



Visual Question Answering Challenge



Yash Goyal
(Georgia Tech)



Aishwarya Agrawal
(Georgia Tech)



Dhruv Batra
(Georgia Tech /
FAIR)



Devi Parikh
(Georgia Tech /
FAIR)



VQA Task



VQA Task



What is the mustache
made of?

VQA Task



AI System

What is the mustache
made of?

VQA Task



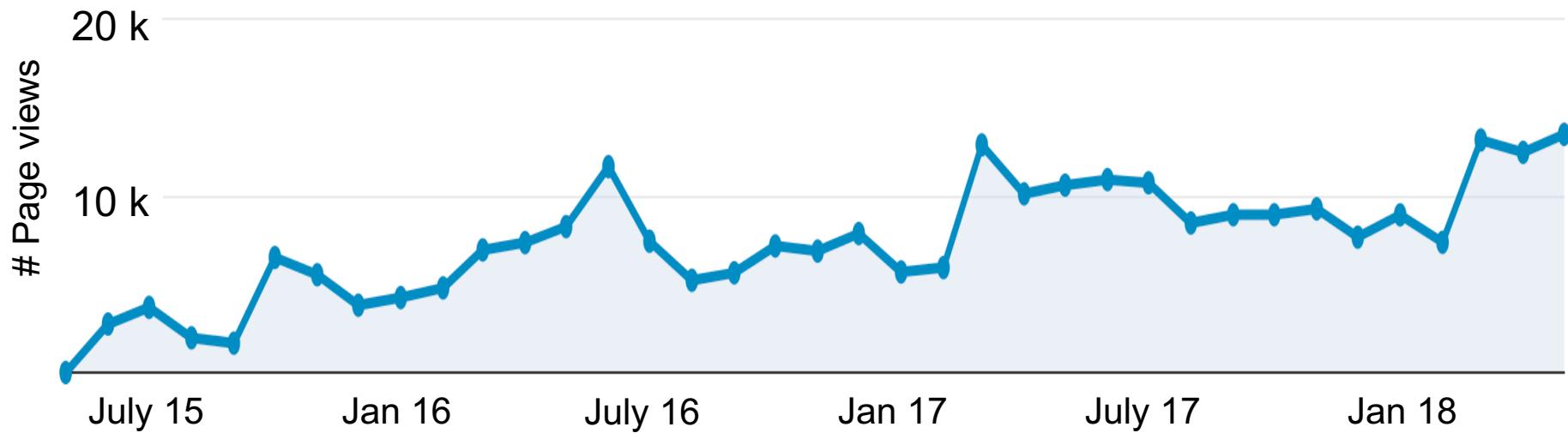
AI System

bananas

What is the mustache
made of?

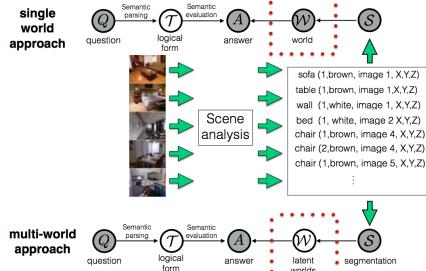
Interest in VQA

(<http://www.visualqa.org/>)



13k page views/month
during VQA Challenge 2018

VQA Timeline



Geman et al., PNAS 14
Malinowski & Fritz, NIPS 14

VQA Real Image Challenge (Open-Ended)
Organized by vqateam
This challenge evaluates algorithms on the VQA Open-Ended task for the dataset built on top of MSCOCO test2015 real images.

Oct 06, 2015-Jun 05, 2016
88 participants

Edit Unpublish Participants Submissions Leaderboard

Featured Challenge
Explore other past, ongoing and upcoming challenges.
[View All](#)

VQA Real Image Challenge (Open-Ended) 2017
Organized by VQA Team
Recent progress in computer vision and natural language processing has demonstrated that lower level tasks are much closer to being solved. We believe that this is ripe to pursue higher-level tasks, one of which is Visual Question Answering (VQA), where the goal is to be able to understand the semantics of scenes well enough to be able to answer open-ended, free-form natural language questions (asked by humans) about images...
Status: In Progress [view more](#)



VQA v1.0 dataset released



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?



Who is wearing glasses?
man woman



Where is the child sitting?
fridge arms

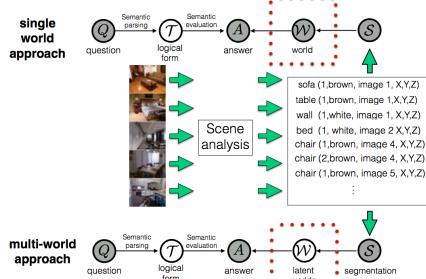


Is the umbrella upside down?
yes no

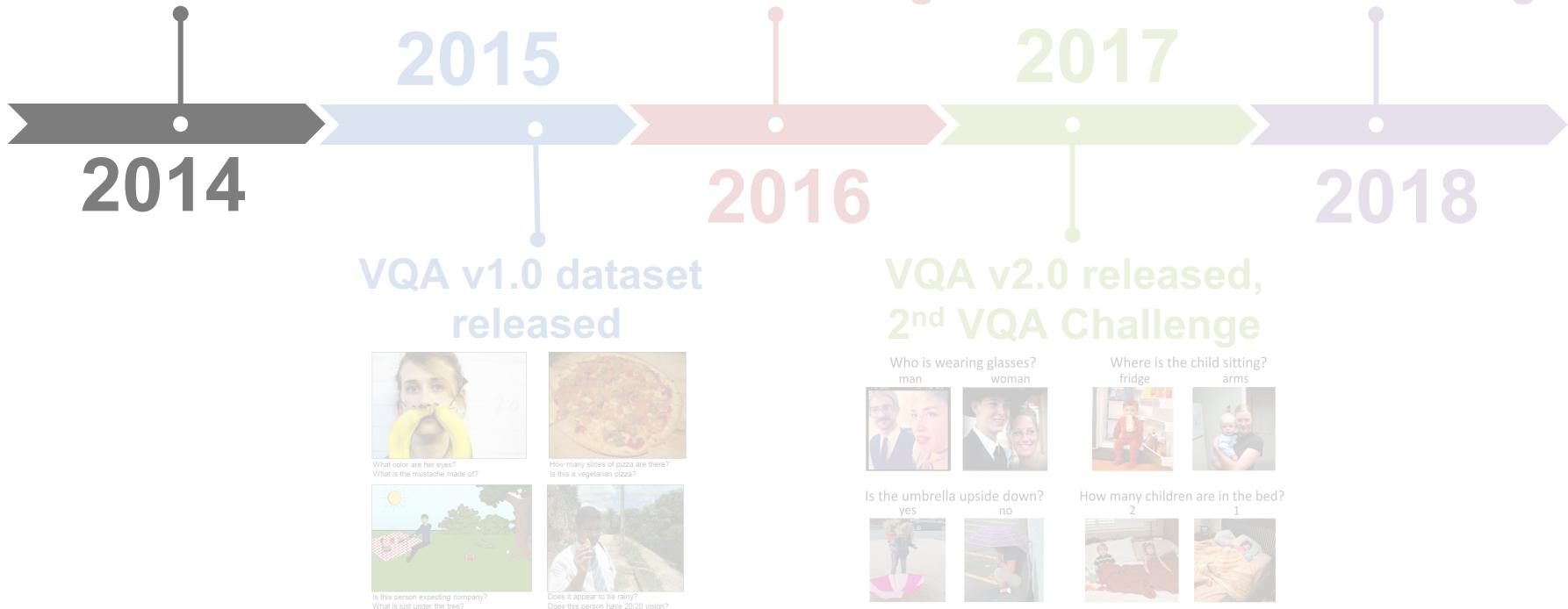


How many children are in the bed?
2 1

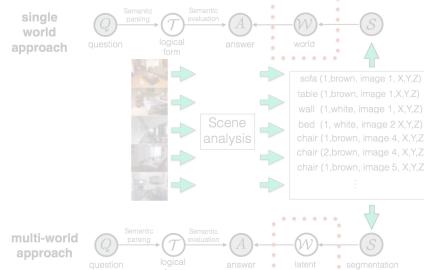
VQA Timeline



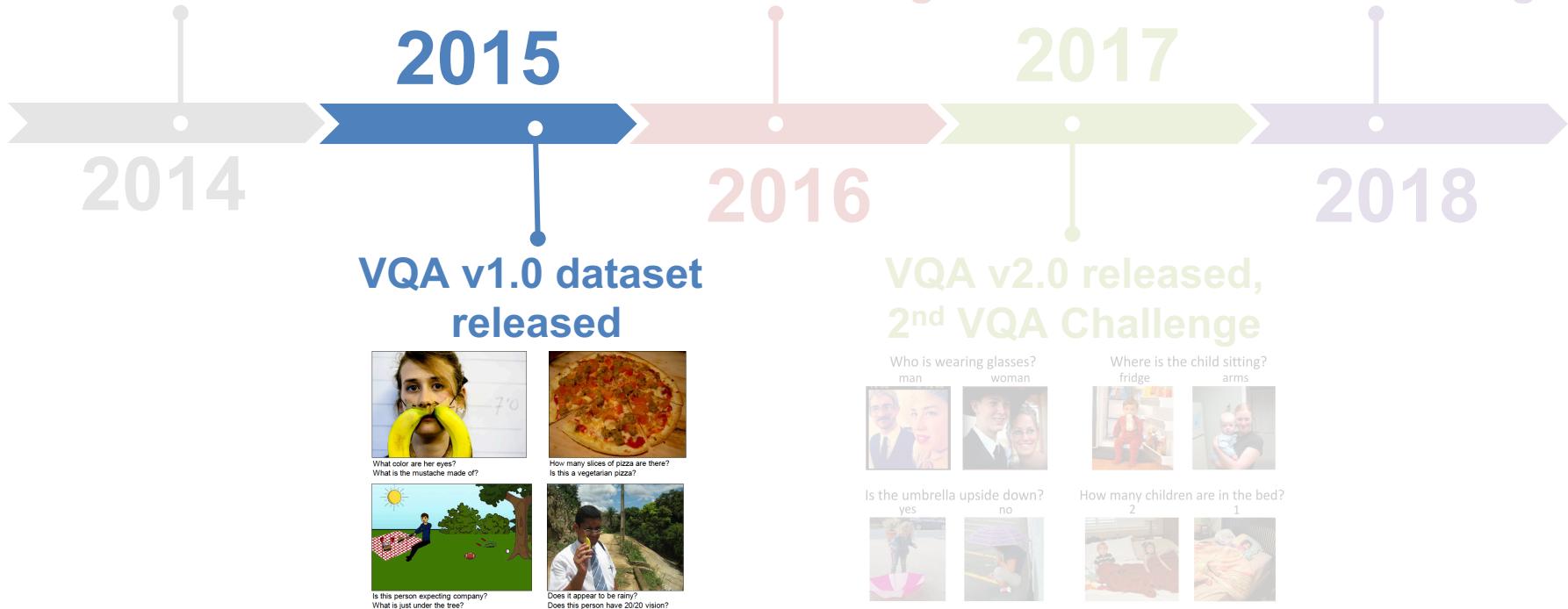
Geman et al., PNAS 14
Malinowski & Fritz, NIPS 14



