From Visual Recognition To Visual Understanding

Laurens van der Maaten

Yapay Öğrenme Yaz Okulu 2020
Bilkent Üniversitesi

facebook
Artificial Intelligence Research

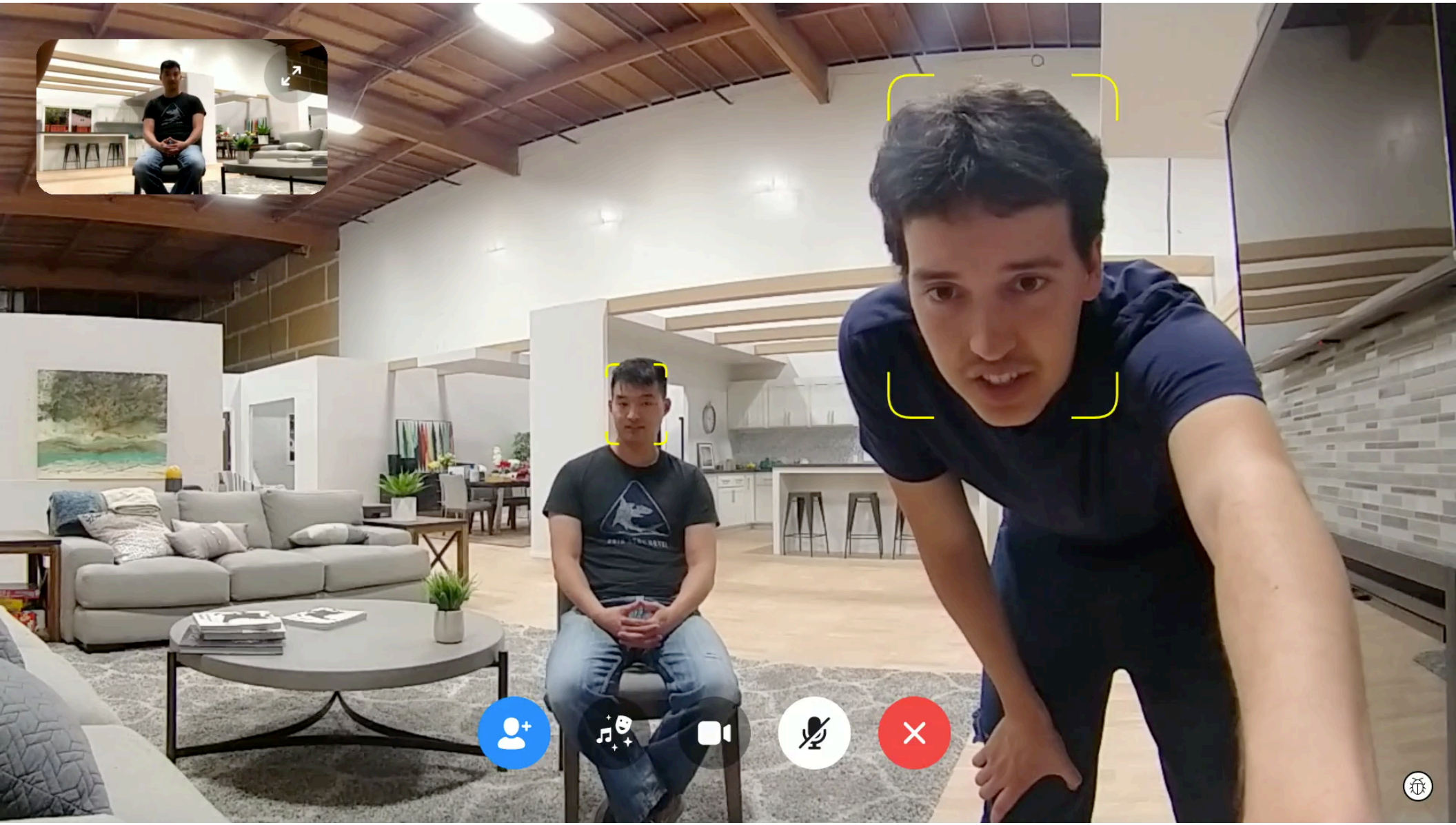
Mission

“Advancing the state-of-the-art in AI through open research for the benefit of all.”

Values

- Openness: Publish and open-source
- Freedom: Researchers have complete control on their agenda
- Collaboration: With internal and external partners
- Excellence: Focus on most impactful projects
- Scale: Operate at large scale

portal



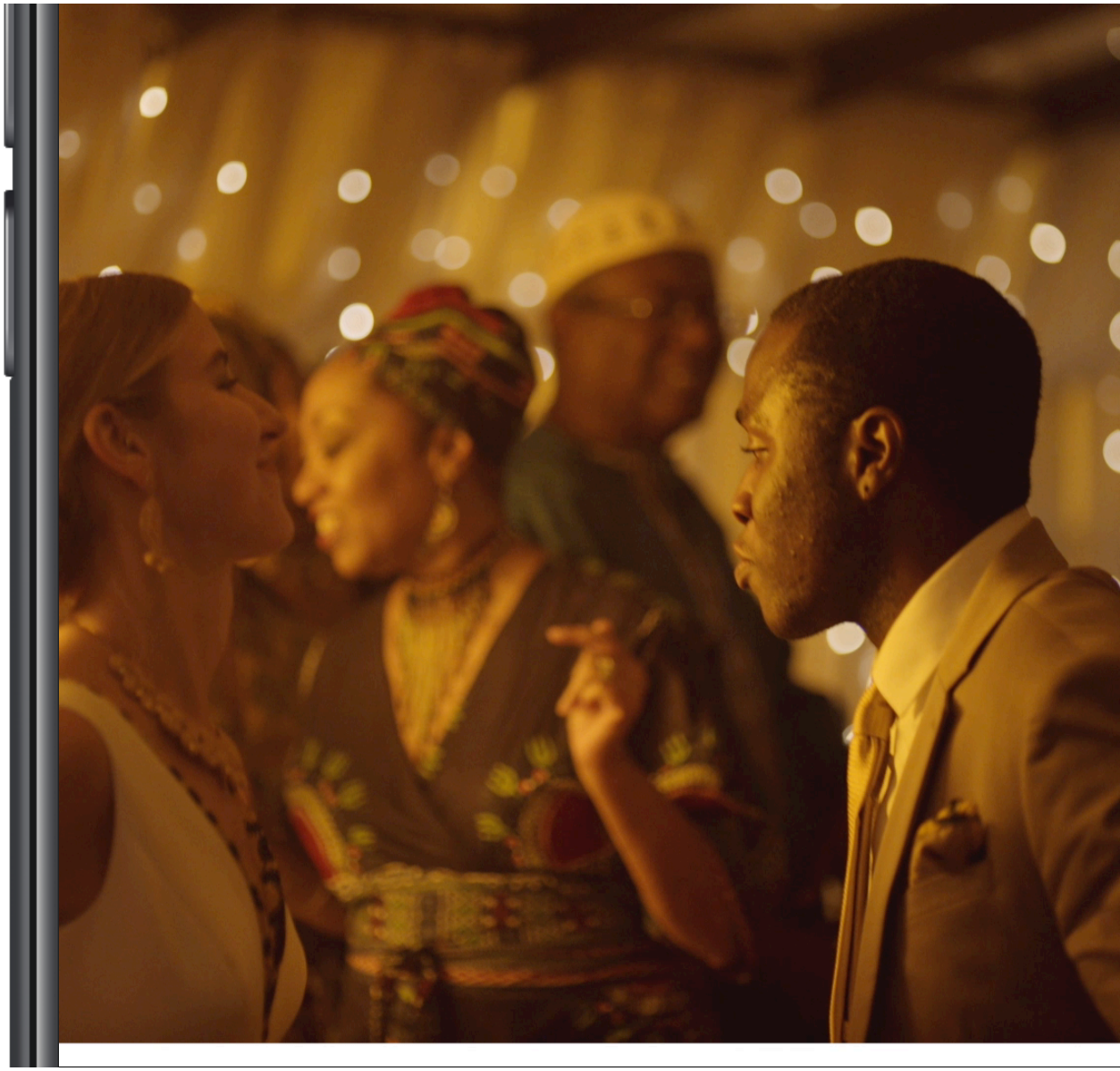
Translation

50%

of users have at least one friend
with a different native language

~6B

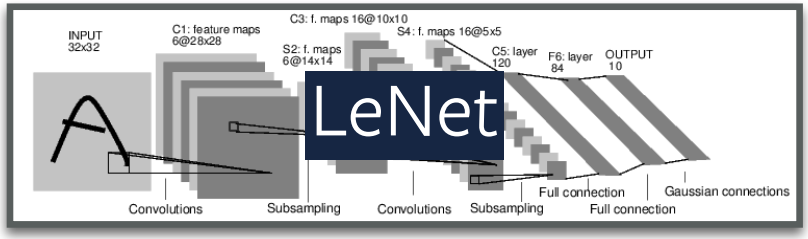
translated posts every day
on Facebook

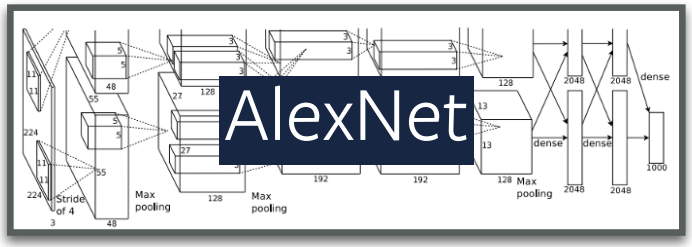


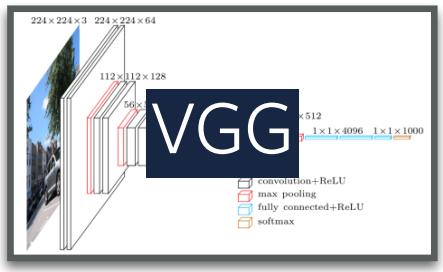
From Visual Recognition...

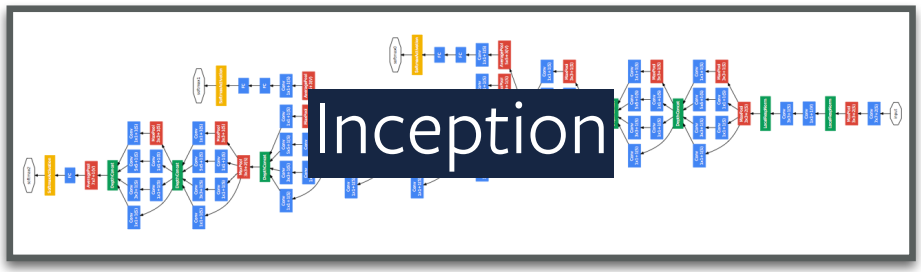
Convolutional Networks

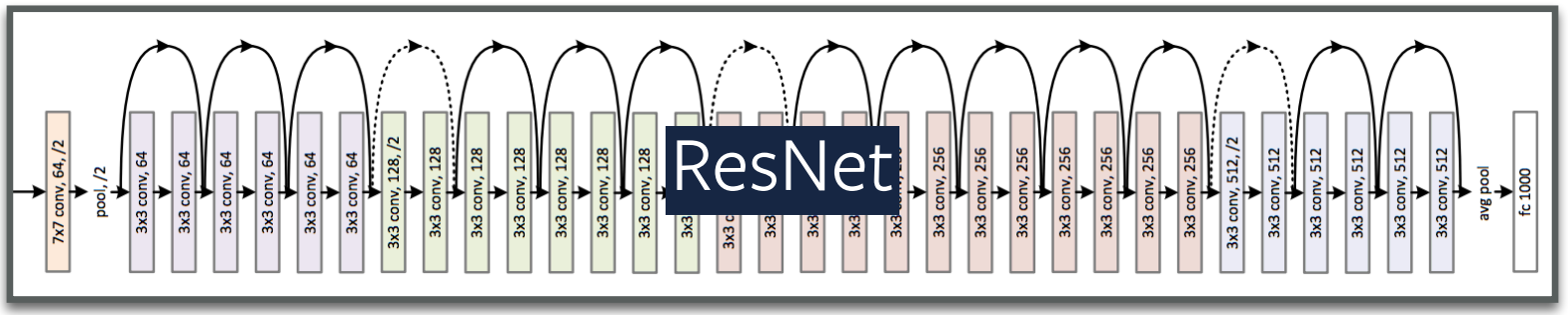
- Repeatedly filters image
- Filters are learned based on labeled training images
- Final representation is "semantic"





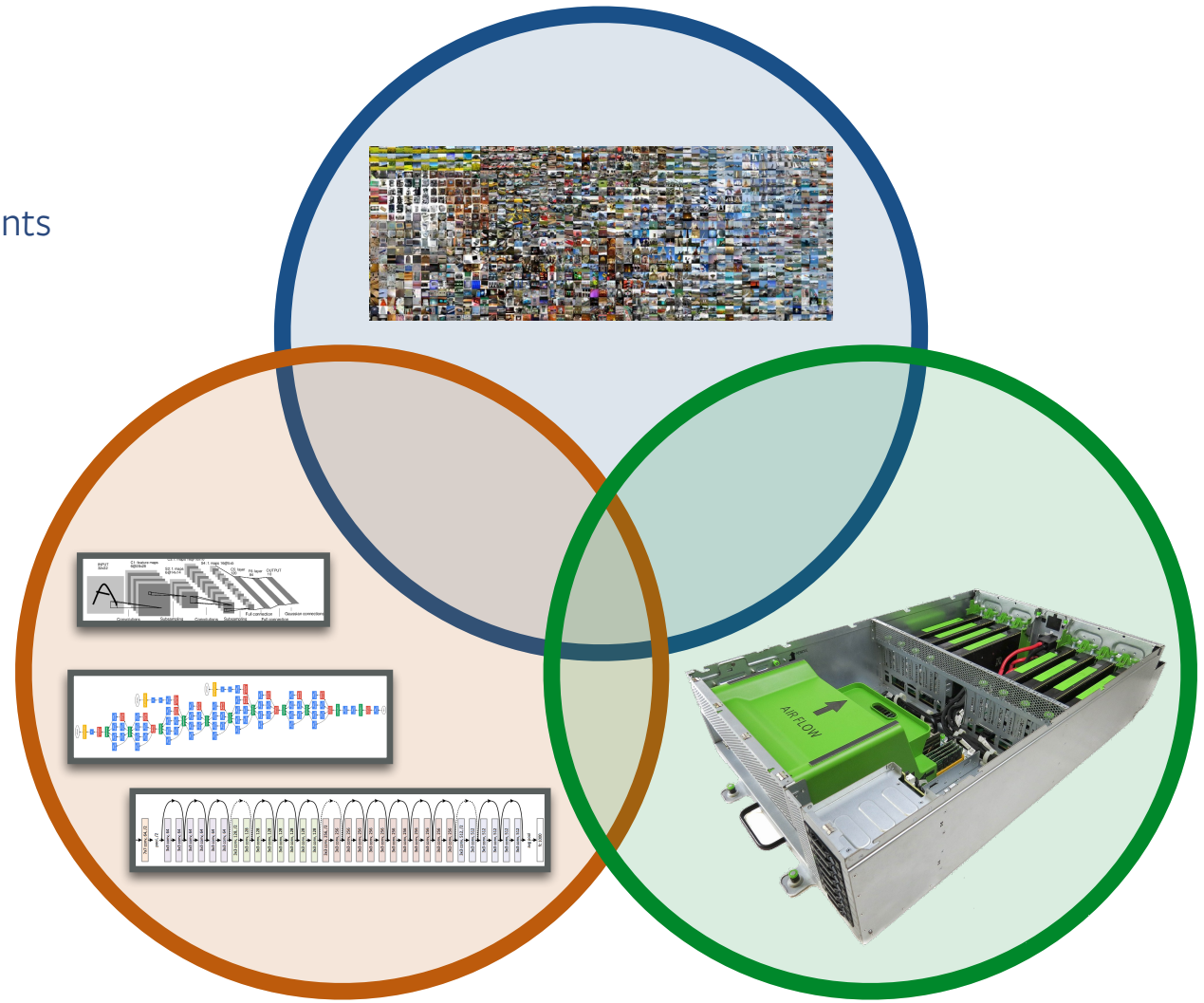






Ingredients

- Three main ingredients came together in recent years



Pretraining Vision Models

- First, train a model on a large "source" dataset (say, ImageNet)

Pretraining Vision Models

- First, train a model on a large "source" dataset (say, ImageNet)

- Finetune on a small "target" dataset

- Measure accuracy on target task

Research question

Can we use large amounts of weakly supervised images for pretraining?

