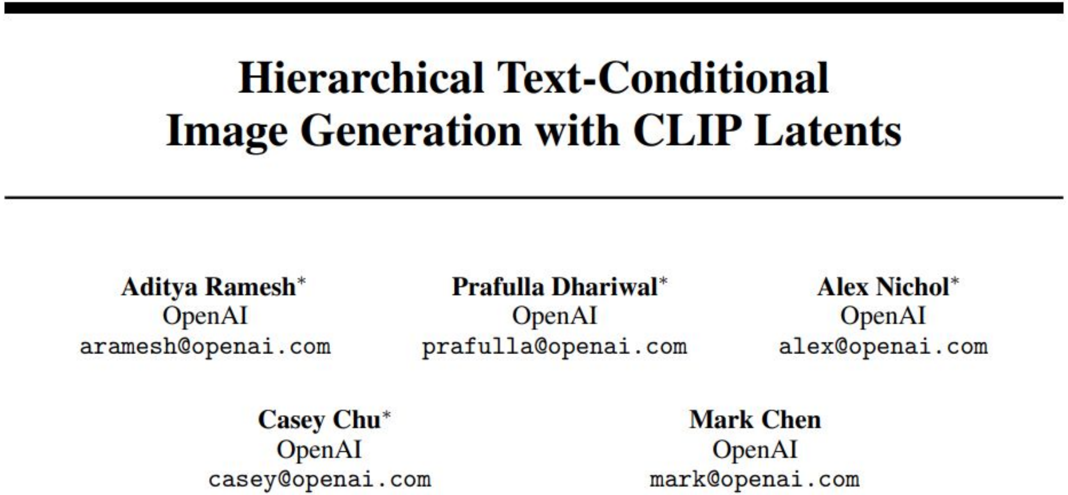
Recognition-style tasks, e.g., ImageNet, food classification, etc.



[Radford et al. 2021
https://github.com/openai/CLIP]

DALL-E 2

"The University of Washington quad with cherry blossoms under the stars, mixed media, 8K trending on artstation"



Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh*
OpenAI
aramesh@openai.com

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Casey Chu*
OpenAI
casey@openai.com

Mark Chen
OpenAI
mark@openai.com

some musings

discussion welcomed!

Bigger models genuinely generalize better: benchmark progress is real!

Modeling papers will continue to look more like **systems papers**

Many, many **new applications** out there.

The purview of "NLP research" is broadening **--- lots of room for creativity!**

There are **/lots/ of ethics and privacy concerns** with training and deploying models

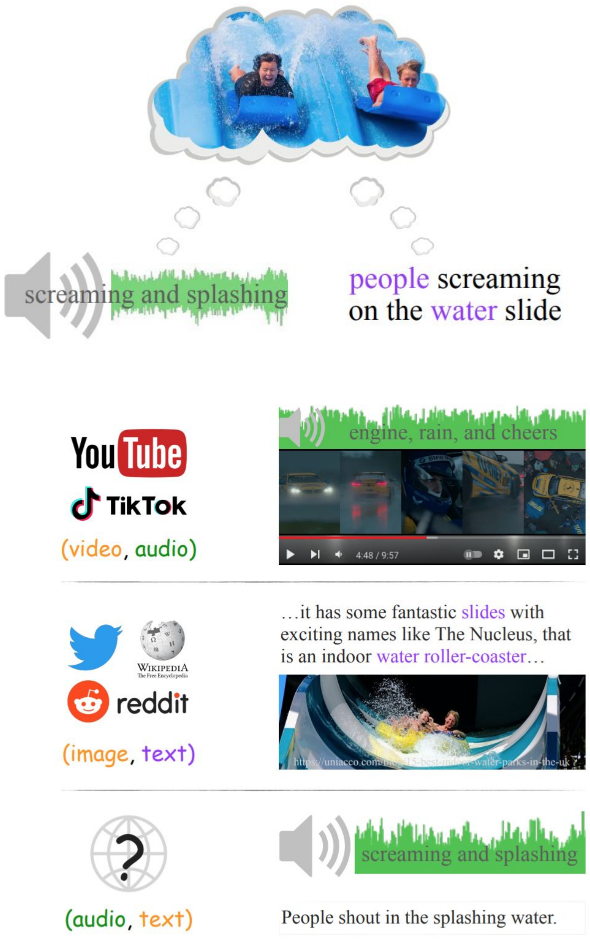
More multimodal work at AI2!

Connecting the Dots between Audio and Text without Parallel Data through Visual Knowledge Transfer

Yanpeng Zhao[♣*] Jack Hessel[♡] Youngjae Yu[♡]
Ximing Lu[♠♡] Rowan Zellers[♠] Yejin Choi[♠♡]
[♣]Institute for Language, Cognition and Computation, University of Edinburgh
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence

Zero resource prediction results:

Model	ESC50	US8K	AS
Supervised	$95.7_{\pm1.4}$	$86.0_{\pm2.8}$	37.9
Wav2CLIP	41.4	40.4	
VIP~ANT++	62.8(55.7)	54.0(47.0)	11.6(12.3)

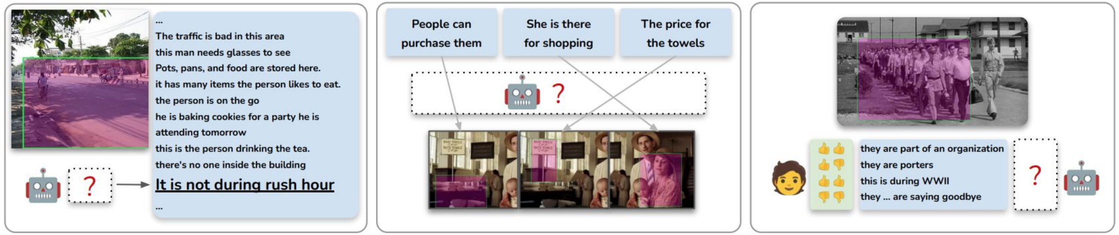


363K (clue, inference) pairs over 103K images!

The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning

Jack Hessel*♣ Jena D. Hwang*♣ Jae Sung Park♡ Rowan Zellers♡
Chandra Bhagavatula♣ Anna Rohrbach♢ Kate Saenko♠ Yejin Choi♣♡

♣Allen Institute for Artificial Intelligence
♡Paul G. Allen School of Computer Science & Engineering, University of Washington
♢University of California, Berkeley ♠ Boston University and MIT-IBM Watson AI



(a) Retrieval of abductive inferences (b) Localization of evidence (c) Comparison of plausibility

	Retrieval			*Localization*	*Comparison*
	im → txt (↓)	txt → im (↓)	$P@1_{im \to txt}$ (↑)	GT-Box/Auto-Box (↑)	Val/Test Human Acc (↑)
CLIP `RN50x64`	19.3	19.7	31.8	86.6/39.5	25.1/26.0
↳ + multitask clue learning	**16.4**	**17.7**	**33.4**	**87.2/40.6**	**26.6/27.1**
Human + (Upper Bound)	-	-	-	92.3/(96.2)	42.3/42.3

Big thanks to my awesome collaborators, and to you for listening!!

"MERLOT on VCR"

Jack Hessel: Research Scientist, AI2. Code, models, papers on my website!

Feel free to reach out: jackh@allenai.org ; www.jmhessel.com; @jmhessel on twitter

If we have more extra time...

Option 1: AMA! Happy to take any questions about me, AI2, multimodal ML, web-scale models, ...

Option 2: We can play with GPT-3 more, and I can talk about a few use cases I've used the model for, in practice.

Option 3: I can talk about a few other projects I'm working on that aren't out yet!