VQA Timeline



Geman et al., PNAS 14
Malinowski & Fritz, NIPS 14



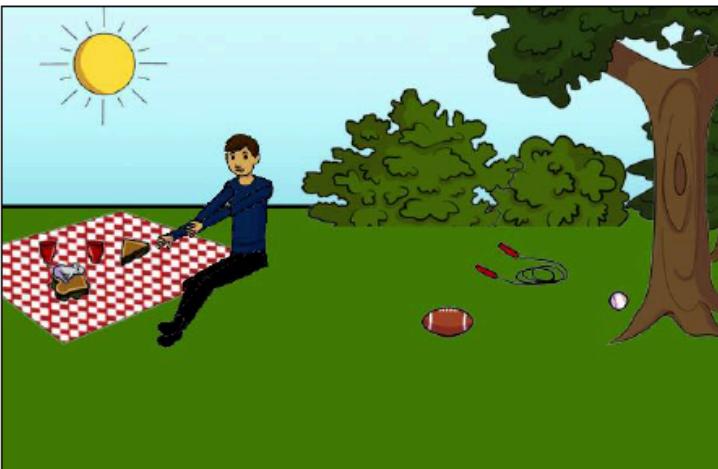
VQA v1.0 Dataset



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset



About
objects

What color are her eyes?

What is the mustache made of?



Is this person expecting company?

What is just under the tree?



How many slices of pizza are there?

Is this a vegetarian pizza?



Does it appear to be rainy?

Does this person have 20/20 vision?

VQA v1.0 Dataset

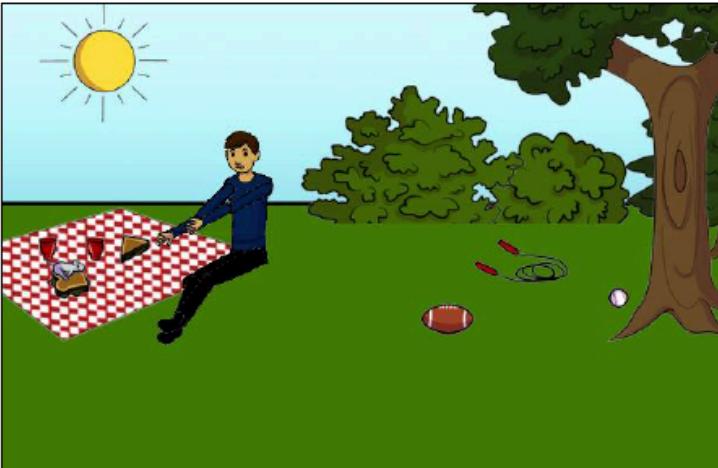


What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?

Fine-grained
recognition



Is this person expecting company?
What is just under the tree?

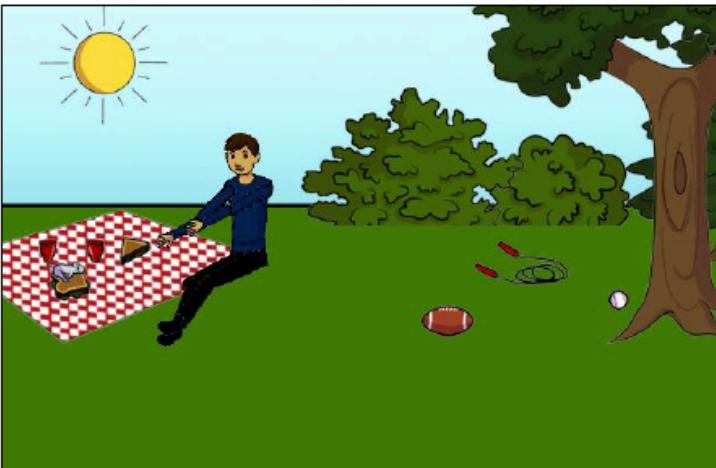


Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset



What color are her eyes?
What is the mustache made of?



Is this person expecting company?
What is just under the tree?



How many slices of pizza are there?
Is this a vegetarian pizza?

Counting



Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset Stats

- >200K images (COCO)

VQA v1.0 Dataset Stats

- >200K images (COCO)
- >600K questions (3 per image)

VQA v1.0 Dataset Stats

- >200K images (COCO)
- >600K questions (3 per image)
- ~8M answers (10 with image + 3 w/o image)

Accuracy Metric

$$\text{Acc}(\textit{ans}) = \min \left\{ \frac{\#\text{humans that said } \textit{ans}}{3}, 1 \right\}$$

1940. COCO_train2014_000000012015



Open-Ended/Multiple-Choice/Ground-Truth

Q: WHAT OBJECT IS THIS

Ground Truth Answers:

- | | |
|----------------|-----------------|
| (1) television | (6) television |
| (2) tv | (7) television |
| (3) tv | (8) tv |
| (4) tv | (9) tv |
| (5) television | (10) television |

Q: How old is this TV?

Ground Truth Answers:

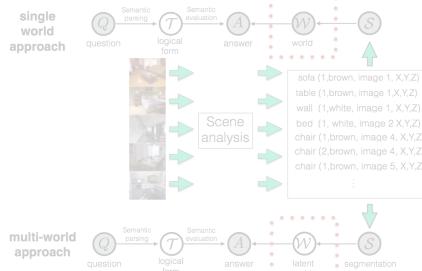
- | | |
|----------------------------|---------------|
| (1) 20 years | (6) old |
| (2) 35 | (7) 80 s |
| (3) old | (8) 30 years |
| (4) more than thirty years | (9) 15 years |
| old | (10) very old |
| (5) old | |

Q: Is this TV upside-down?

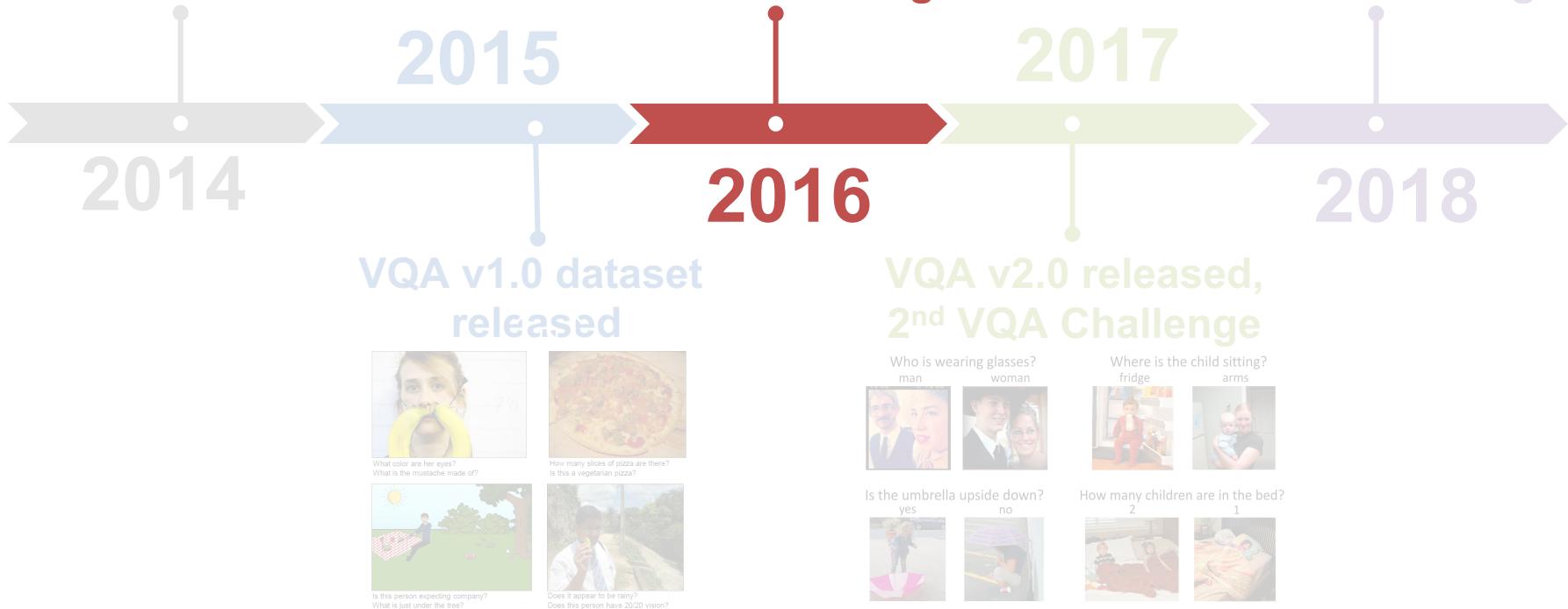
Ground Truth Answers:

- | | |
|---------|----------|
| (1) yes | (6) yes |
| (2) yes | (7) yes |
| (3) yes | (8) yes |
| (4) yes | (9) yes |
| (5) yes | (10) yes |

VQA Timeline



Geman et al., PNAS 14
Malinowski & Fritz, NIPS 14



VQA Challenge 2016

on www.codalab.org



VQA Real Image Challenge (Open-Ended)