Highlights

- We pretrain models by predicting relevant hashtags for images
- We pretrain models to predict 17.5K hashtags for 3.5B images
- After finetuning, we beat the state-of-the-art on, *e.g.*, ImageNet


Dhruv Mahajan


Ross Girshick


Vignesh Ramanathan


Kaiming He


Manohar Paluri


Yixuan Li

Ashwin Bharambe

facebook
Artificial Intelligence Research

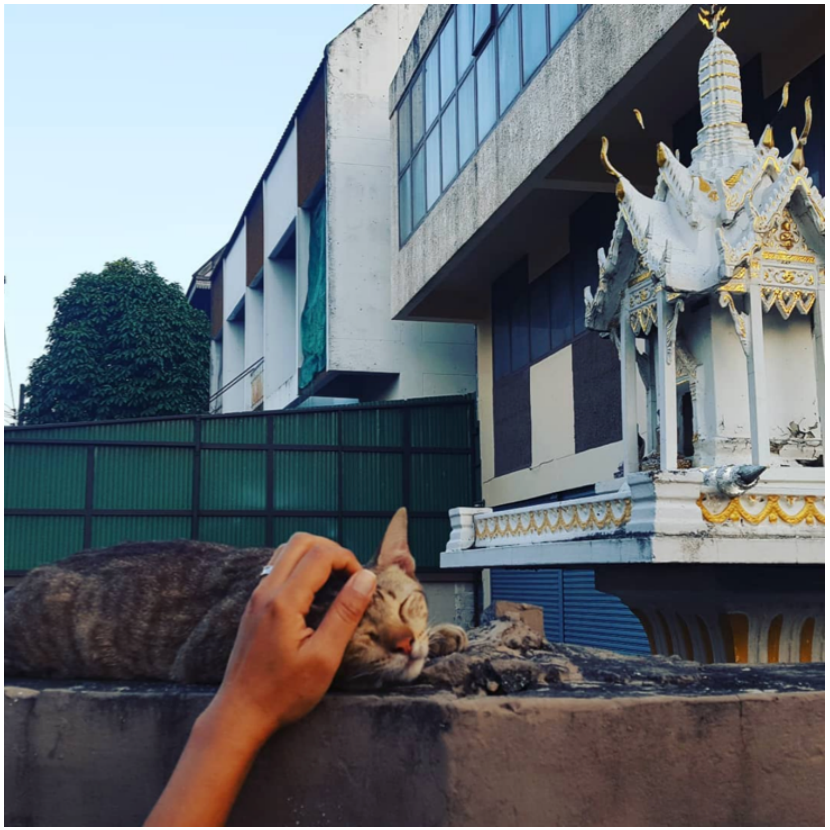
Hashtag Supervision

- It is easy to get billions of public images and hashtags
- Hashtags are more structured than captions
- Hashtags were often assigned to make images "searchable"

#cheesecake #birthday

Hashtag Supervision

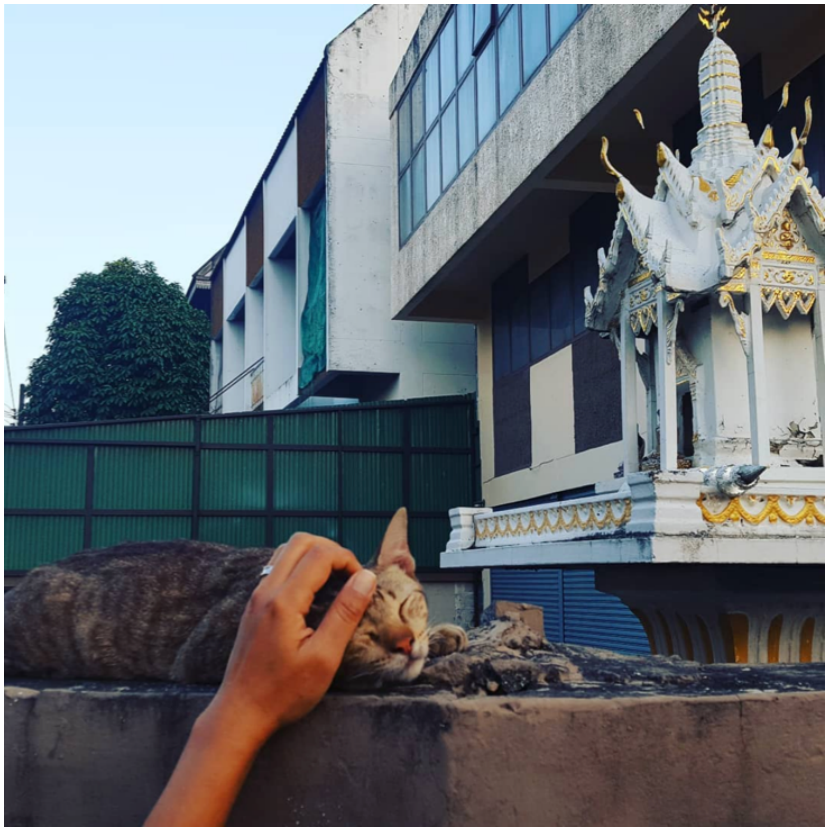
- But hashtags are not perfect supervision



#cat #travel #thailand #family

Hashtag Supervision

- But hashtags are not perfect supervision
- Some hashtags are not visually relevant



#cat #travel #thailand #family

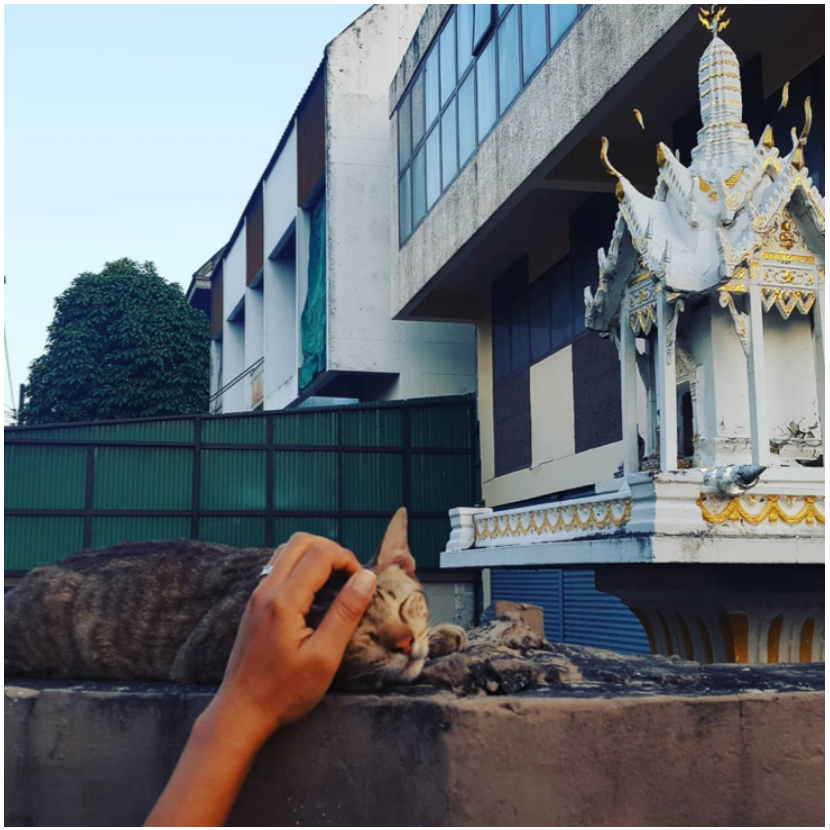
Hashtag Supervision

- But hashtags are not perfect supervision
- Some hashtags are not visually relevant
- Other hashtags are not in the photo

#cat #travel #thailand #family

Hashtag Supervision

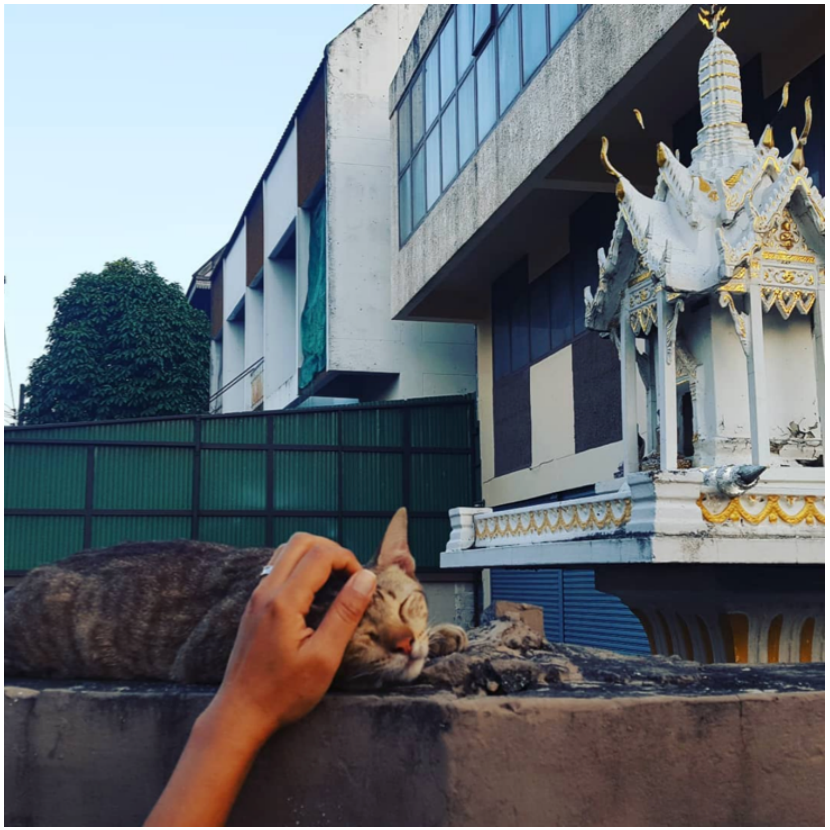
- But hashtags are not perfect supervision
- Some hashtags are not visually relevant
- Other hashtags are not in the photo
- And there are many false negatives



#cat #travel #thailand #family
#building #fence #...

Hashtag Supervision

- But hashtags are not perfect supervision
- Some hashtags are not visually relevant
- Other hashtags are not in the photo
- And there are many false negatives
- Is scaling up sufficient to make up for this noise?



#cat #travel #thailand #family
#building #fence #...

Experiments

- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge `#brownbear` and `#ursusarctos`)

Experiments

- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge #brownbear and #ursusarctos)
- The final list has 17,517 hashtags

1	aar	44	accommodation	17474	yurt
2	aardvark	45	accompaniment	17475	zabaglione
3	aardwolf	46	accordion	17476	zambeziriver
4	aba	47	accoutrement	17477	zamboni
5	abaca	48	accumulator	17478	zamia
6	abacus	49	ace	17479	zantac
7	abalone	50	aceofclubs	17480	zantedeschia
8	abatis	51	aceofdiamonds	17481	zap
9	abaya	52	aceofhearts	17482	zapper
10	abbey	53	aceofspades	17483	zarf
11	abele	54	acer	17484	zea
12	abelia	55	acerjaponicum	17485	zebra
13	abies	56	acerola	17486	zebrafinch
14	abila	57	acerpalmatum	17487	zebrawood
15	abm	58	acerrubrum	17488	zebu
16	abortus	59	acetaminophen	17489	zero
17	abronia	60	acetate	17490	zeus
18	absinth	61	acheron	17491	zhujiang
19	absinthe	62	acherontia	17492	ziggurat
20	abstraction	63	acherontiaatropos	17493	zill
21	abstractionism	64	achillea	17494	zimmerframe
22	abutilon	65	achilleamillefolium	17495	zinfandel
23	abutment	66	achimenes	17496	zing
24	abyss	67	acid	17497	zingiber
25	abyssinian	68	acidophilus	17498	zinnia
26	acacia	69	acinonyxjubatus	17499	zipgun
27	acaciadealbata	70	acinus	17500	zipper
28	academy	71	ackee	17501	zither
29	acalypha	72	aconcagua	17502	ziti
30	acanthaceae	73	aconite	17503	ziziphus
31	acanthurus	74	aconitum	17504	zizz
32	acanthus	75	acorn	17505	zodiac
33	acanthusmollis	76	acornsquash	17506	zoloft
34	acapulcogold	77	acousticguitar	17507	zombi
35	acarus	78	acoustics	17508	zoologicalgarden
36	accelerator	79	acrididae	17509	zoom
37	accelerometer	80	acrobates	17510	zooplankton
38	access	81	acropolis	17511	zootsuit
39	accessory	82	acropora	17512	zori
40	accident	83	acrylic	17513	zoysia
41	accipiter	84	acrylicpaints	17514	zuiderzee
42	accipiternisus	85	actias	17515	zygnema
43	accipitridae	86	actiasluna	17516	zygocactus
				17517	zygoptera

...