Organized by vqateam

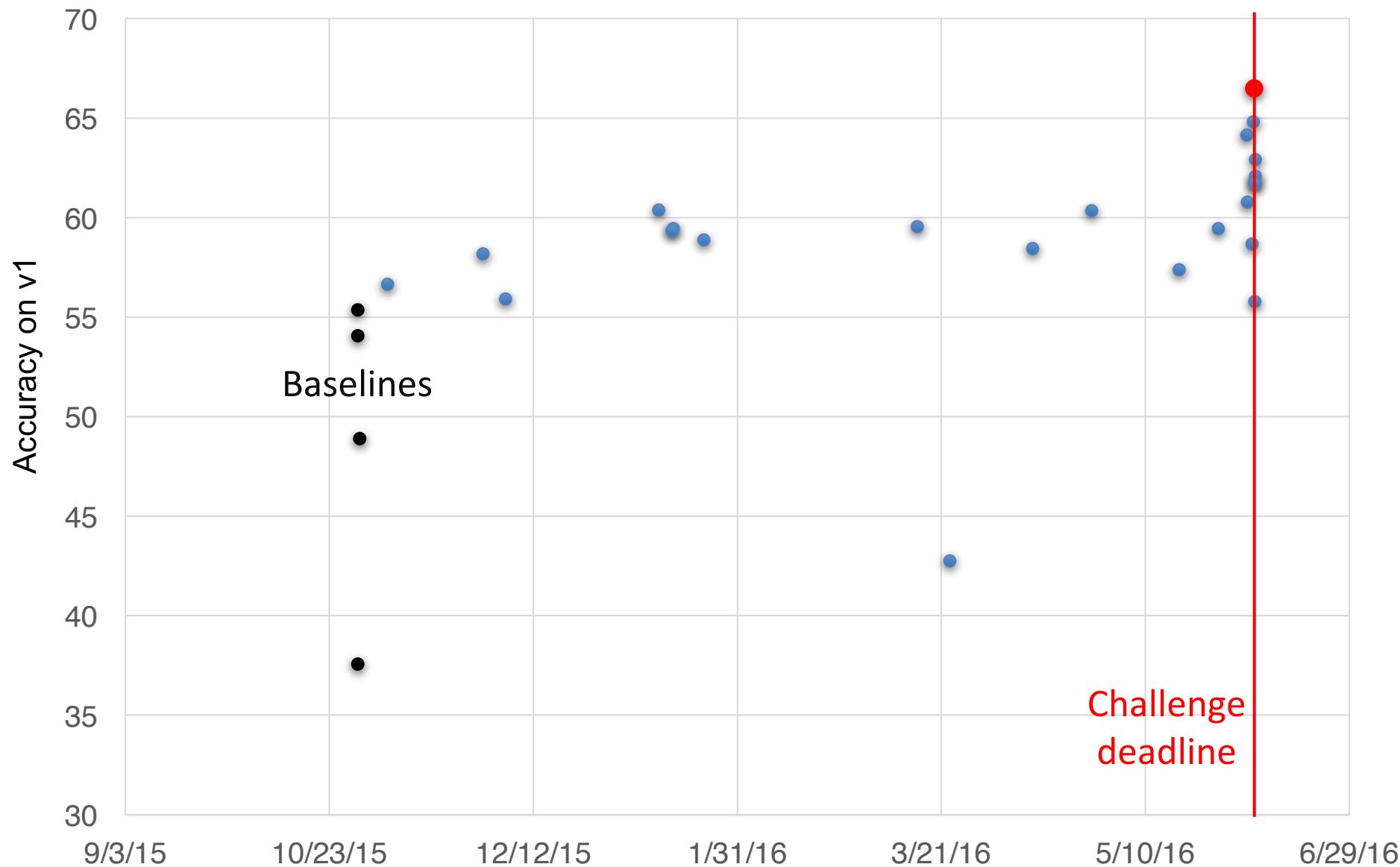
This challenge evaluates algorithms on the VQA Open-Ended task for the dataset built on top of MSCOCO test2015 real images.