Experiments

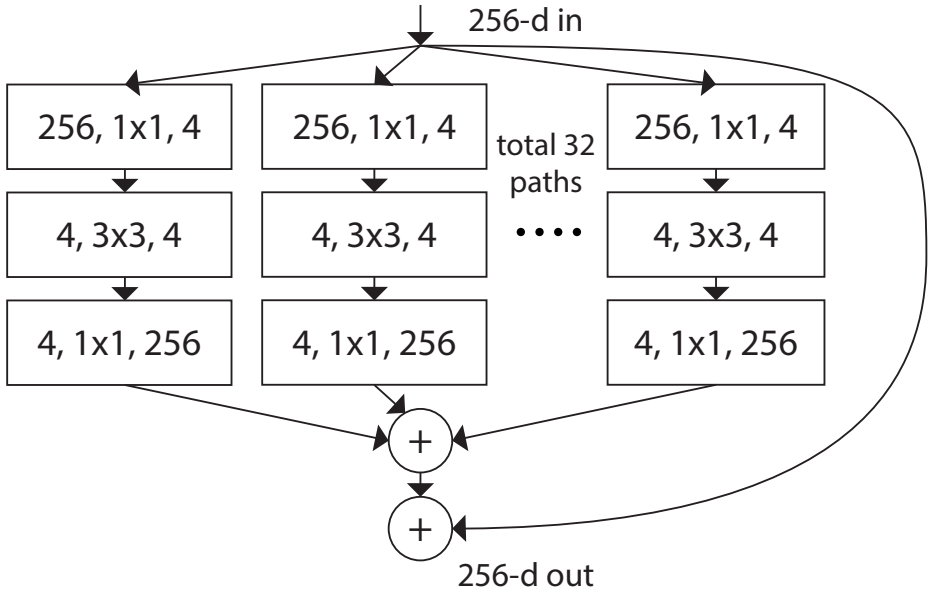
- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge `#brownbear` and `#ursusarctos`)
- The final image set has ~3.5B images

I cannot show you the images, but you can look...

https://www.instagram.com/explore/tags/brownbear/
https://www.instagram.com/explore/tags/crane/
https://www.instagram.com/explore/tags/...

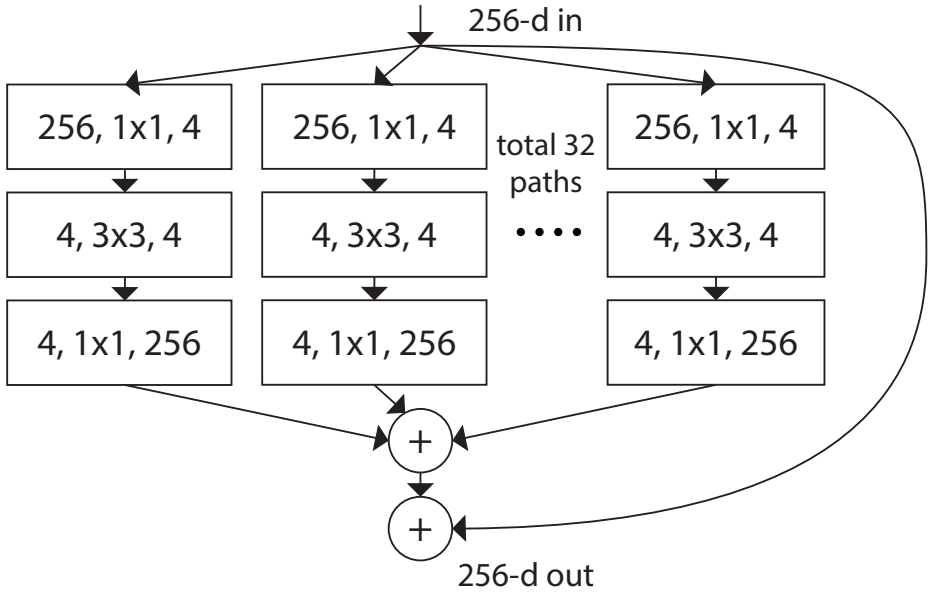
Experiments

- Train ResNeXt-32xCd convolutional networks
- Use *c*-of-*K* vector to represent multiple labels
- Train to minimize multi-class logistic loss



Experiments

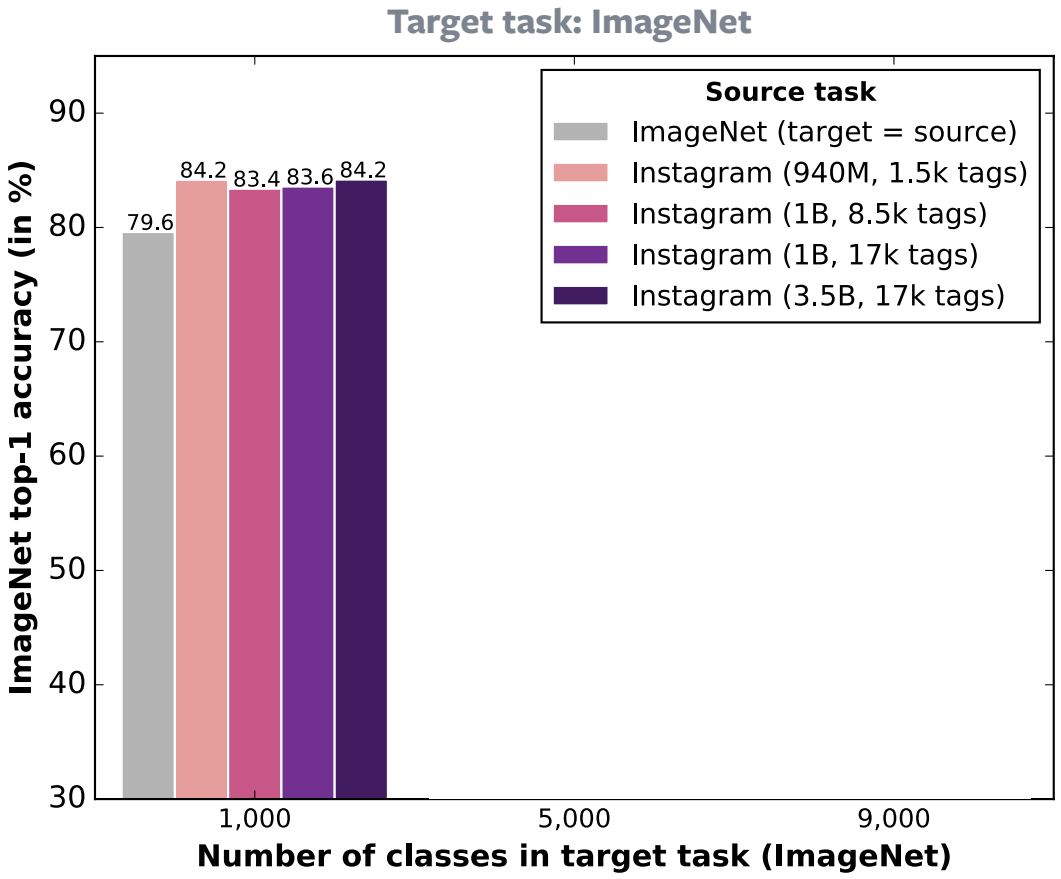
- Train ResNeXt-32xCd convolutional networks
- Use *c*-of-*K* vector to represent multiple labels
- Train to minimize multi-class logistic loss



most experiments use ResNeXt-101 32x16d

Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

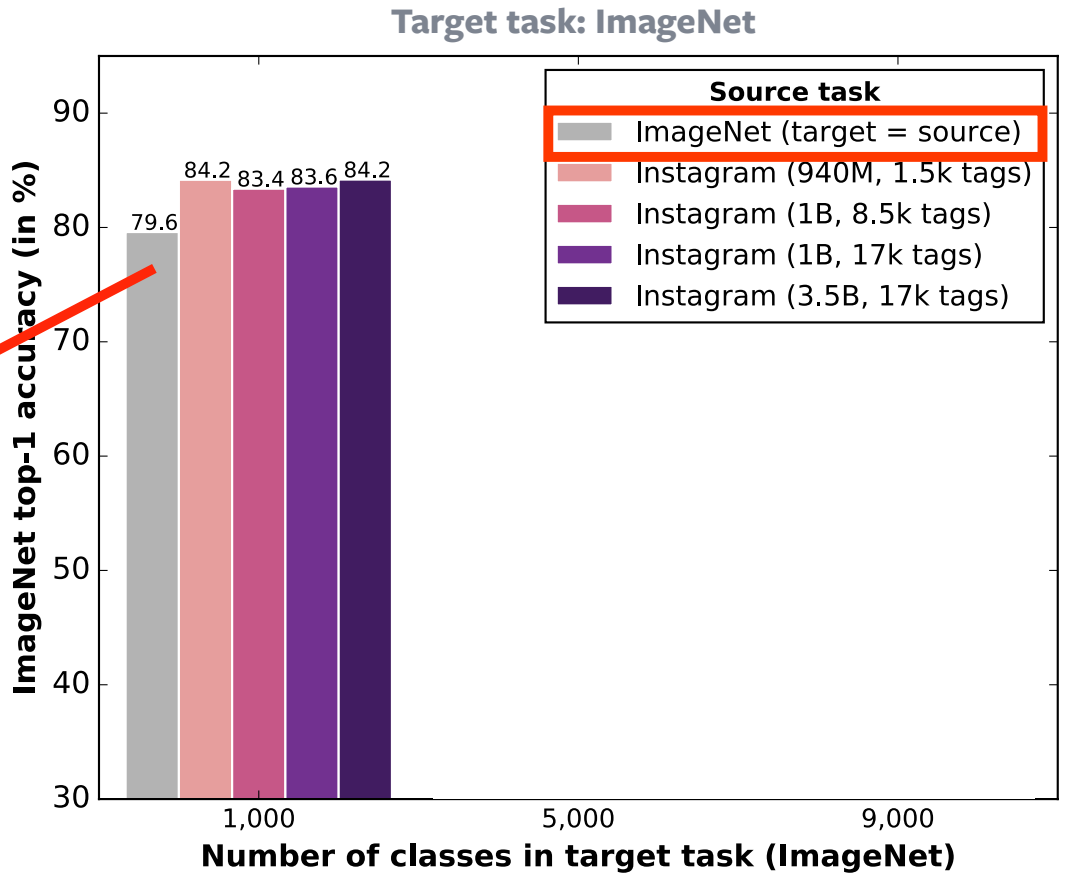


Models available in PyTorch Hub!

Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

"standard" ImageNet training

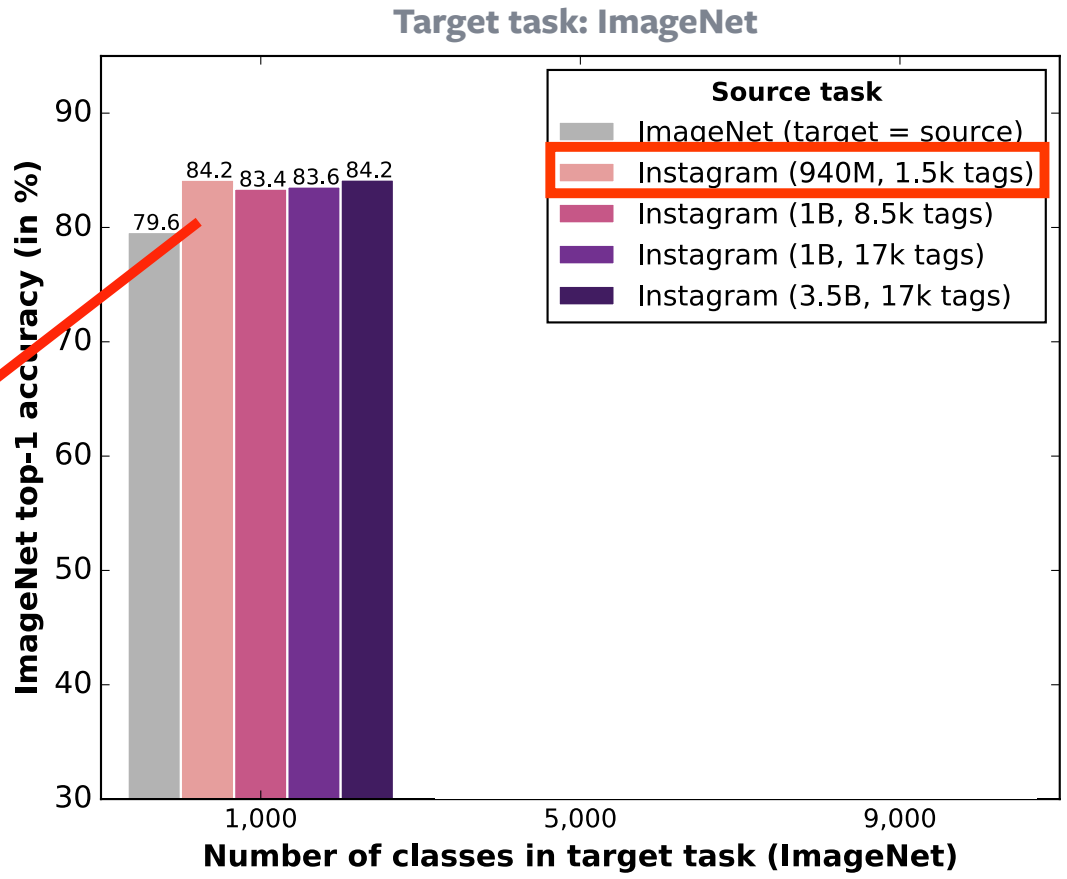


Models available in PyTorch Hub!

Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

pre-training on 1B Instagram images, selected to match ImageNet classes

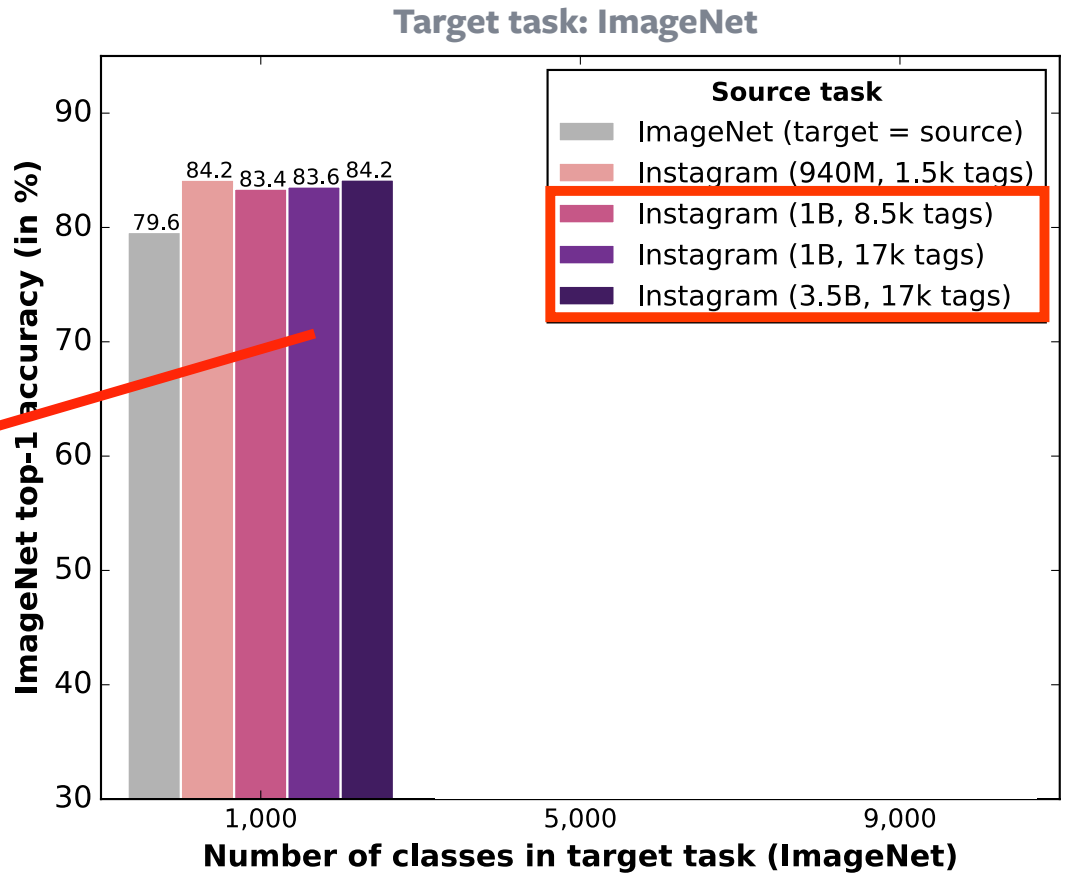


Models available in PyTorch Hub!

Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

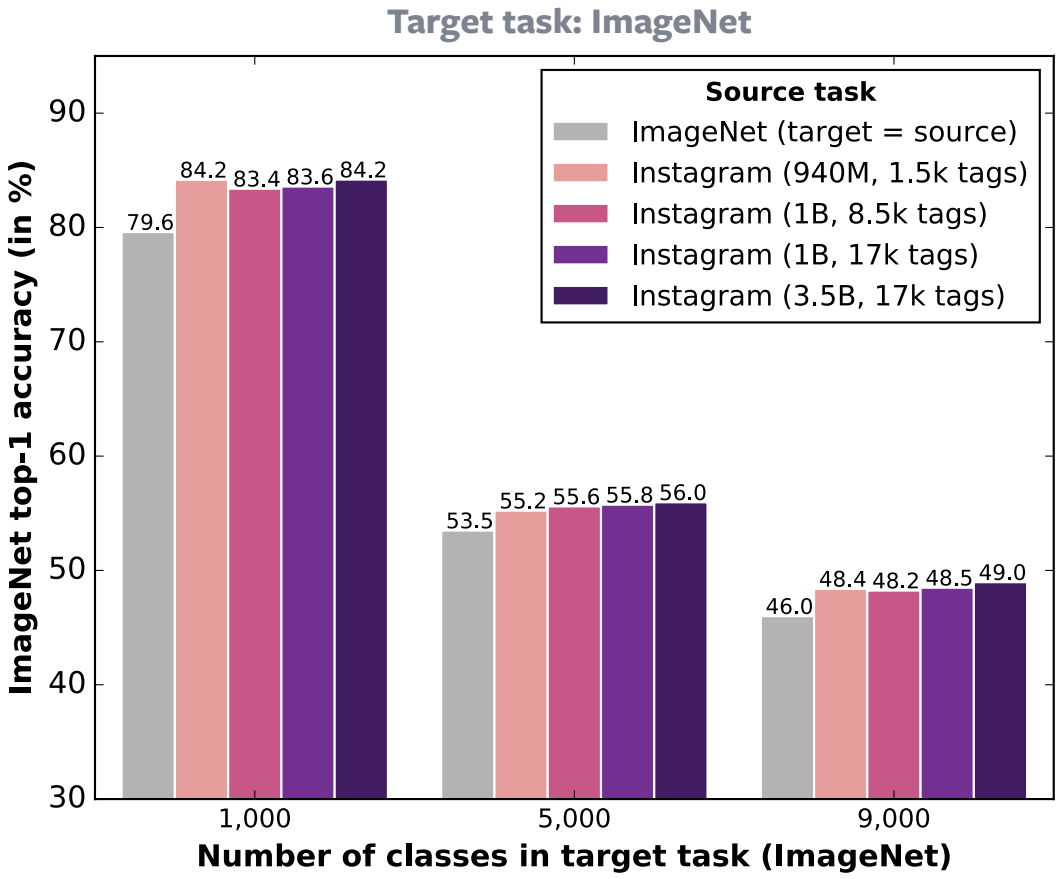
pretraining on 1-3.5B Instagram images, without selection



Models available in PyTorch Hub!

Fix Model; Vary Data

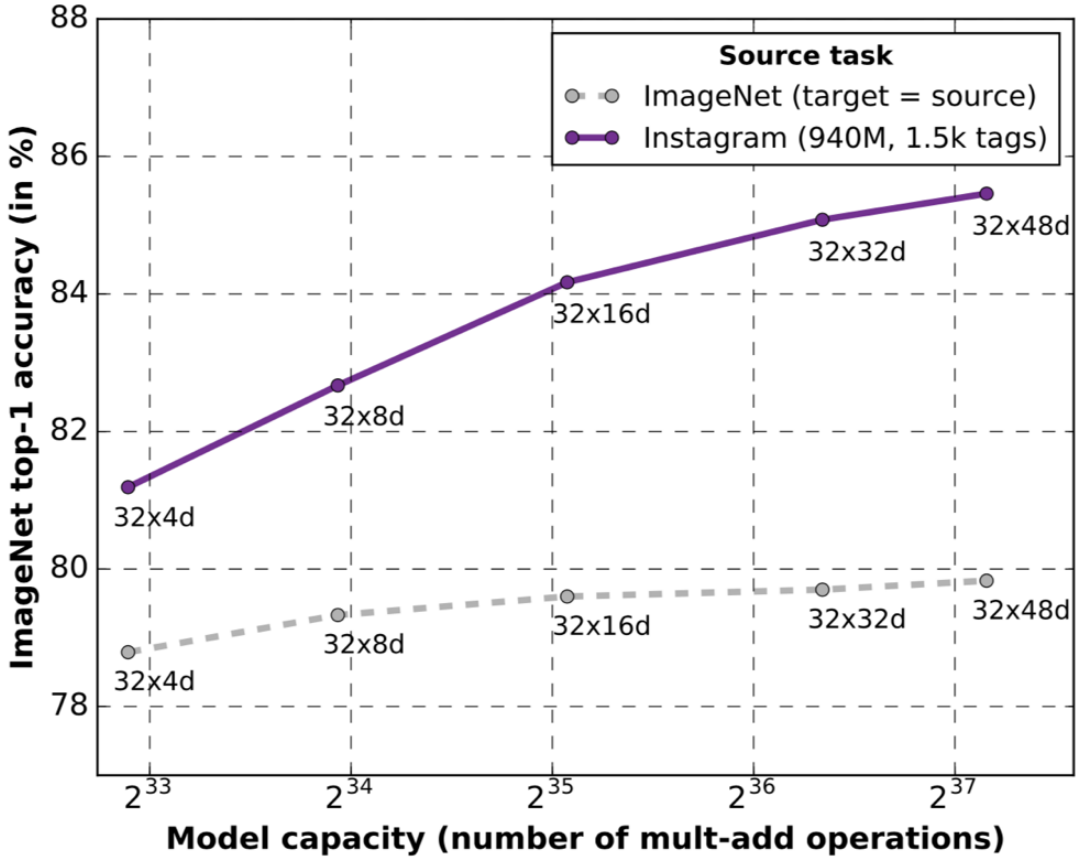
- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet
- Similar results on larger versions of ImageNet



Models available in PyTorch Hub!

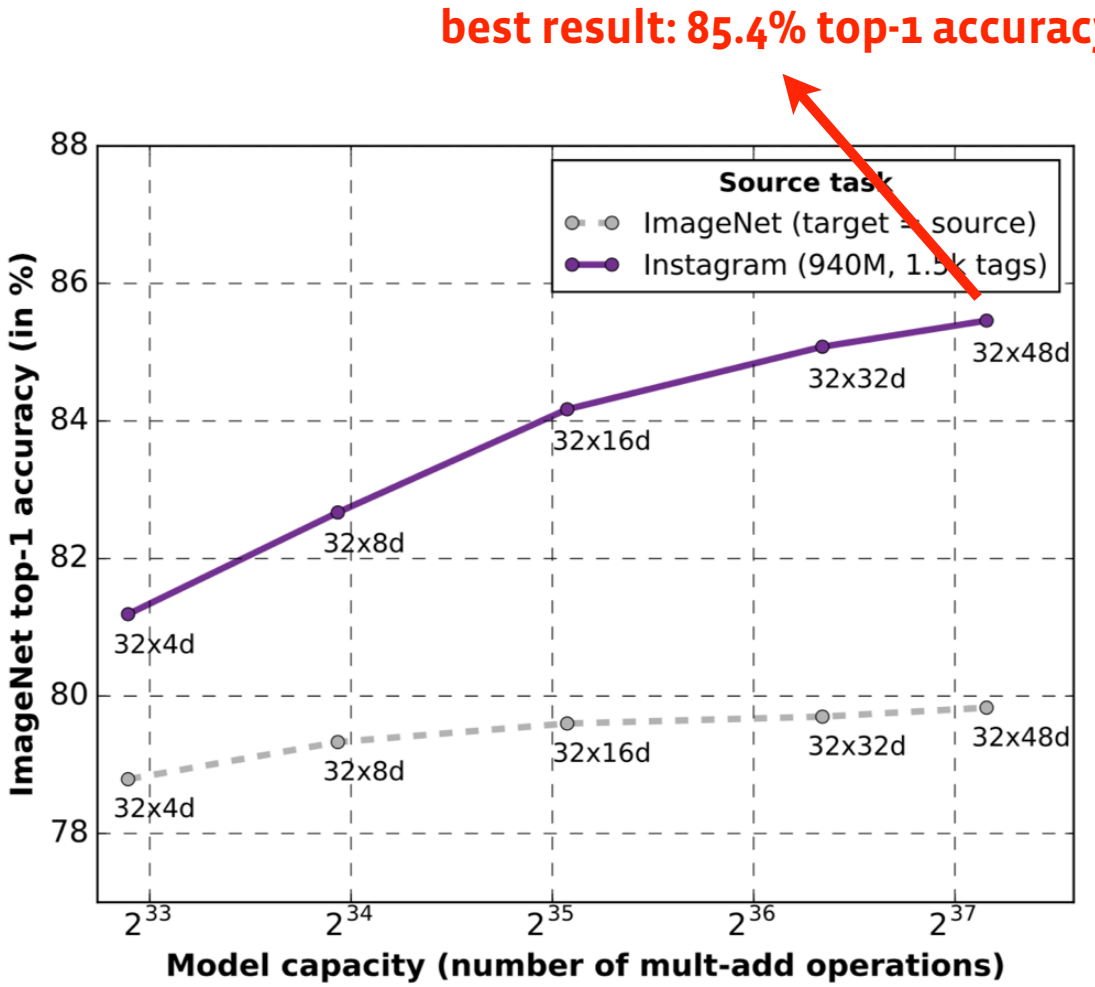
Fix Data; Vary Model

- Increasing model capacity has a larger positive effect
- Even lower error rates may be possible?



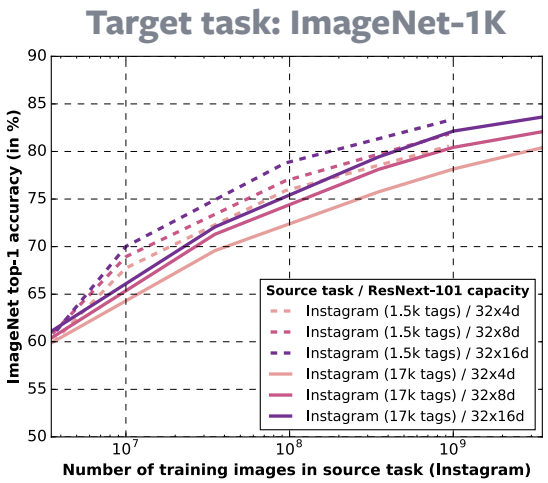
Fix Data; Vary Model

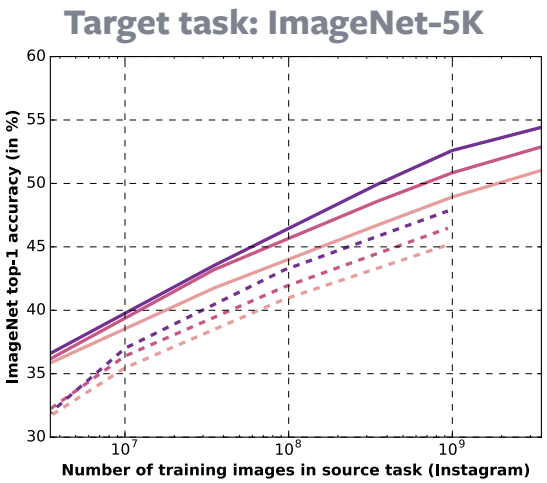
- Increasing model capacity has a larger positive effect
- Even lower error rates may be possible?

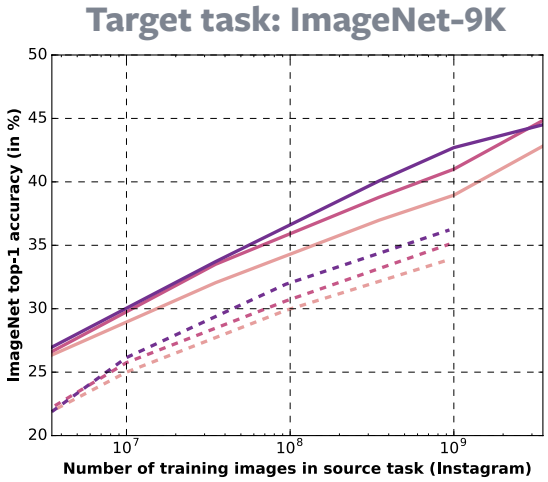


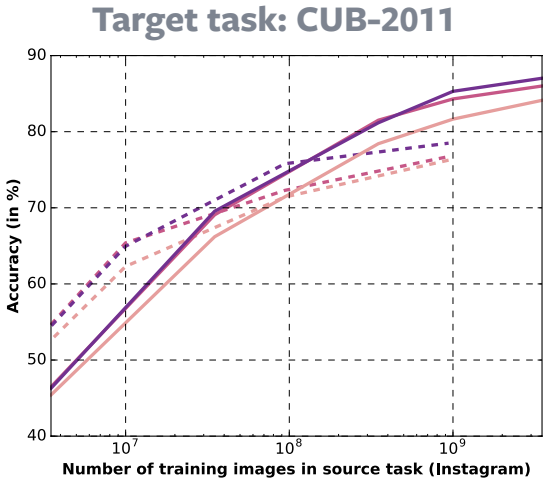
Learning Curves

- Accuracy on target task improves (almost) log-linearly with data size
- Matching hashtags to target task may help









Research question

Does image recognition work for everyone?