Oct 06, 2015-Jun 05, 2016

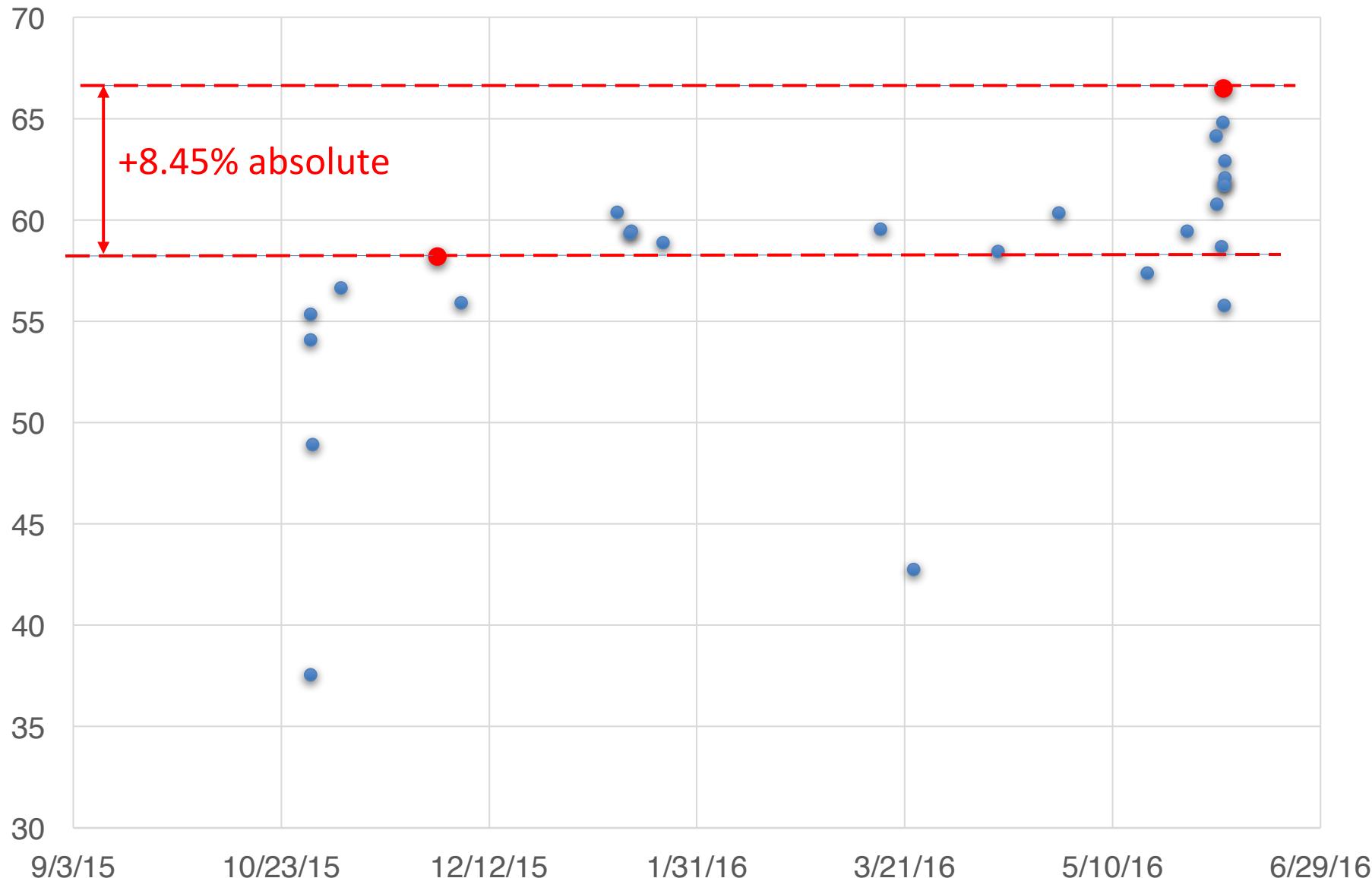
438 participant s



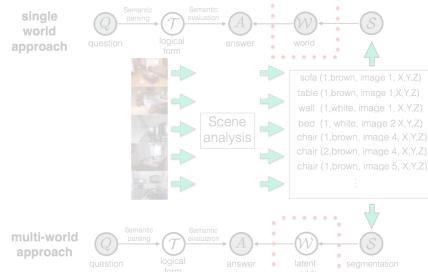
VQA Challenge 2016



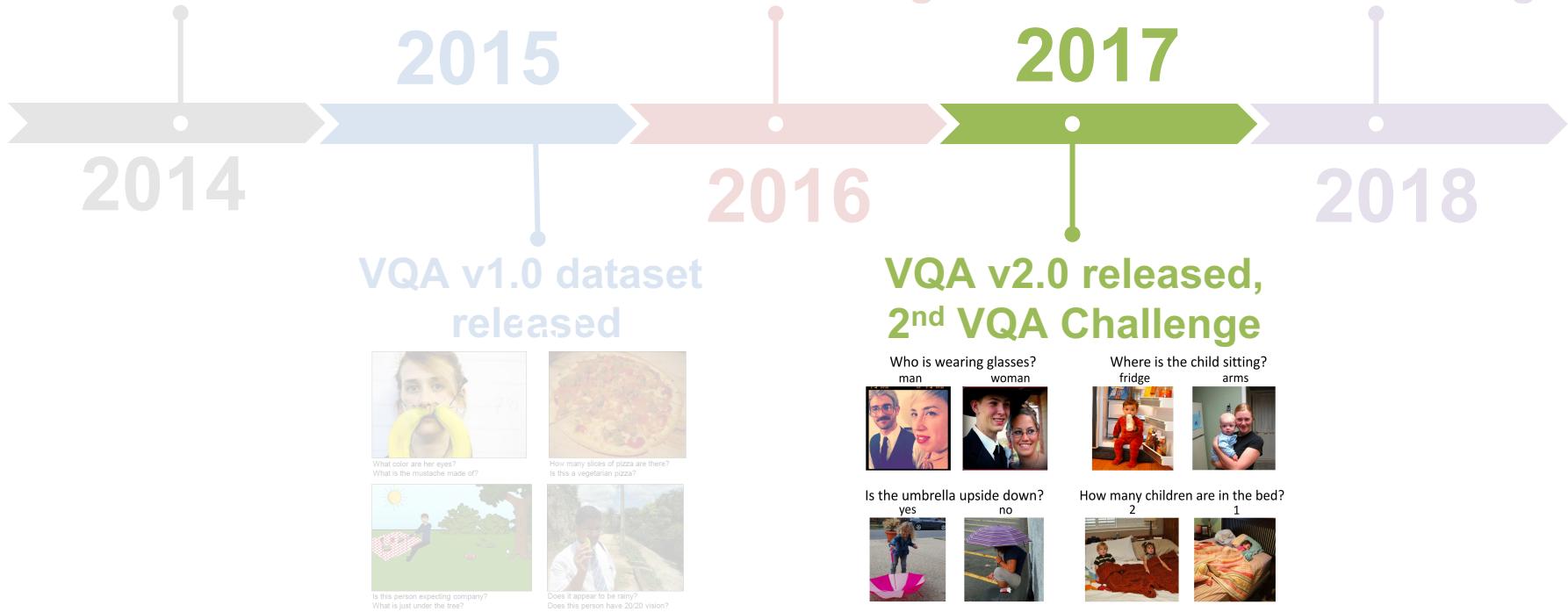
VQA Challenge 2016



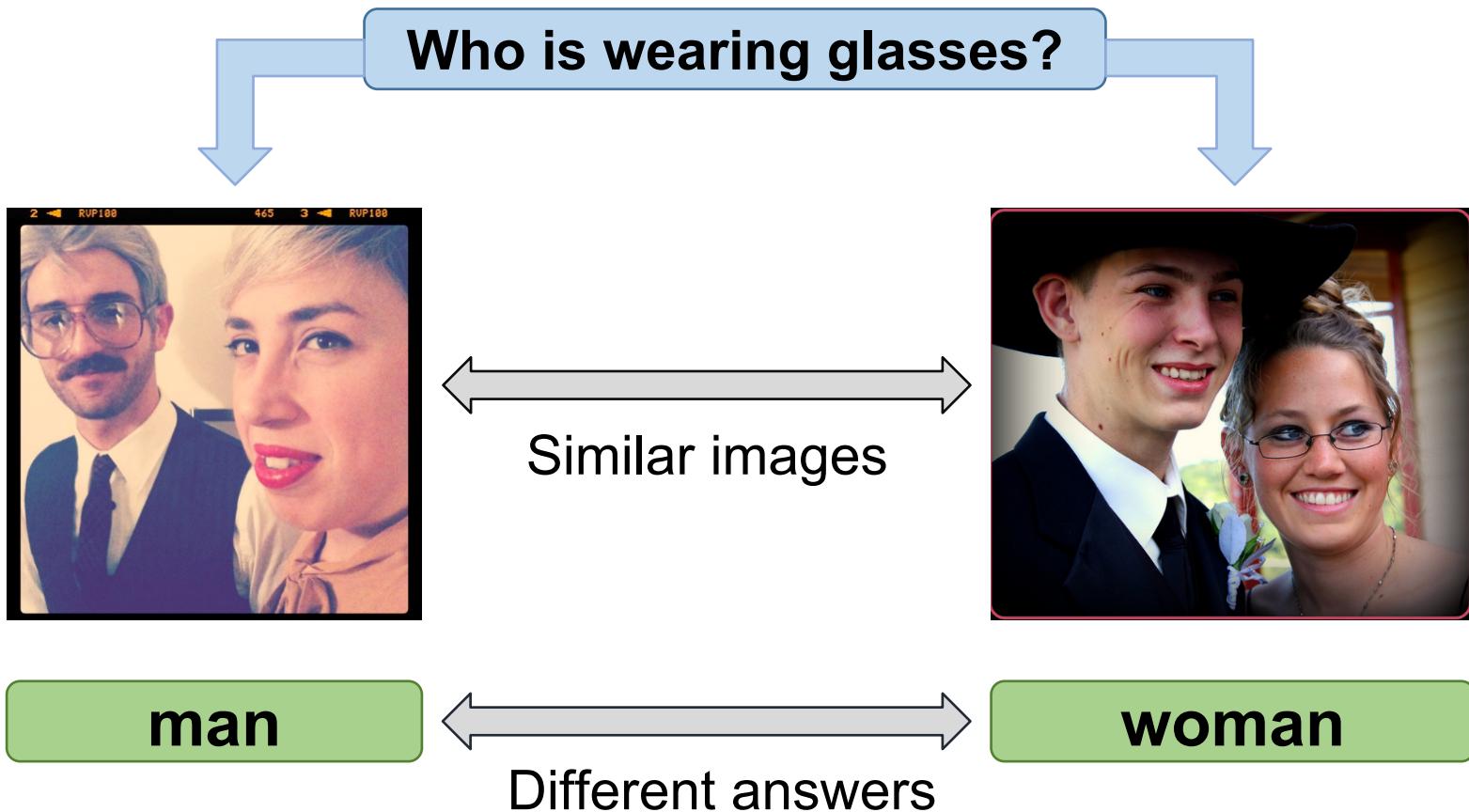
VQA Timeline



Geman et al., PNAS 14
Malinowski & Fritz, NIPS 14



VQA v2.0 Dataset



VQA v2.0 Dataset Stats

- >200K images

VQA v2.0 Dataset Stats

- >200K images
- >1.1M questions

VQA v2.0 Dataset Stats

- >200K images
- >1.1M questions
- >11M answers

VQA v2.0 Dataset Stats

- >200K images
- >1.1M questions
- >11M answers

1.8 x VQA v1.0

VQA Challenge 2017 on

<https://evalai.cloudcv.org/>

Featured Challenge

Explore other past, ongoing and upcoming challenges.

[View All](#)

Challenge Accuracy: 69.00



Derek Hoiem, VQA workshop

VQA VQA Real Image Challenge
(Open-Ended) 2017

Organized by: VQA Team

Recent progress in computer vision and natural language processing has demonstrated that lower-level tasks are much closer to being solved. We believe that the time is ripe to pursue higher-level tasks, one of which is Visual Question Answering (VQA), where the goal is to be able to understand the semantics of scenes well enough to be able to answer open-ended, free-form natural language questions (asked by humans) about images....

Status: In Progress [view more](#)



EvalAI: Evaluating state of the art in AI

Open source platform that helps researchers to create, collaborate and participate in various AI challenges

The screenshot shows the VQA Challenge 2018 page on the EvalAI platform. At the top, there's a logo for "VQA" and the text "VQA Challenge 2018 Organized by: VQA Team". Below this, there are tabs for "Overview", "Evaluation", "Phases", "Participate", and "Leaderboard". The "Overview" tab is selected. The main content area displays several questions with image examples and multiple-choice answers. For example, one question asks "Who is wearing glasses?" with options "man" and "woman", and another asks "Where is the child sitting?" with options "fridge", "arms", and "fridge". There are also questions like "Is the umbrella upside down?" with "yes" and "no" options, and "How many children are in the bed?" with "2" and "1" options. At the bottom, there's a note about the progress in computer vision and natural language processing.

VQA Challenge 2018
Organized by: VQA Team

Overview Evaluation Phases Participate Leaderboard

VQA Challenge 2018

Who is wearing glasses?
man woman

Where is the child sitting?
fridge arms

Is the umbrella upside down?
yes no

How many children are in the bed?
2 1

Recent progress in computer vision and natural language processing has demonstrated that lower-level tasks are much closer to being solved. We believe that the time is ripe to pursue higher-level tasks, one of which is Visual Question Answering (VQA), where the goal is to be able to understand the semantics of scenes well enough to be able to answer open-ended, free-form natural language questions (asked by humans) about images.

The screenshot shows the "Who uses EvalAI?" section of the EvalAI website. At the top, there are "Sign Up" and "Log In" buttons. Below that, the heading "Who uses EvalAI?" is followed by the text "Several partner organizations already host their challenges on EvalAI:". A grid of logos for various organizations is displayed, including Facebook, Google, Carnegie Mellon University, Stanford University, Georgia Tech, Mapillary Research, Virginia Tech, UMBC, University of Pittsburgh, and Draper.

Sign Up Log In

Who uses EvalAI?

Several partner organizations already host their challenges on EvalAI:

facebook Google Carnegie Mellon University Stanford University

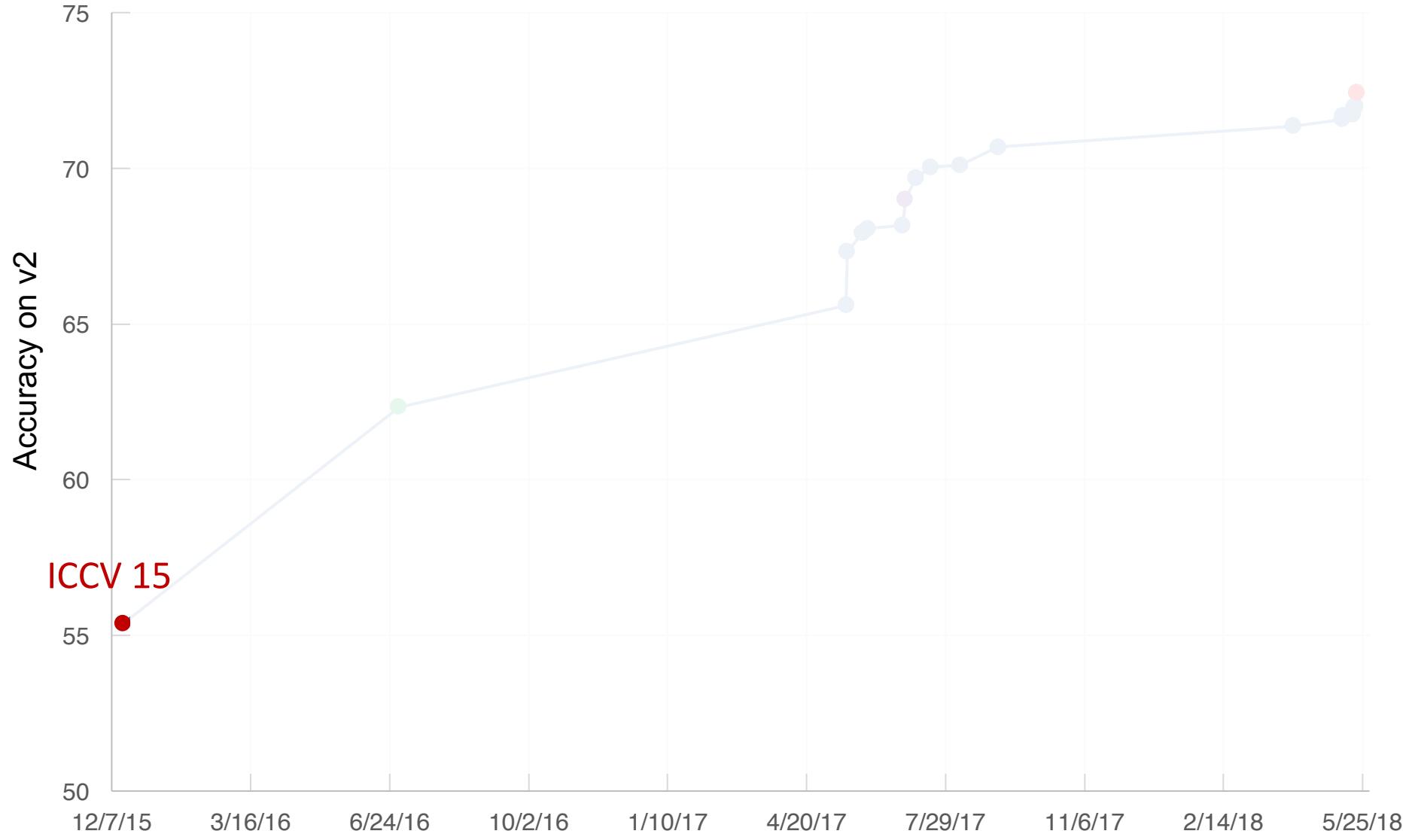
Georgia Tech Mapillary Research Virginia Tech UMBC

University of Pittsburgh Draper

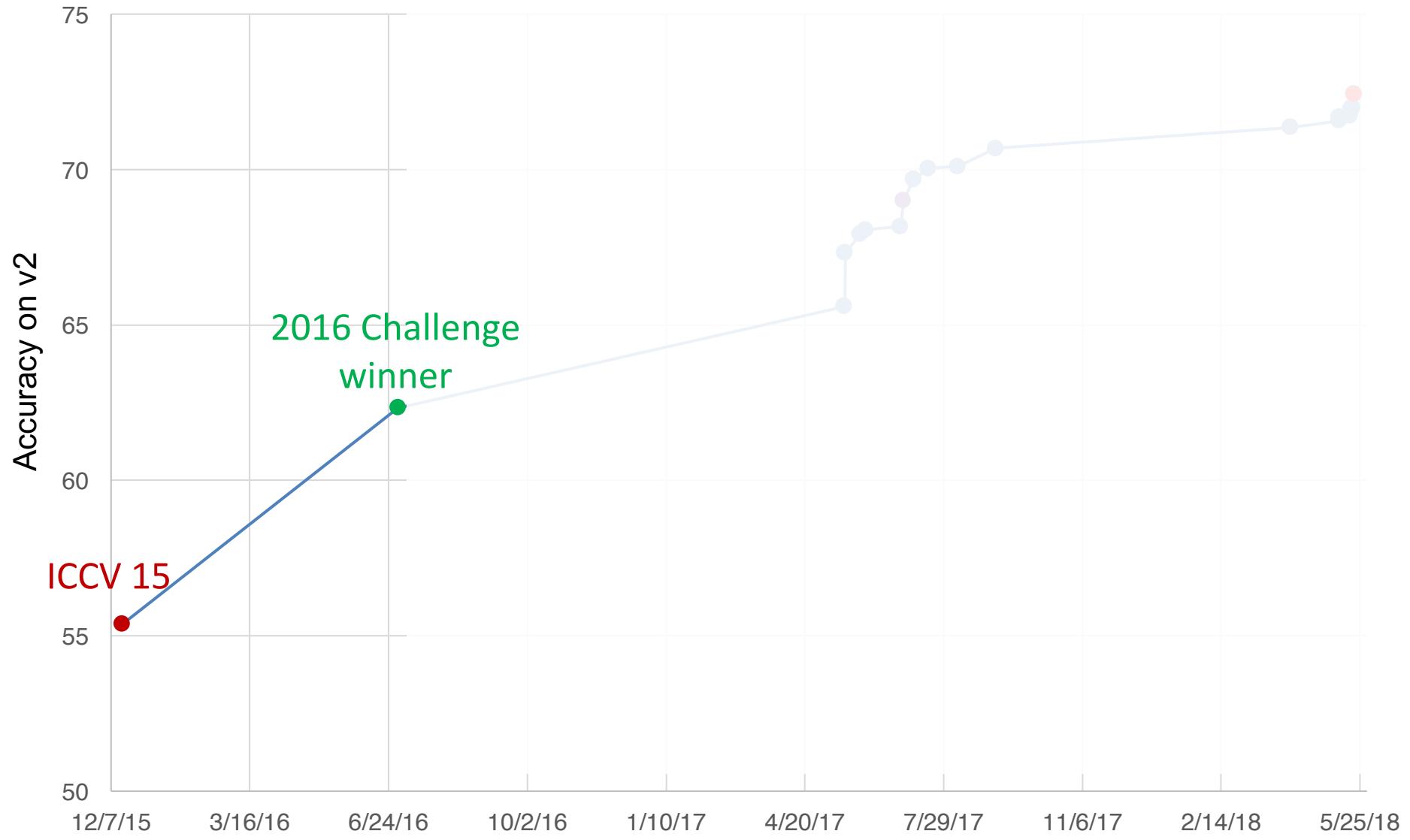
Contact: deshraj@gatech.edu

<https://evalai.cloudcv.org/>

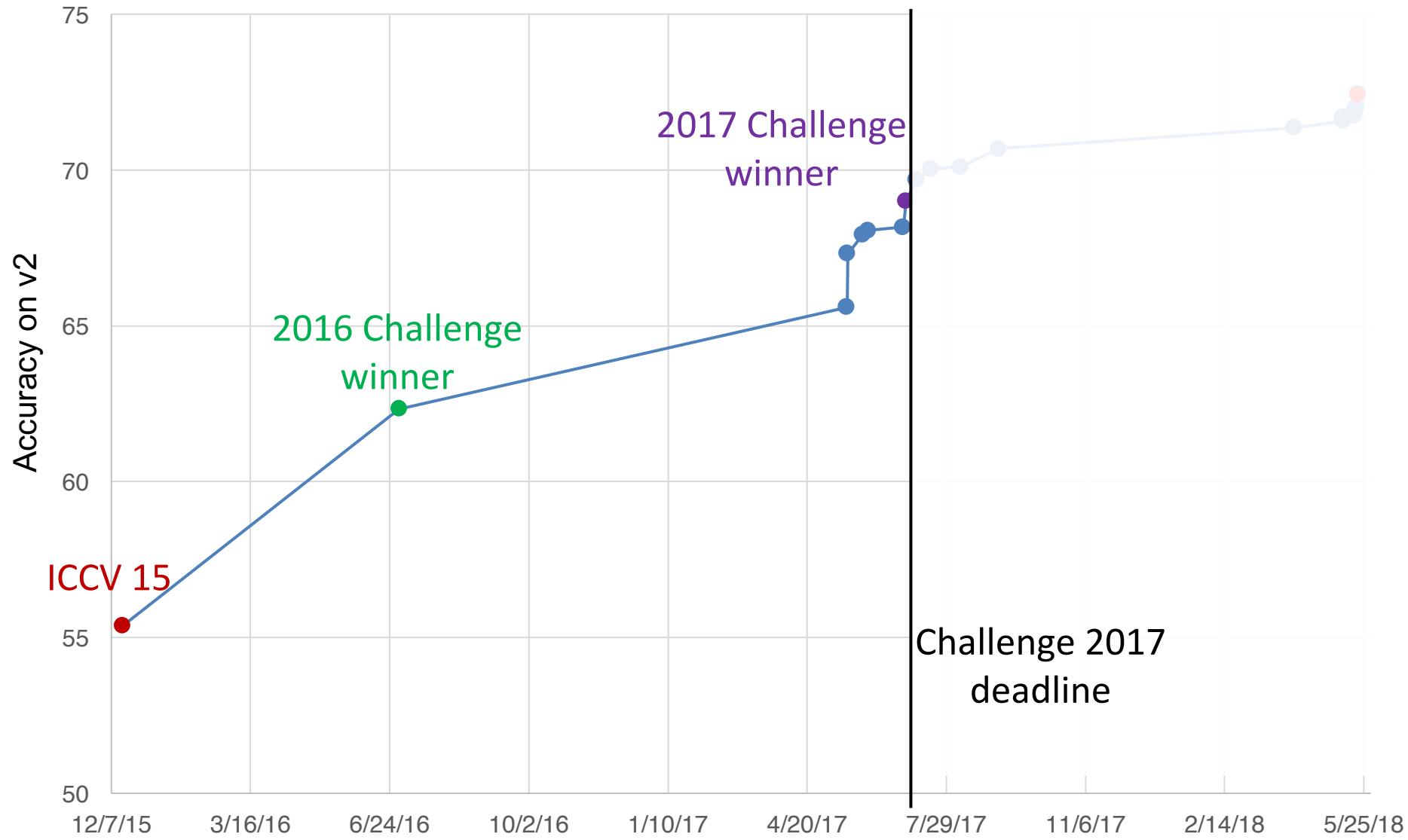
Benchmarking on VQA v2.0



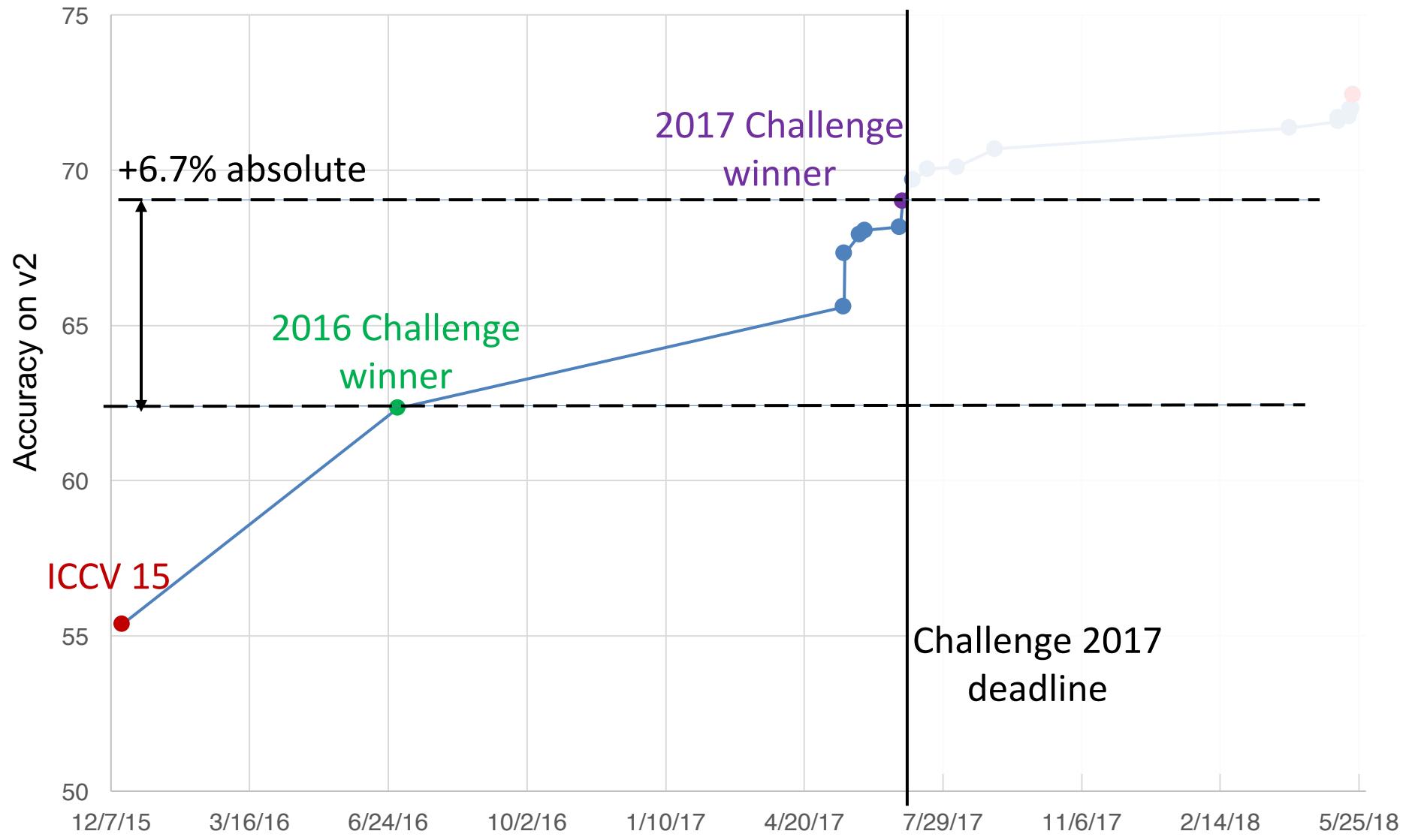
Benchmarking on VQA v2.0