Highlights

- We test cloud services for image recognition on household items
- Images used for test come from across the world
- Services work better in some countries than others


Terrance DeVries


Changhan Wang


Ishan Misra

Image Classification

- Is image classification solved? Yes?

Soap

Country of Origin: UK
Prediction: Toiletry

Spices

Country of Origin: USA
Prediction: Spice

Toothpaste

Country of Origin: USA
Prediction: Toothpaste

Image Classification

- Is image classification solved? Yes? **No!**

Soap

Country of Origin: UK
Prediction: Toiletry

Spices

Country of Origin: USA
Prediction: Spice

Toothpaste

Country of Origin: USA
Prediction: Toothpaste

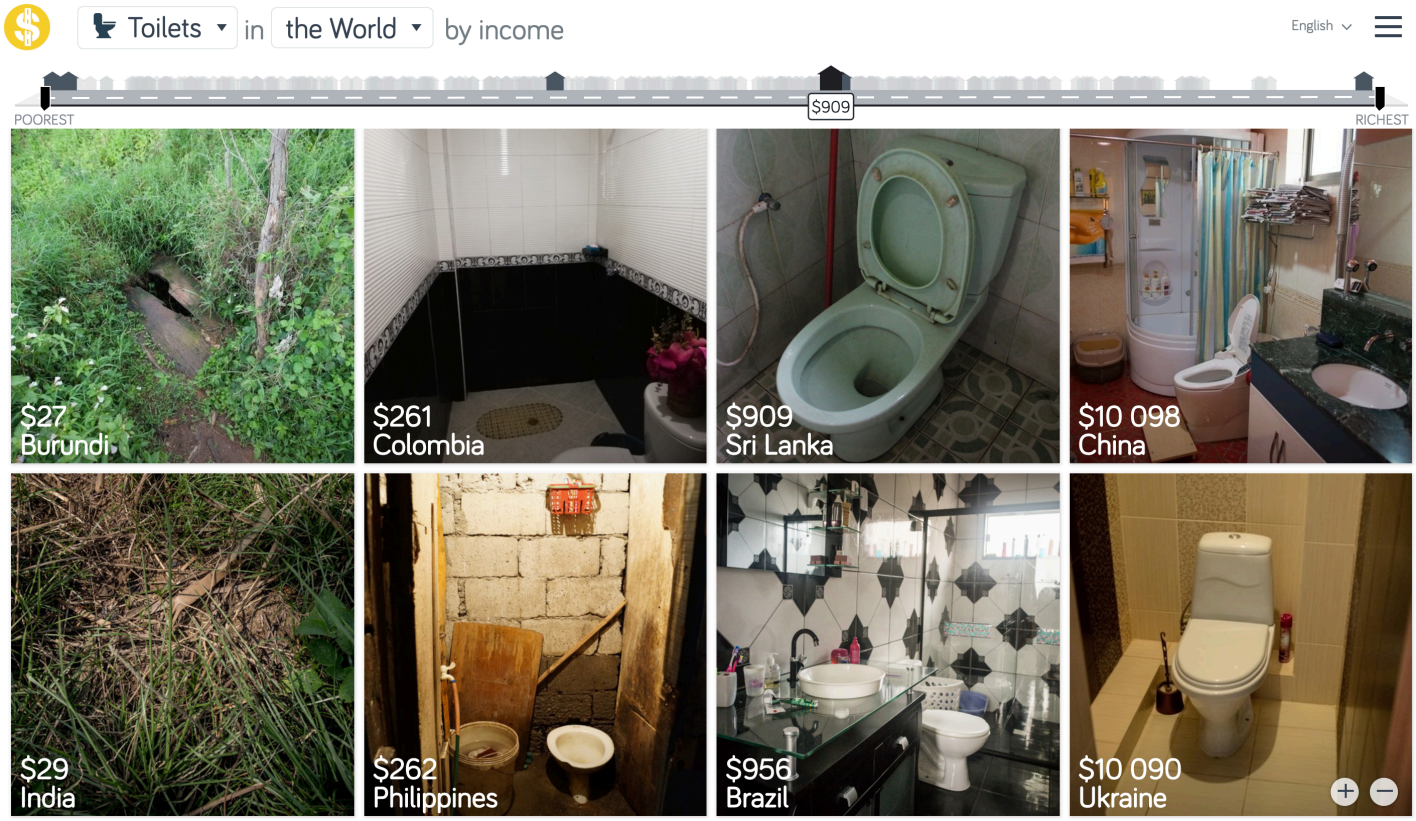
Country of Origin: Nepal
Prediction: Food

Country of Origin: Philippines
Prediction: Beer

Country of Origin: Burundi
Prediction: Wood

Dollar Street

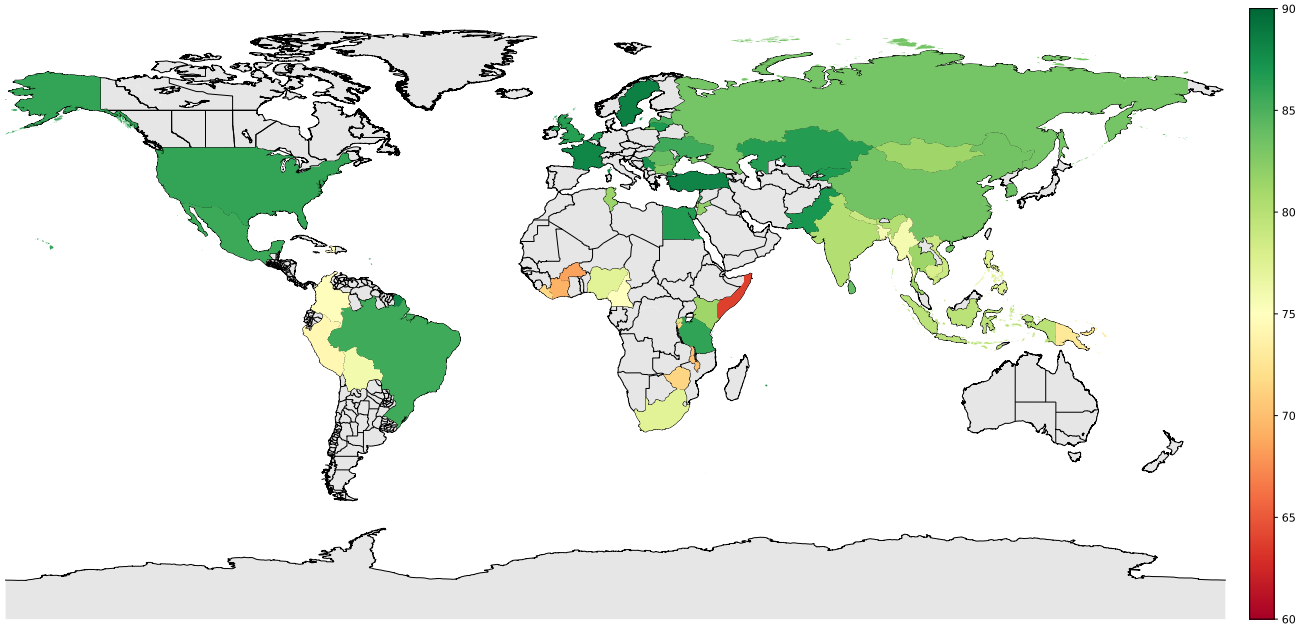
- Photo collection gathered by Gapminder to show how people across the world live
- Photos are annotated with object class, country, and family income
- We used ~20,000 photos from 117 classes



https://www.gapminder.org/dollar-street

Results

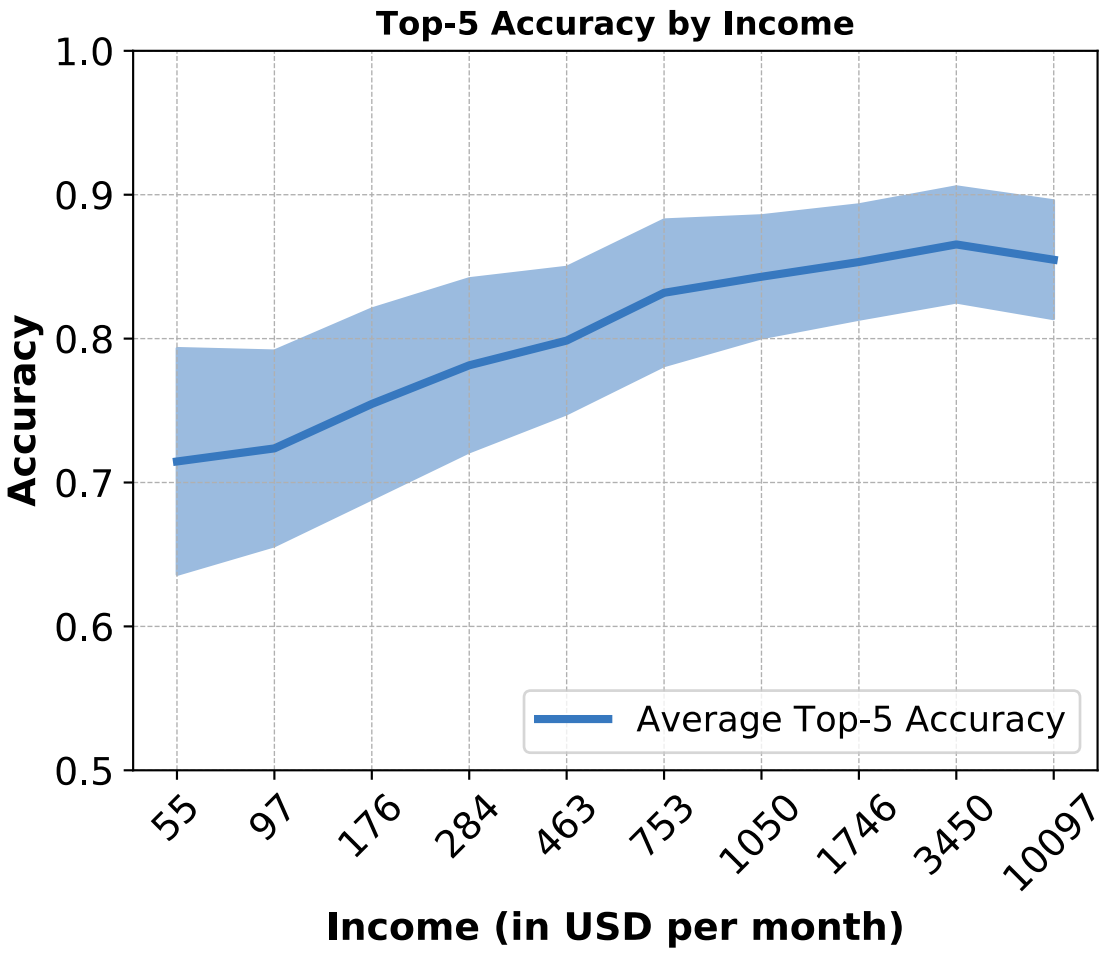
- Average top-5 accuracy of all image recognition systems:
  - Amazon Rekognition
  - Google Cloud Vision
  - Clarifai
  - Microsoft Azure
  - IBM Watson
- Accuracy varies per country
- Results are consistent across all services analyzed



Red: 60%; Yellow: 75%; Green: 90%

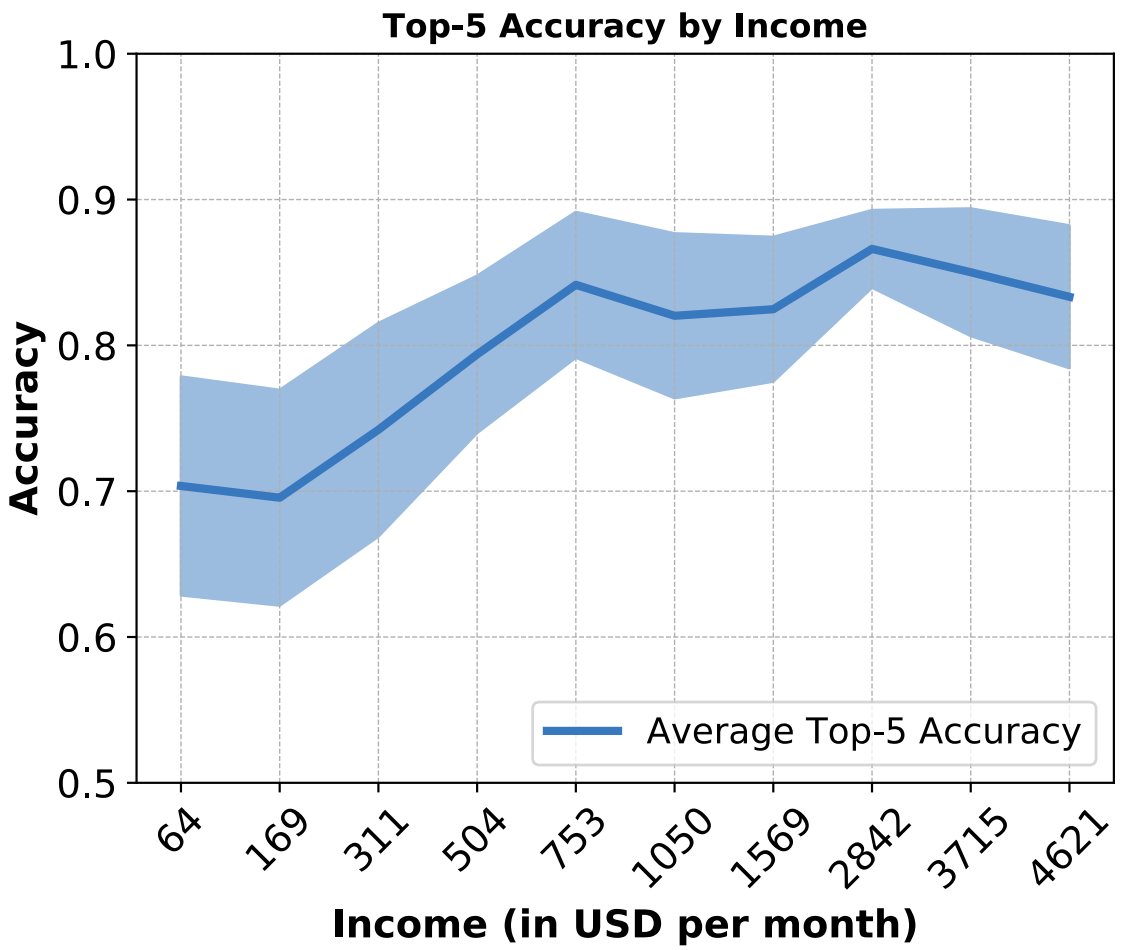
Results

- Average top-5 accuracy of all image recognition systems:
  - Amazon Rekognition
  - Google Cloud Vision
  - Clarifai
  - Microsoft Azure
  - IBM Watson
- Accuracy varies per country
- Results are consistent across all services analyzed



Results

- Analysis of only India shows this is not only due to income correlating with location



Main problems

- Dataset collection relies on services that are primarily popular in the West

Main problems

- Dataset collection relies on services that are primarily popular in the West
- Most dataset collections start with English queries

Top results for “Wedding”

Top results for “शादी” (Wedding in Hindi)

Main problems

- Dataset collection relies on services that are primarily popular in the West
- Most dataset collections start with English queries

Top results for “Wedding”

Top results for “शादी” (Wedding in Hindi)

Top results for “Spices”

Top results for “मसाले” (Spices in Hindi)

... to Visual Understanding

facebook
Artificial Intelligence Research

. \ |

Research question

How do we measure image understanding?