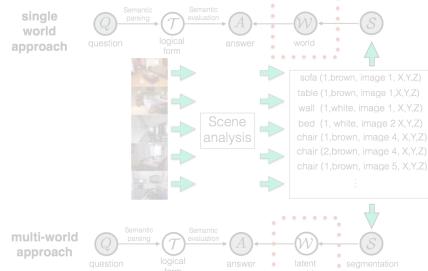
VQA Challenge 2017



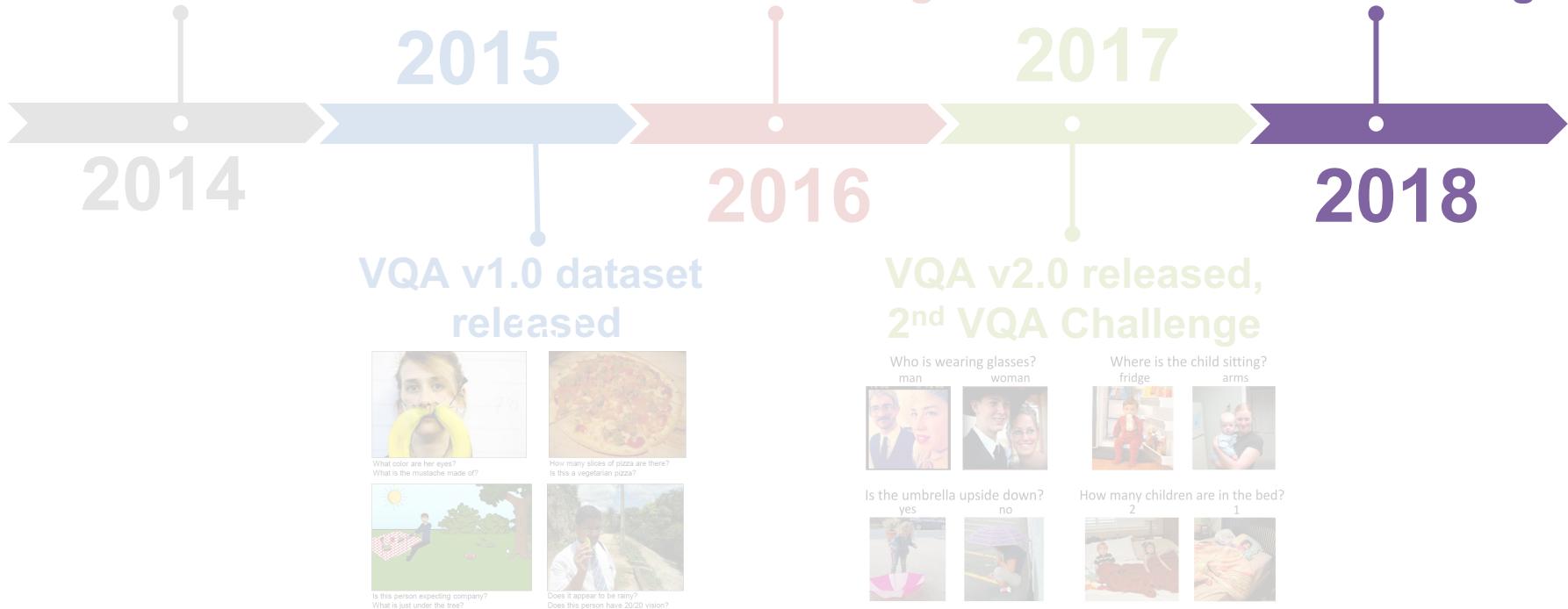
VQA Challenge 2017



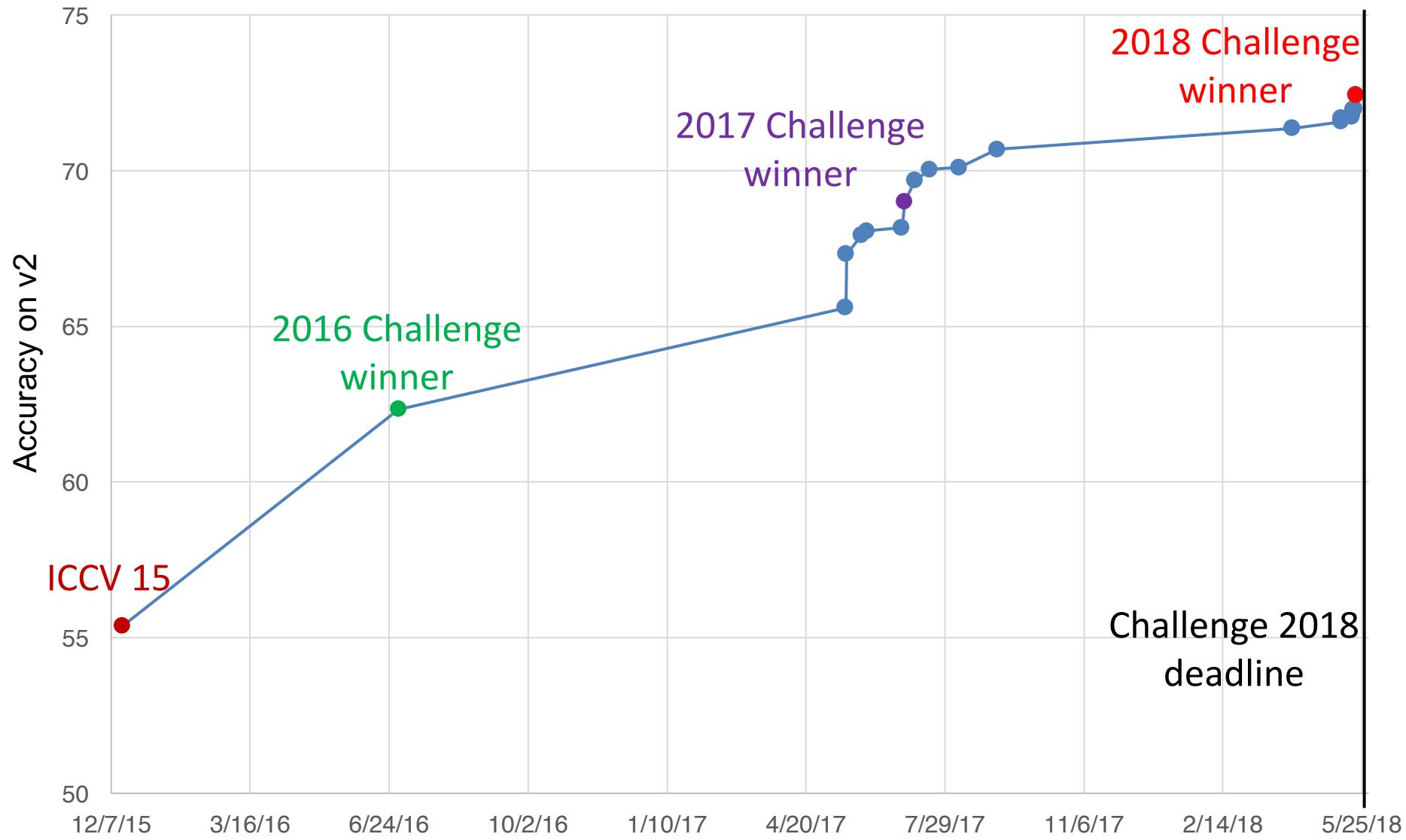
VQA Timeline



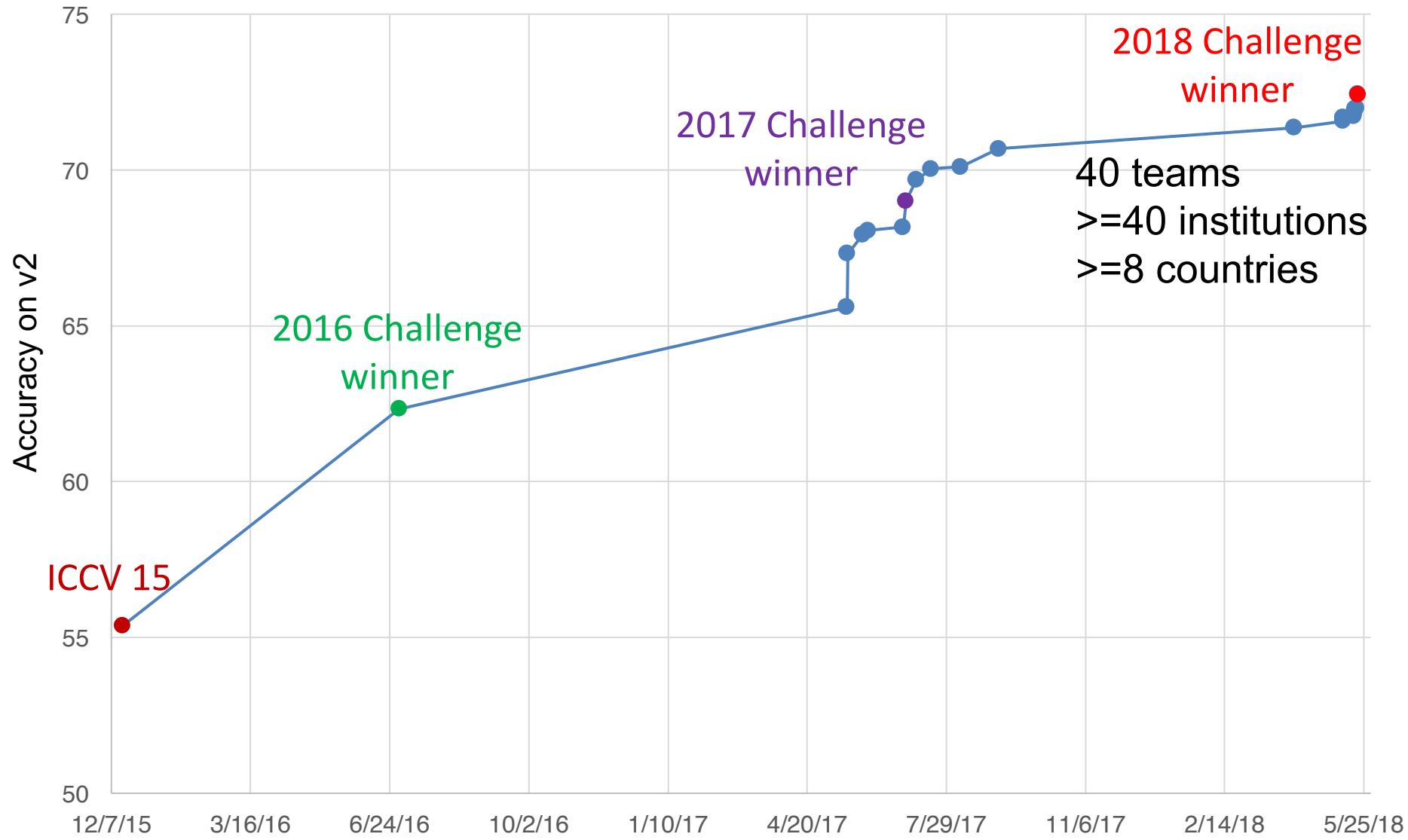
Geman et al., PNAS 14
Malinowski & Fritz, NIPS 14