Highlights

- We find problems in common benchmarks for visual understanding

Allan Jabri

Armand Joulin

facebook
Artificial Intelligence Research

Visual Question Answering

- How do we measure understanding?

Visual Question Answering

- How do we measure understanding?

- Proposal:
  Given an image, answer questions about that image

What color is the jacket?
-Red and blue.
-Yellow.
-Black.
-Orange.

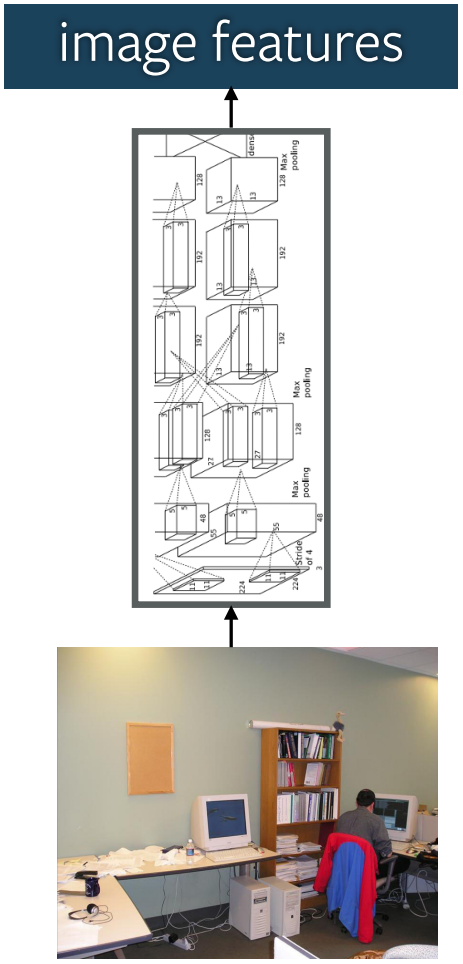
How many cars are parked?
-Four.
-Three.
-Five.
-Six.

What event is this?
-A wedding.
-Graduation.
-A funeral.
-A picnic.

When is this scene taking place?
-Day time.
-Night time.
-Evening.
-Morning.

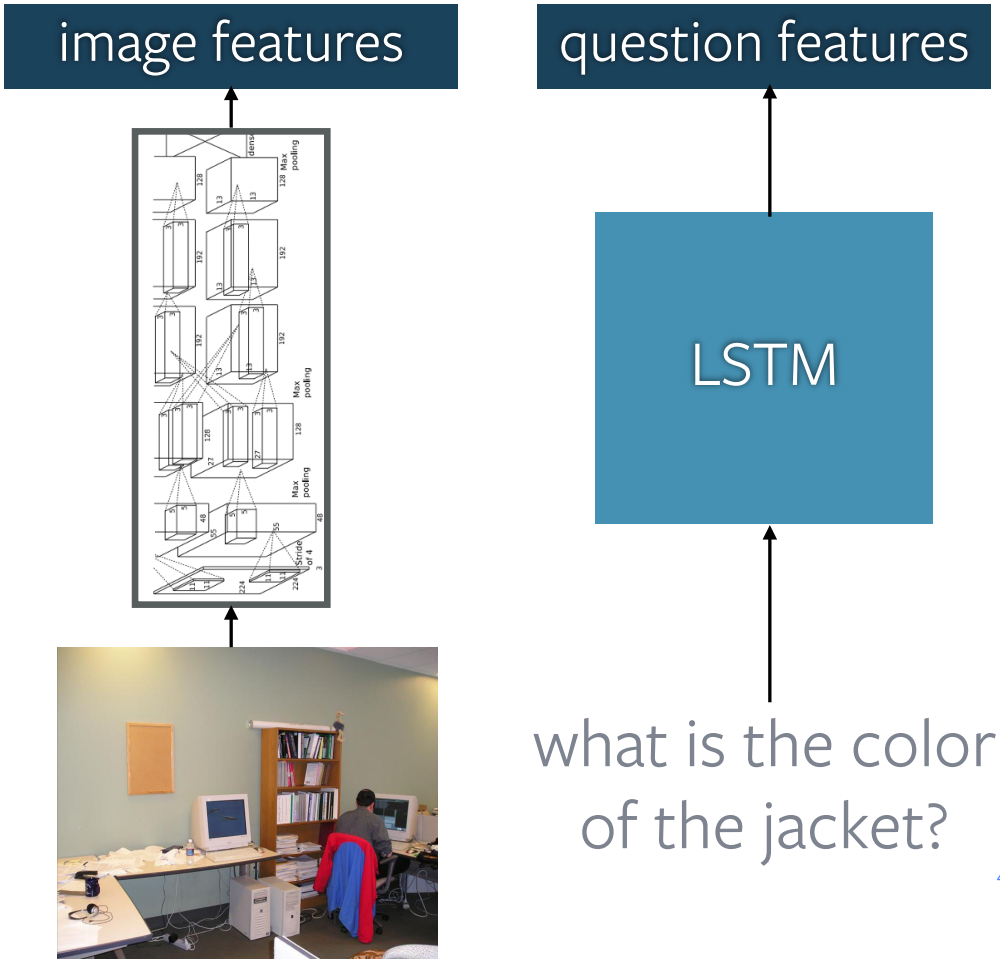
Models for VQA

- Feed image through convolutional network



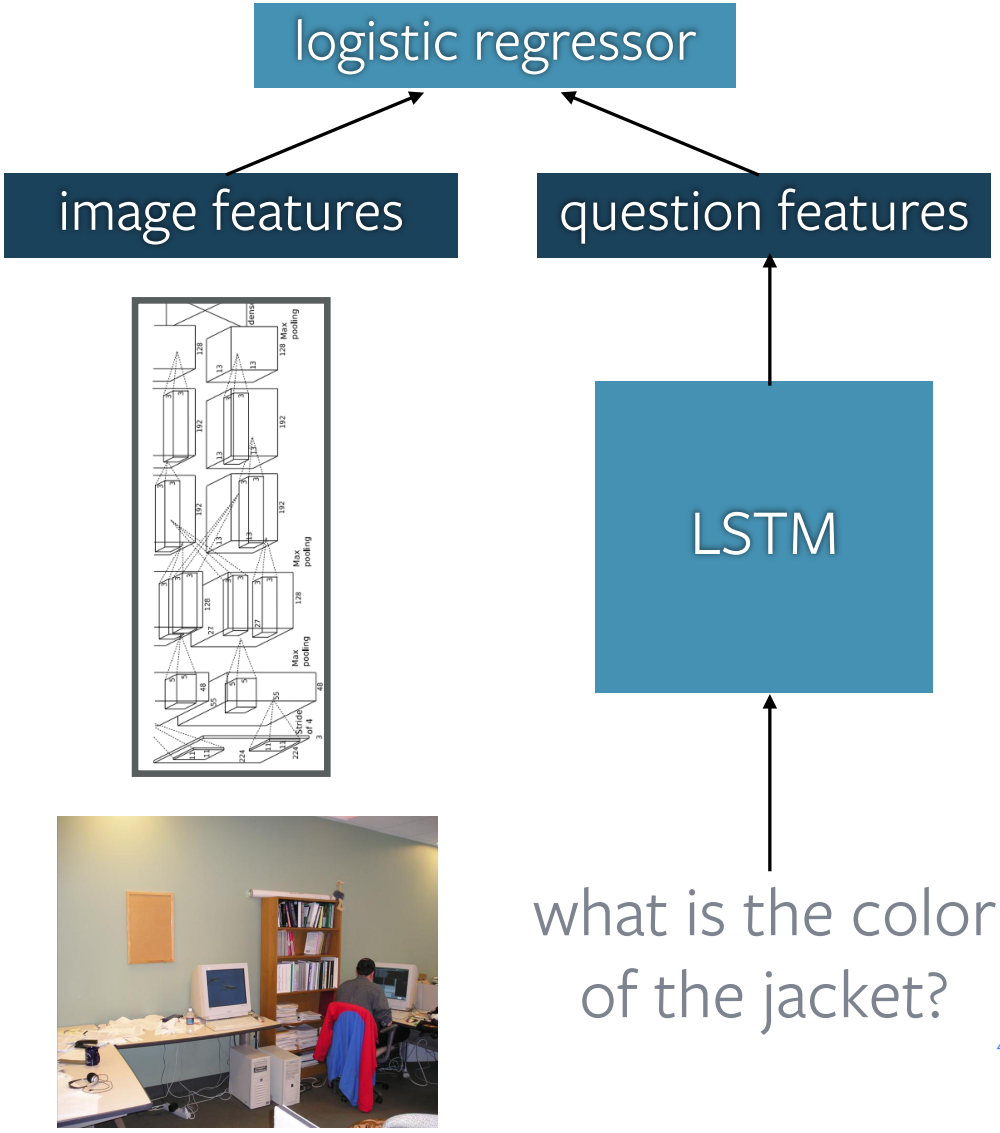
Models for VQA

- Feed image through convolutional network
- Feed question through recurrent network



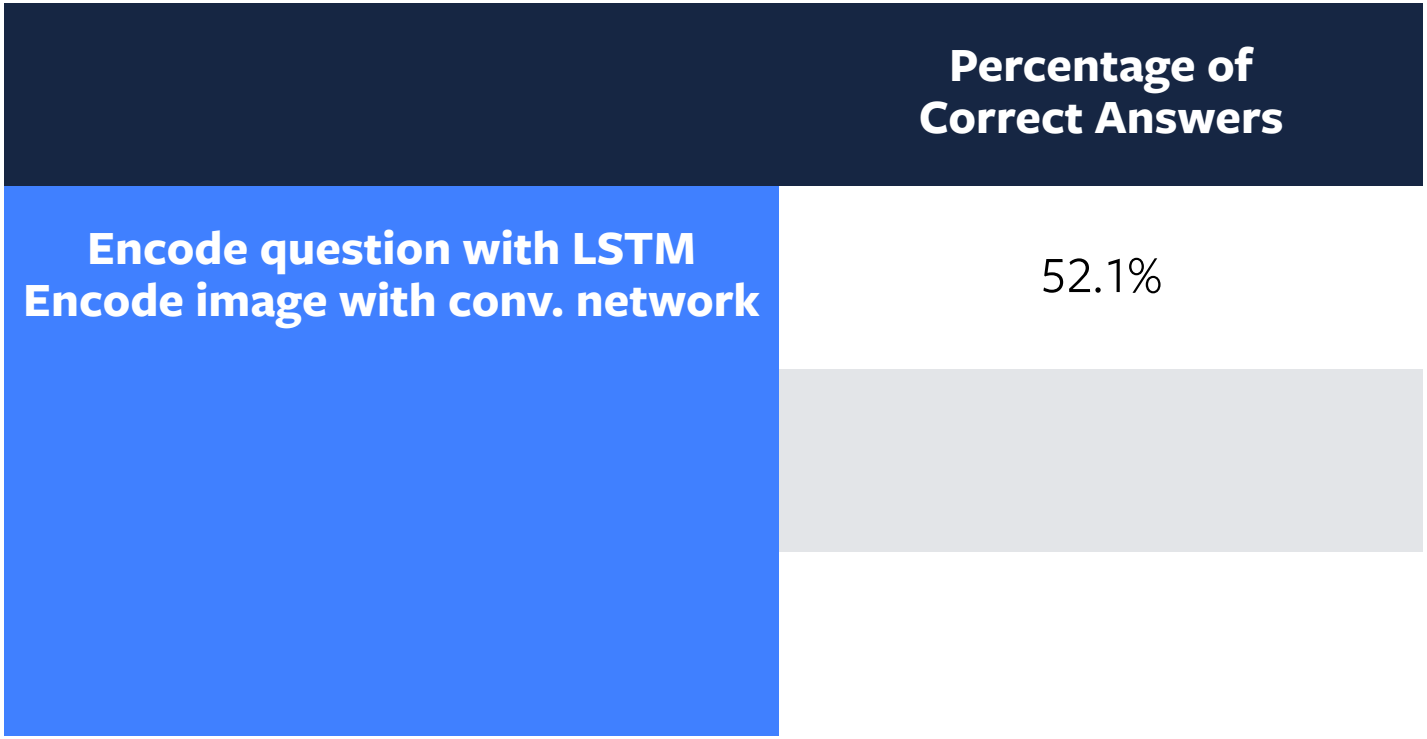
Models for VQA

- Feed image through convolutional network
- Feed question through recurrent network
- Apply multi-class logistic regressor to predict answer



Visual Question Answering

- Train on a collection of 70K multiple-choice questions
- Measure accuracy on 40K held-out questions



Clever Hans

Are we building horses?