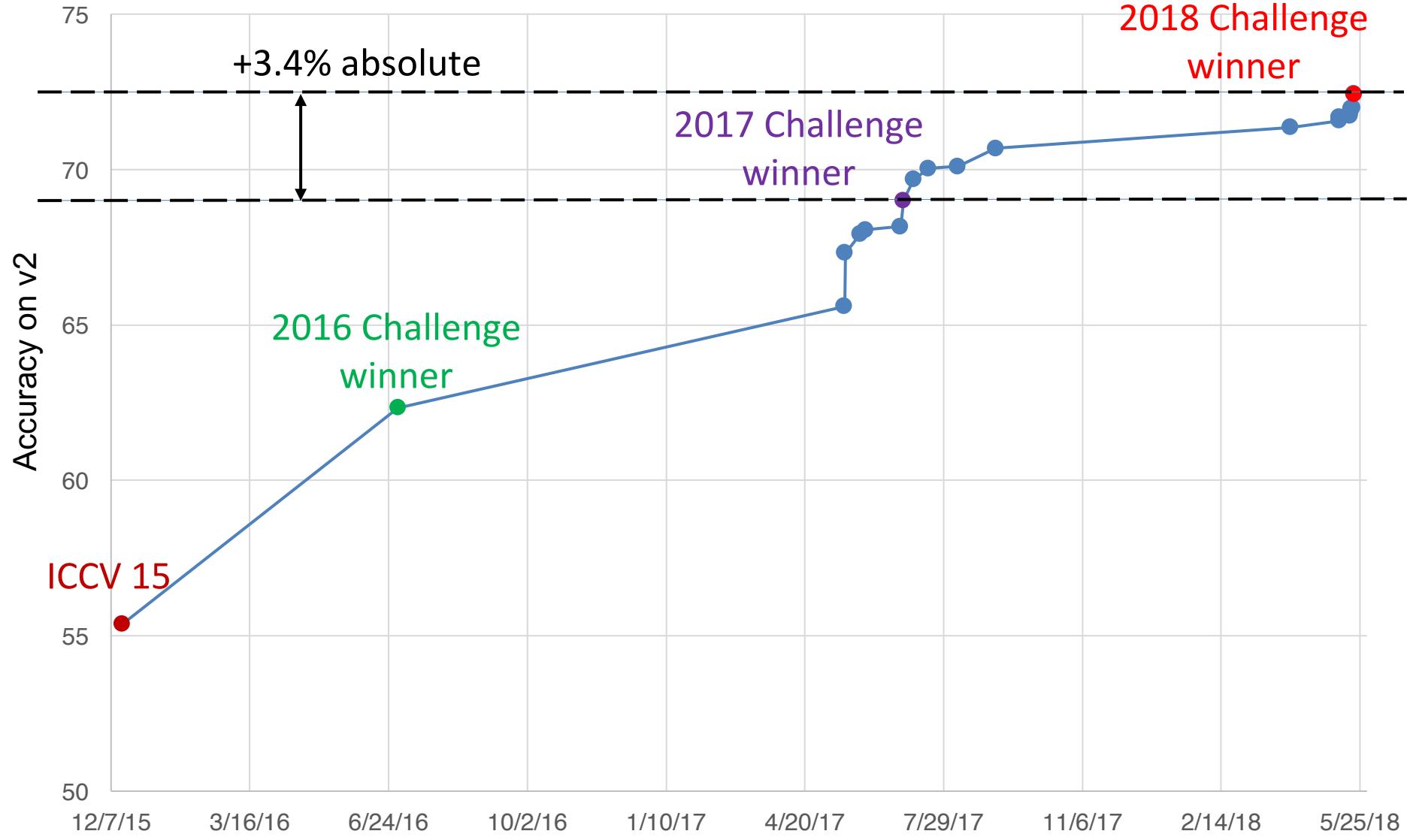
VQA Challenge 2018



VQA Challenge 2018



VQA Challenge 2018

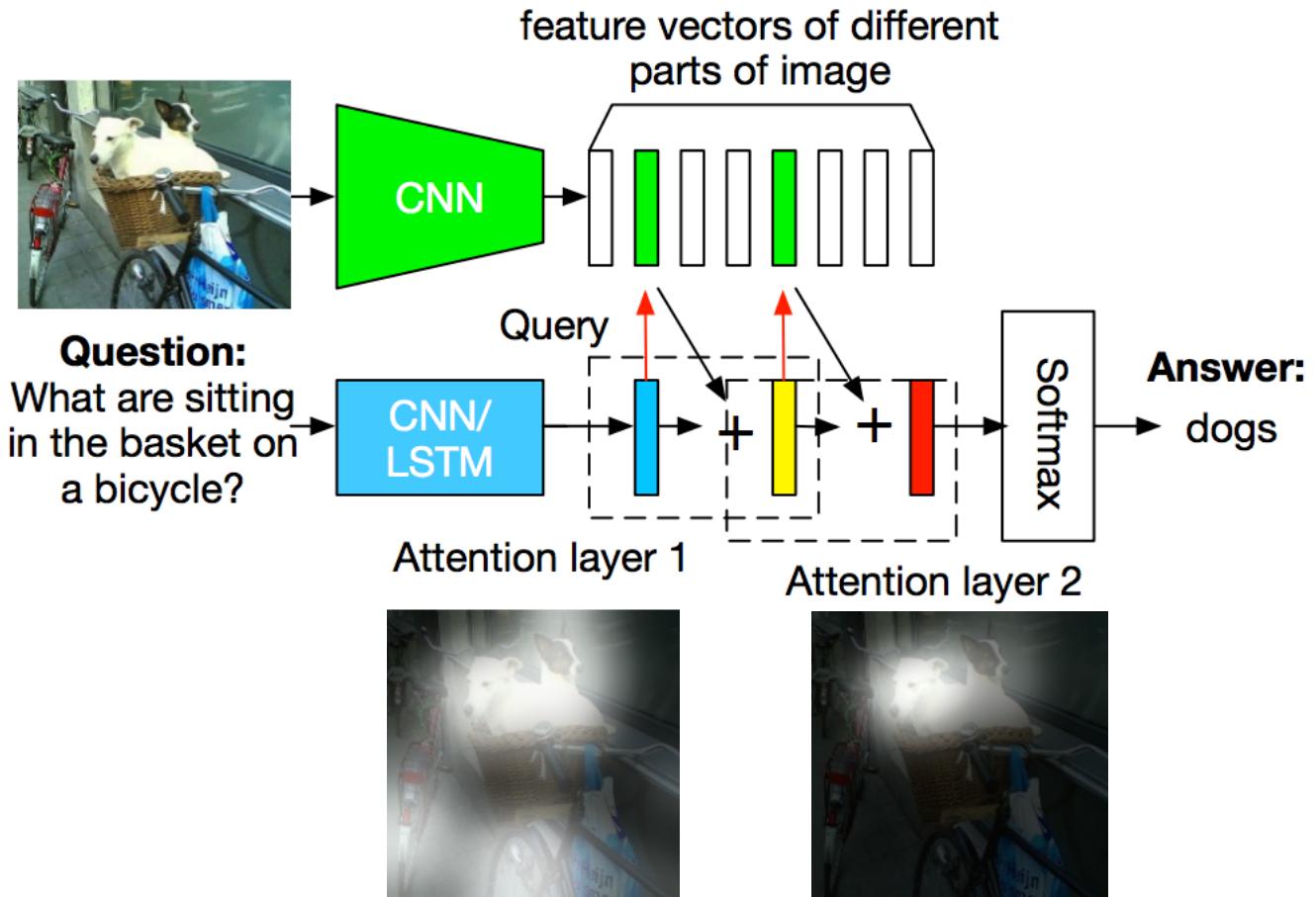


What has the community gained?

What has the community gained?

- Attention Networks

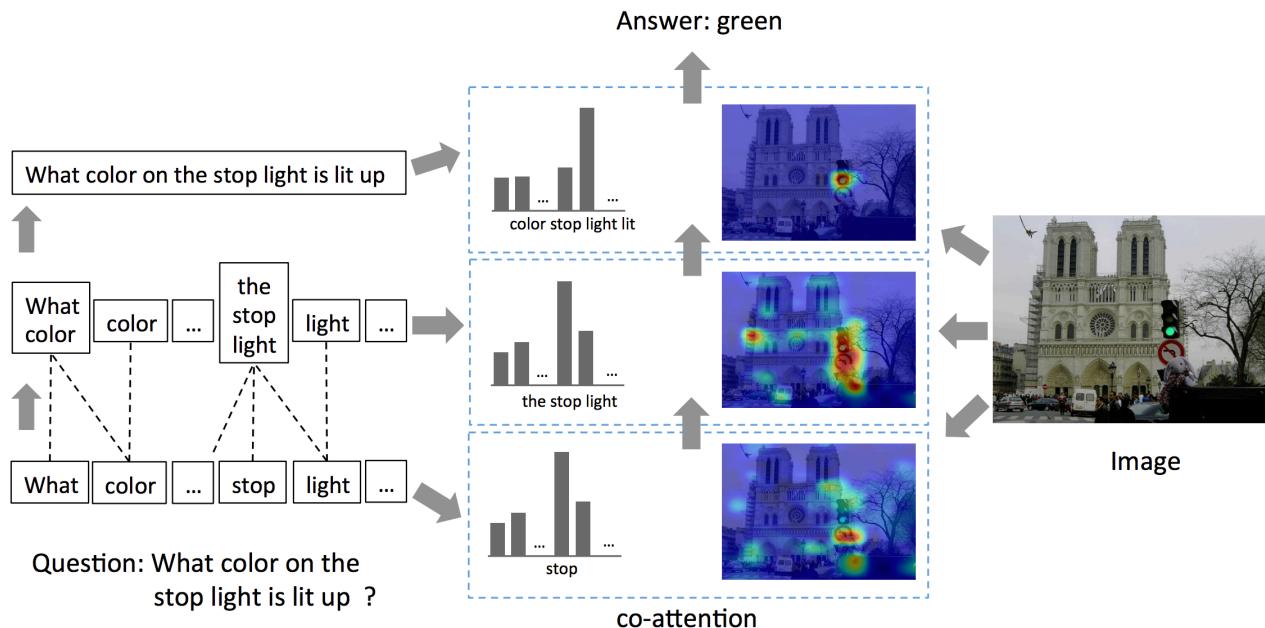
Stacked Attention Networks. Yang et al., CVPR 16



What has the community gained?