- Take multiple-choice visual question
- Throw away both image and question
- Encode answer using word2vec features
- Train binary classifier to predict whether or not answer is correct
- Test time: Predict highest-scoring answer

Visual Question Answering

- Our simple baseline outperforms the state-of-the-art

	Percentage of Correct Answers
Encode question with LSTM Encode image with conv. network	52.1%
Simple model	52.9%

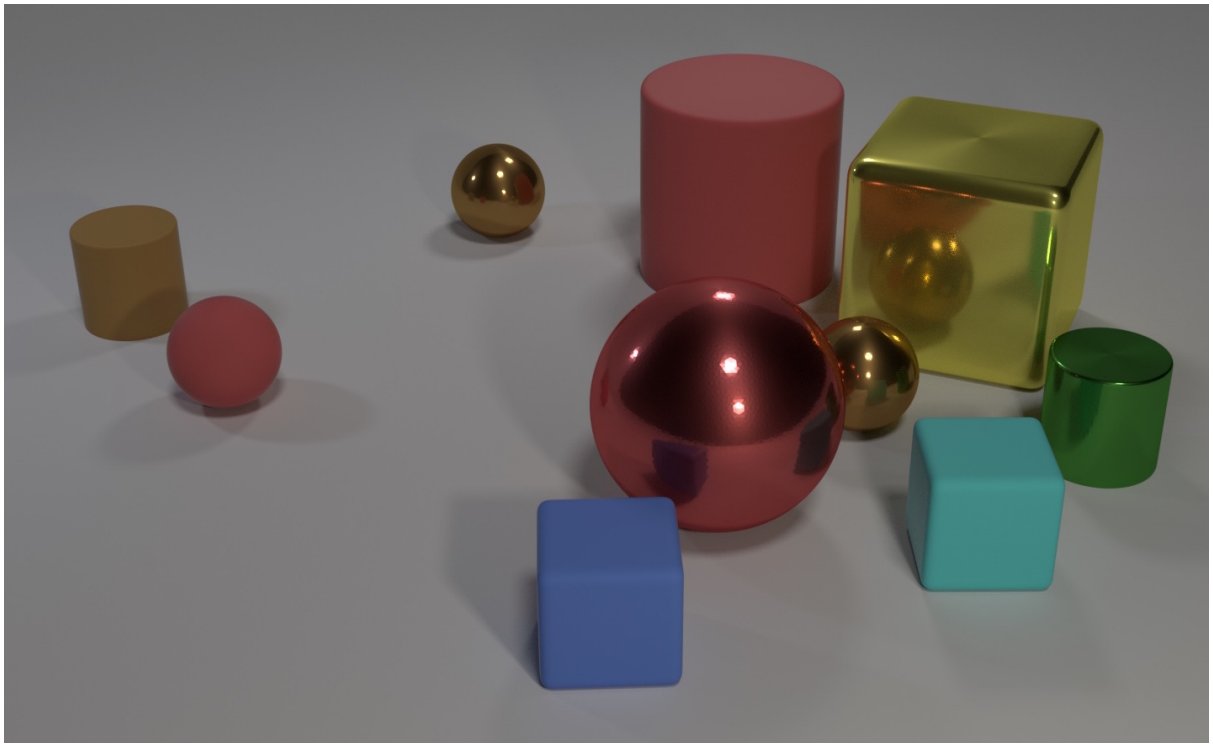
Visual Question Answering

- Does looking at the image and question help? Sure.

	Percentage of Correct Answers
Encode question with LSTM Encode image with conv. network	52.1%
Simple model	52.9%
Simple model++	67.1%

CLEVR

- Visual reasoning benchmark that cannot be solved by a "horse"



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

Research question

How do we develop better benchmarks for image captioning?

Highlights

- Image captioning systems are difficult to evaluate
- We propose binary image selection as an alternative evaluation

Hexiang Hu

Ishan Misra

facebook
Artificial Intelligence Research

Image Captioning

- Evaluating relevance of captions to images is very difficult

- Does a captioning system really possess visual understanding?

Image Captioning

- Evaluating relevance of captions to images is very difficult
- Does a captioning system really possess visual understanding?

Image Captioning

- Evaluating relevance of captions to images is very difficult
- Does a captioning system really possess visual understanding?

A bunch of luggage sitting on top of a floor.

Image Captioning

- Evaluating relevance of captions to images is very difficult
- Does a captioning system really possess visual understanding?

A bunch of luggage sitting on top of a floor.

CIDEr-D: 55.96

Image Captioning

- Evaluating relevance of captions to images is very difficult
- Does a captioning system really possess visual understanding?

A pile of garbage sitting next to a trash can.

Image Captioning

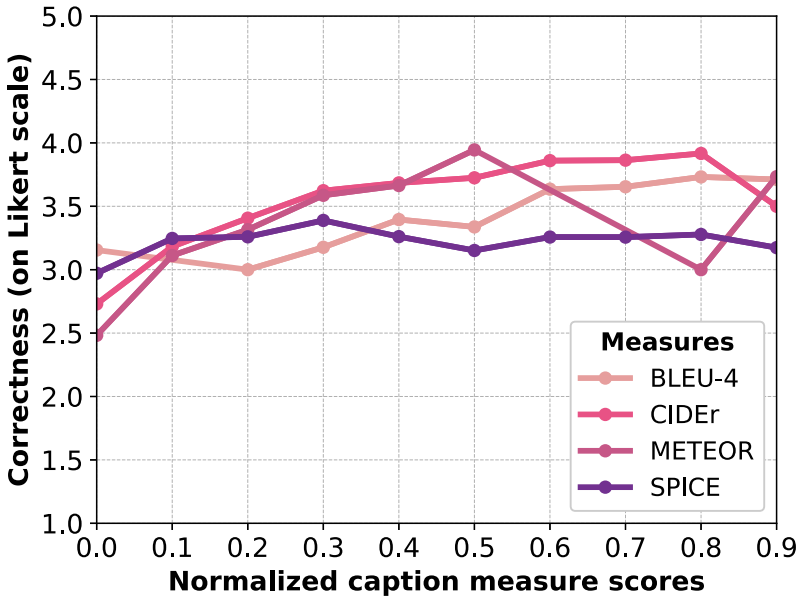
- Evaluating relevance of captions to images is very difficult
- Does a captioning system really possess visual understanding?

A pile of garbage sitting next to a trash can.

CIDEr-D: 0.14

Image Captioning

- Captioning scores correlate poorly with human evaluations of *correctness*:



* Annotations gathered using COCO guidelines for human evaluation.

Image Retrieval

- Is text-based image retrieval a good alternative?
- No: does not provide true negatives.

Text query: A person wearing a banana headdress and necklace.

Retrieved image

Correct image

Text query: There is a green clock in the street.

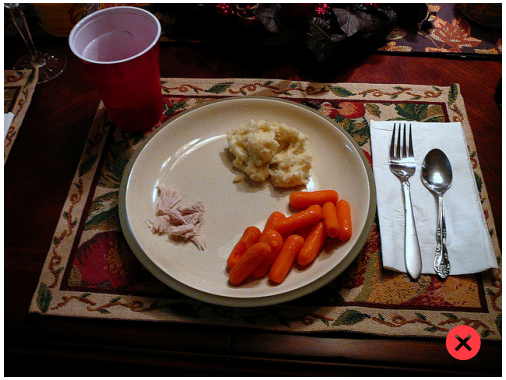
Retrieved image

Correct image

BISON: Binary Image SelectiON

- Given a text query, pick one of two images
- BISON is designed such that both images are similar, but one is a negative for the query

Text query: Plates filled with carrots and beets on a white table.



Negative image

Positive image

Text query: Yellow shirted tennis player looking for incoming ball.

Positive image

Negative image

Analyzing Systems

- Performance of captioning and retrieval systems on COCO-val

		COCO-validation			COCO-BISON
		BLEU-4	**CIDEr-D**	**Recall@1**	**BISON**
Captioning	Show & Tell (Vinyals *et al.*, 2015)	32.35	97.20	-	78.59
	Show & Tell + Attention (Xu *et al.*, 2015)	33.49	101.55	-	82.04
	UpDown (Anderson *et al.*, 2018)	**34.53**	**105.40**	-	84.04
Retrieval	Convnet + BoW	-	-	45.19	81.48
	Convnet + BiGRU (Faghri *et al.*, 2018)	-	-	49.34	85.46
	SCAN i2t (Lee *et al.*, 2018)	-	-	52.35	86.40
	SCAN t2i (Lee *et al.*, 2018)	-	-	**54.10**	87.50
	Human	21.70*	85.40*	-	**100.0**

* Human scores computed on COCO test set.

Try it yourself at http://github.com/facebookresearch/binary-image-selection

Research question

How do we move beyond image understanding to physical reasoning?

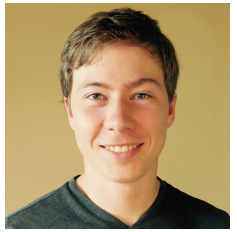
Highlights

- PHYRE is a new benchmark for physical reasoning


Anton Bakhtin


Laura Gustafson

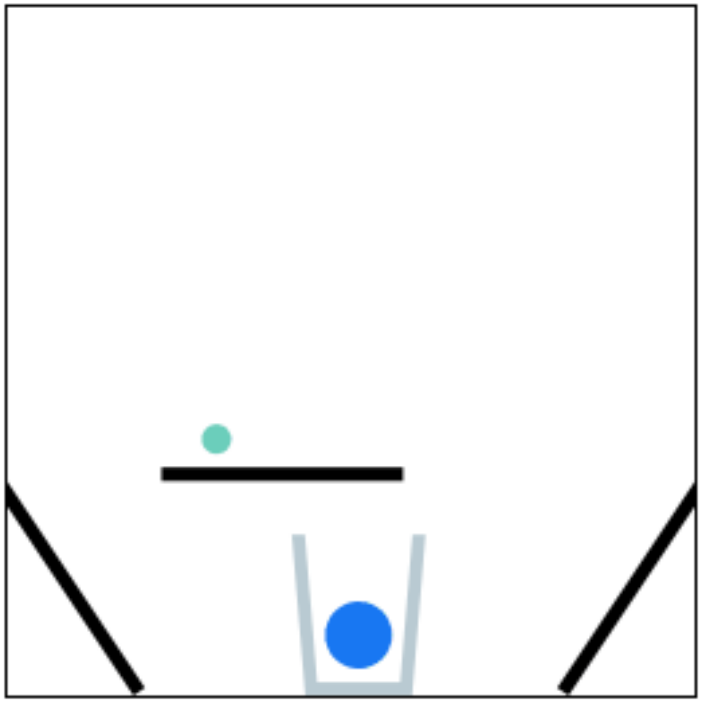

Ross Girshick


Justin Johnson

facebook
Artificial Intelligence Research

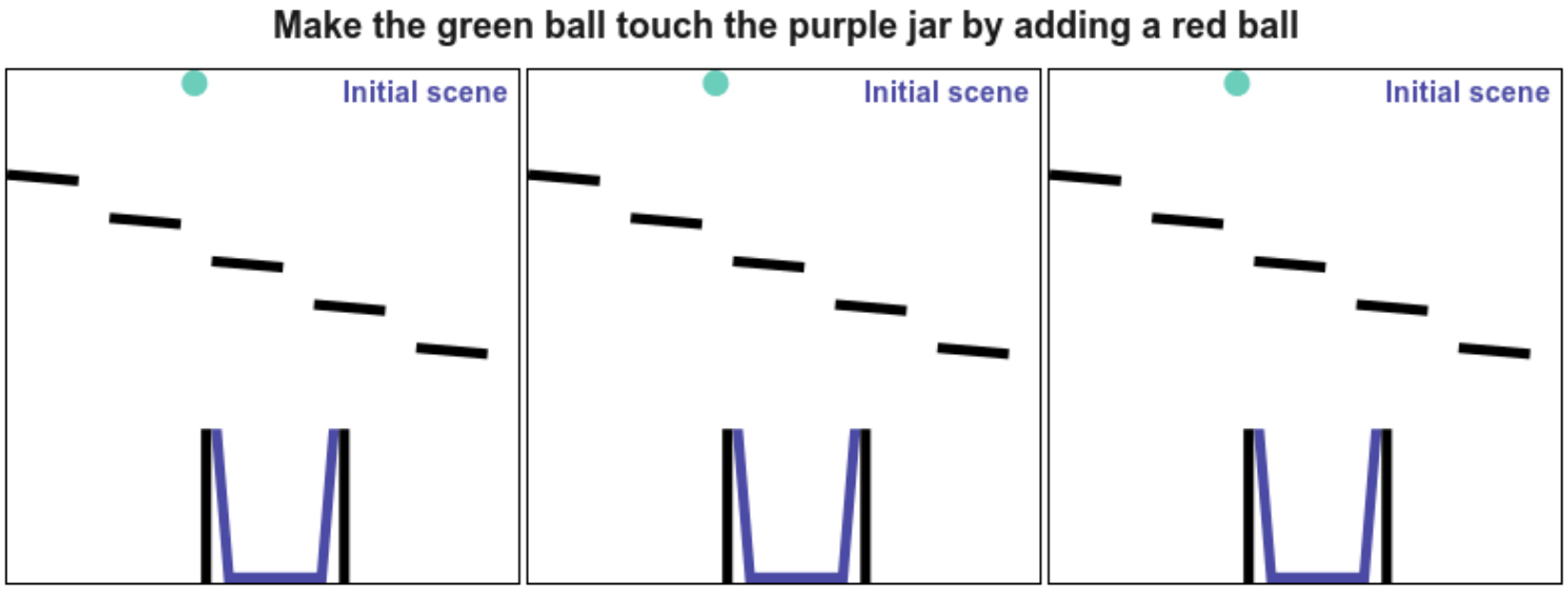
PHYRE

- New benchmark for test ability of AI agent to perform physical reasoning:



PHYRE

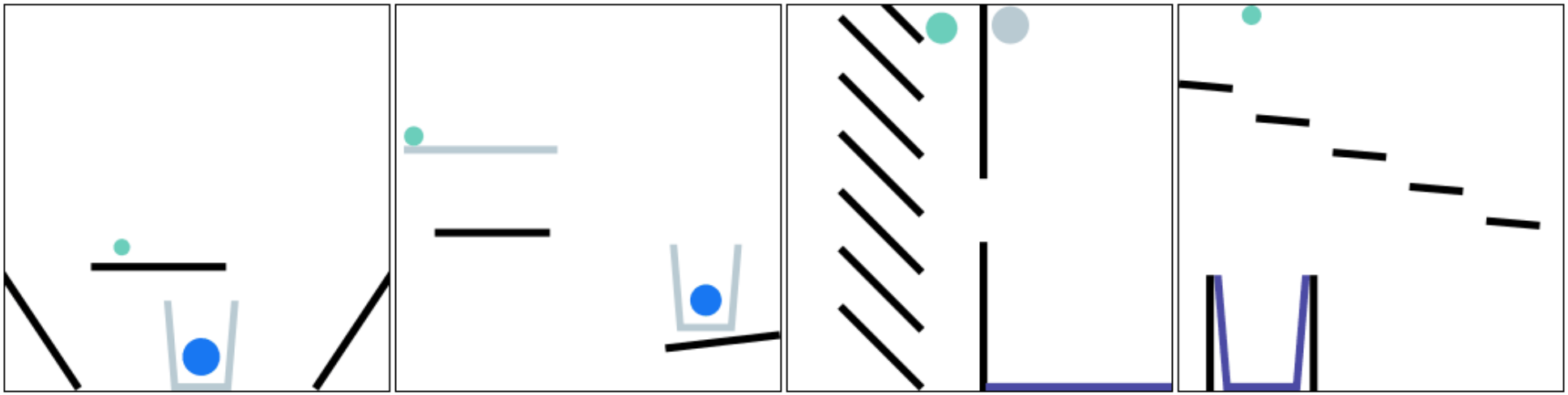
- Overview of the stages in the PHYRE benchmark:



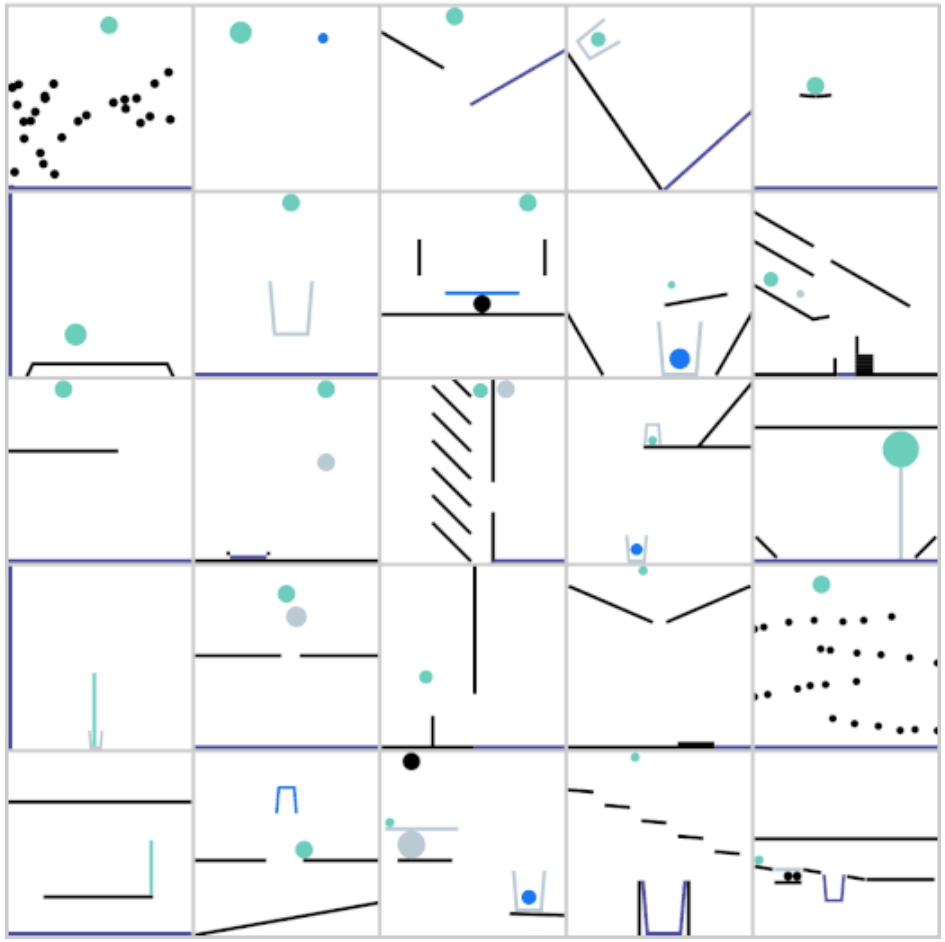
PHYRE

- Solution strategies the agent needs to use are very diverse:

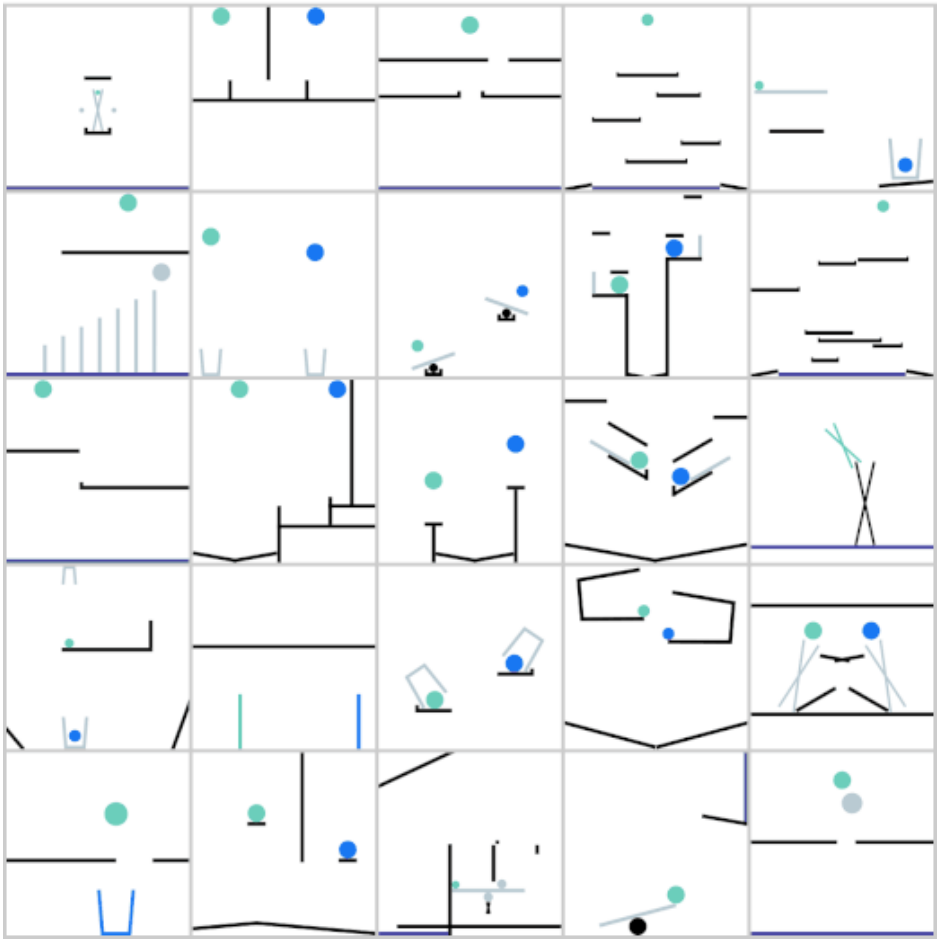
Make the green ball touch the blue/purple object by adding red objects



PHYRE-B Tier



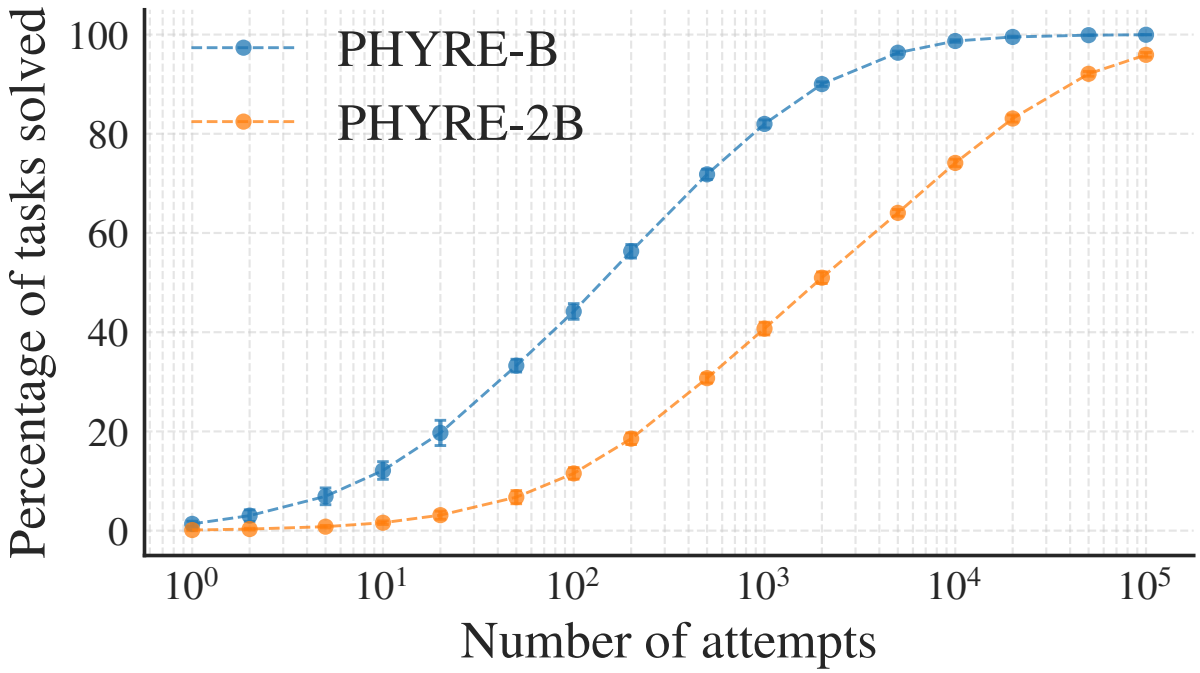
PHYRE-2B Tier



http://www.phyre.ai

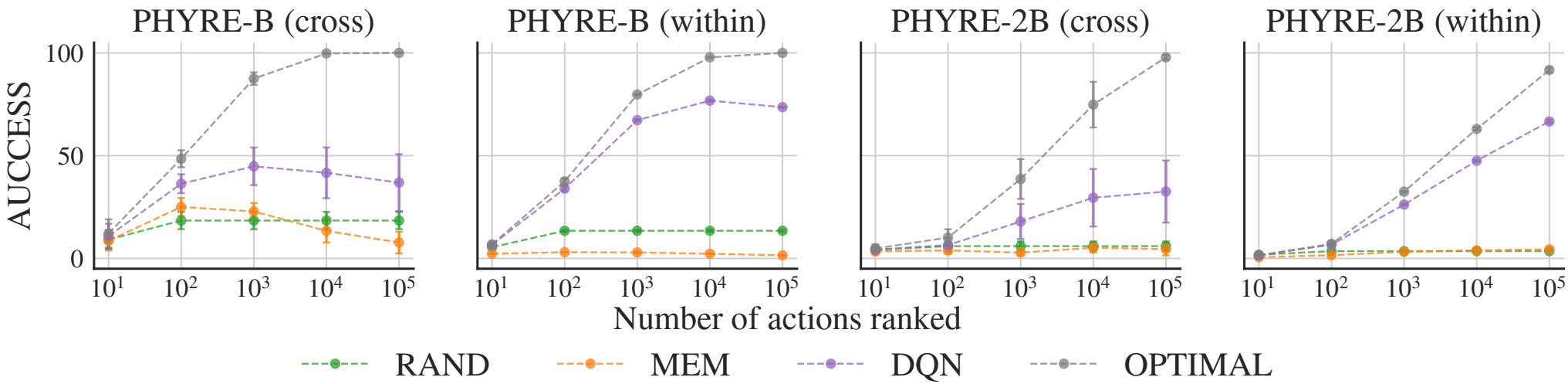
PHYRE

- Tasks cannot be solved well by random search



PHYRE

- Agents are still far away from doing optimal ranking (which is non-optimal itself):

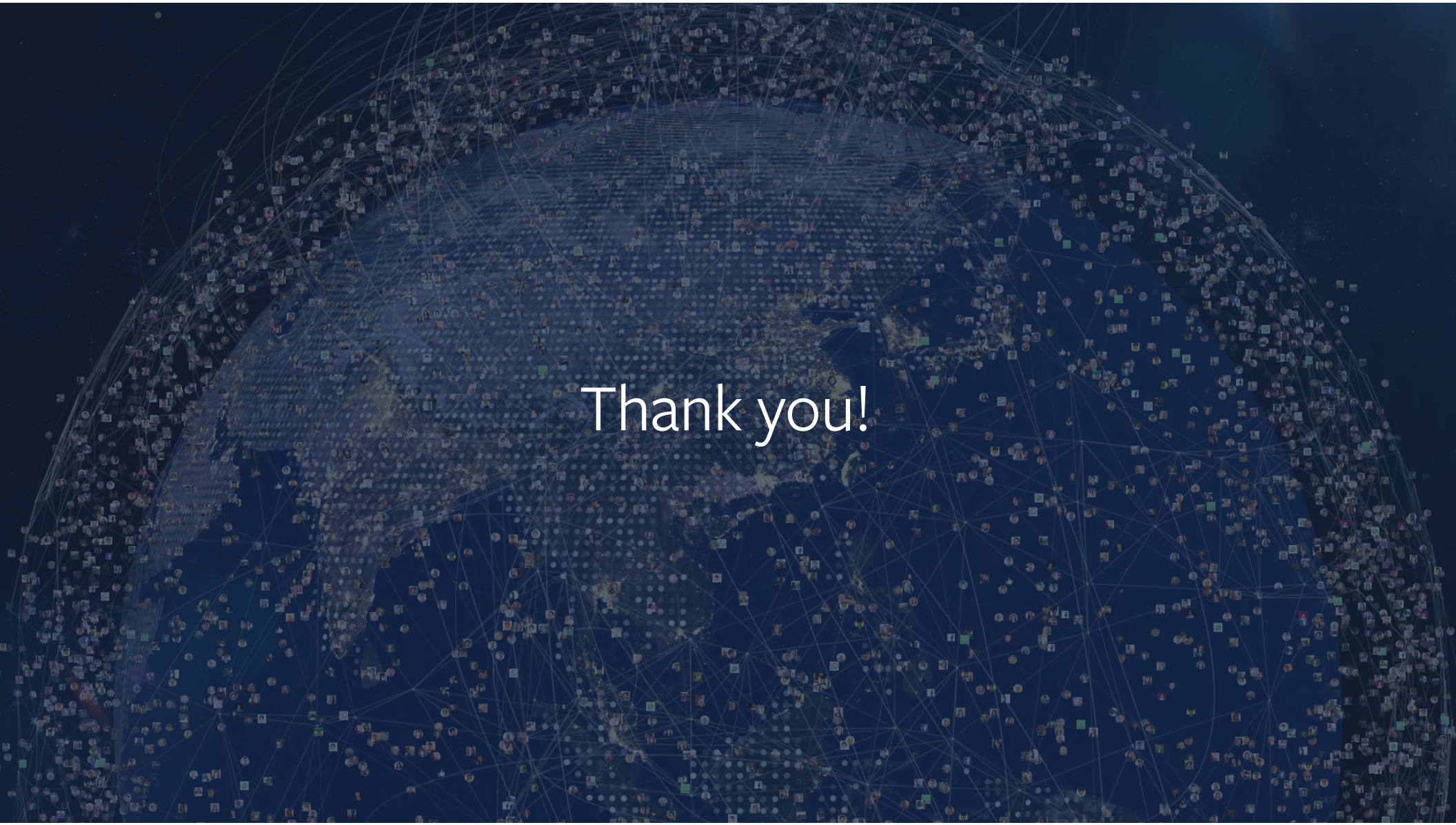


Conclusion

Visual recognition works but we still have a way to go towards visual understanding.

Details

- Large-scale training of modern convolutional networks works great
- Care is needed to prevent networks from having undesired biases
- Visual understanding is still difficult and work-in-progress


Thank you!