- Attention Networks
- Hierarchical Co-attention

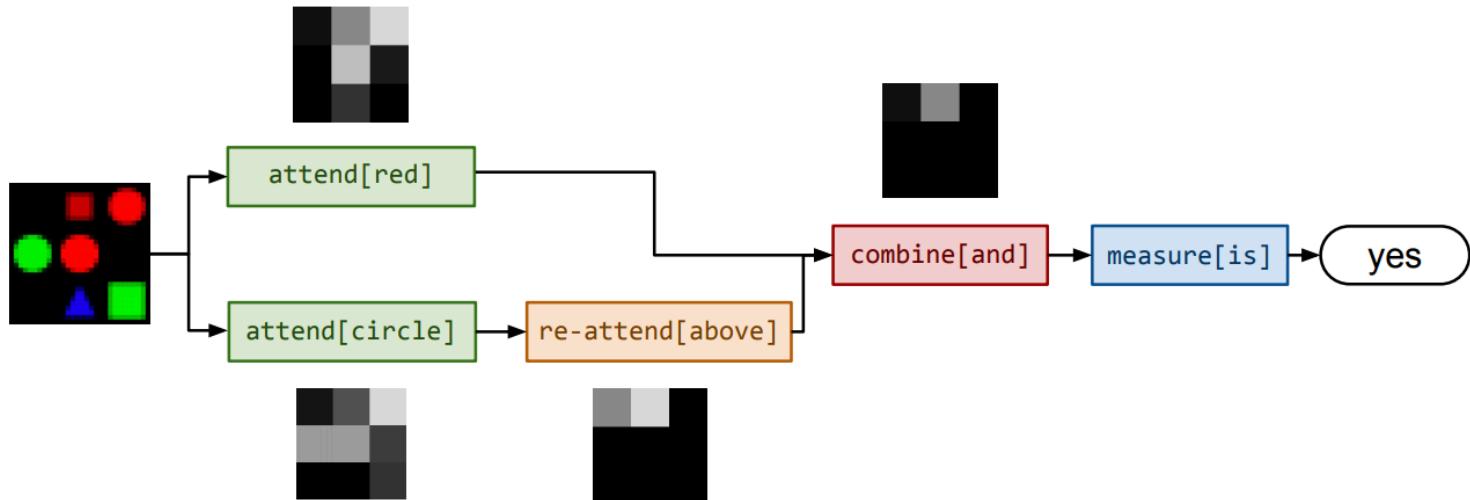
Hierarchical Question-Image Co-Attention Lu et al., NIPS 16



What has the community gained?

- Attention Networks
- Compositionality
- Hierarchical Co-attention

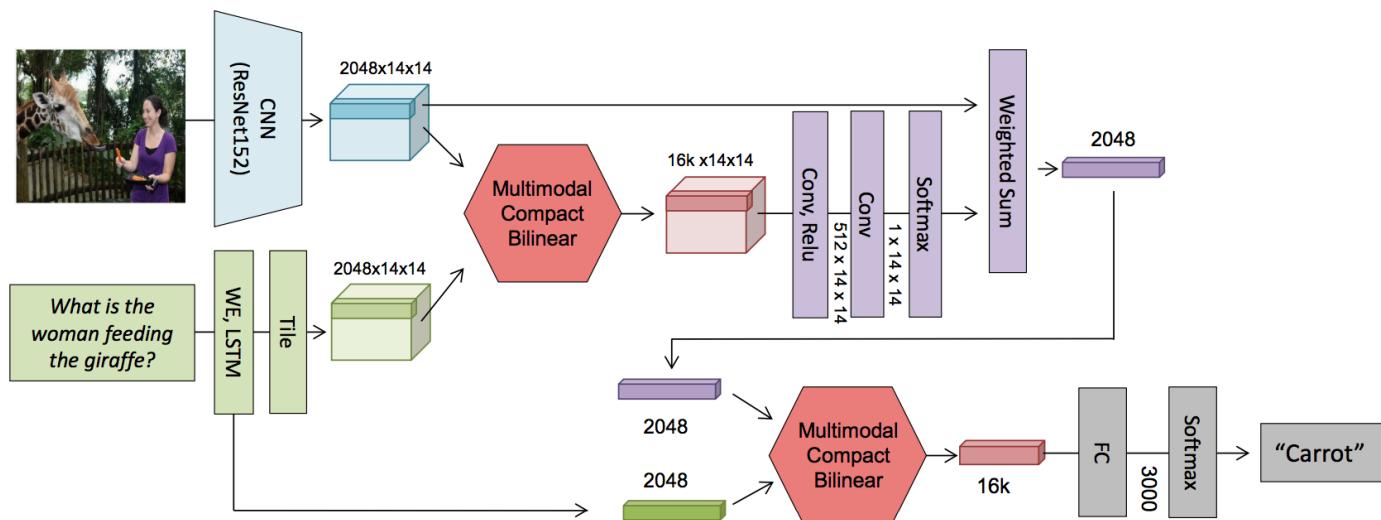
Neural Module Networks
Andreas et al., CVPR 16



What has the community gained?

- Attention Networks
- Compositionality
- Hierarchical Co-attention
- Multi-modal Pooling

Multimodal Compact Bilinear Pooling Fukui et al., EMNLP 16

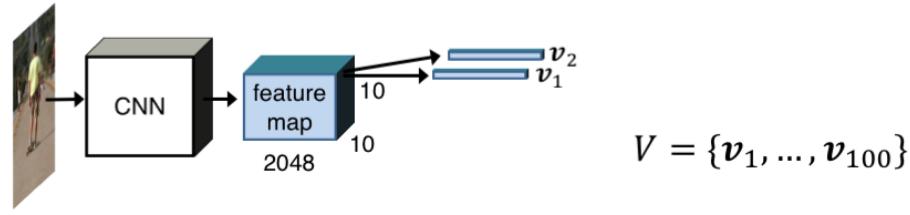


What has the community gained?

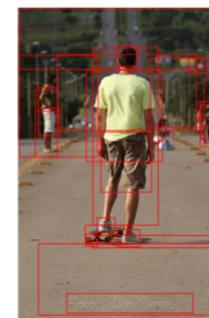
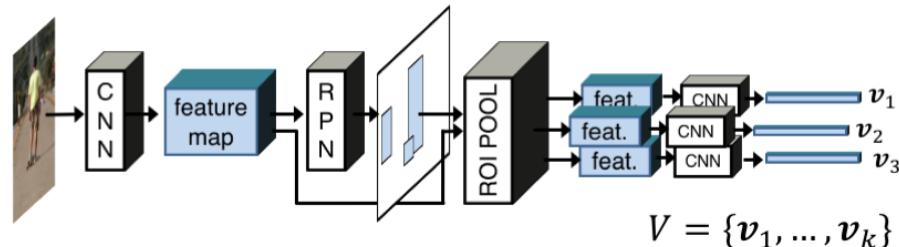
- Attention Networks
- Compositionality
- Bottom-up Top-down attention
- Hierarchical Co-attention
- Multi-modal Pooling

Bottom-Up and Top-Down Attention Anderson et al., CVPR 18

Spatial output of a CNN:



Bottom-up attention (using Faster R-CNN):



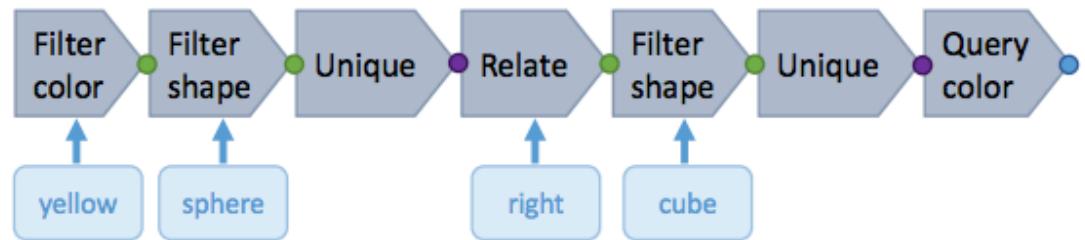
What has the community gained?

- Attention Networks
- Compositionality
- Bottom-up Top-down attention
- Hierarchical Co-attention
- Multi-modal Pooling
- Visual Reasoning

CLEVR
Johnson et al., CVPR 17



Sample chain-structured question:



What color is the cube to the right of the yellow sphere?

Visual Dialog Challenge 2018

Visual Dialog

A screenshot of the Visual Dialog challenge interface. At the top, there's a blue header bar with the title "Visual Dialog". Below it is a photograph of a brown tabby cat standing on a table, drinking water from a white mug. To the right of the image, a dark blue rounded rectangle contains the caption: "A cat drinking water out of a coffee mug.". The main area is a conversation log between a user and a robot. The user (represented by a small robot icon) asks: "What color is the mug?", "Are there any pictures on it?", "Is the mug and cat on a table?", and "Are there other items on the table?". The robot (also represented by a small robot icon) responds: "White and red", "No, something is there can't tell what it is", "Yes, they are", and "Yes, magazines, books, toaster and basket, and a plate". At the bottom of the interface, there are two icons: a camera and a microphone, followed by the placeholder text "Start typing question here ...".

VisDial v1.0

- ~130k images (COCO)
- 10-round dialog / image
- ~1.3 million QA pairs
- Evaluation
 - Automatic metrics
 - Human annotations

- Deadline: **mid-August, 2018**
- Results: **September 8th, 2018** at ECCV 2018

visualdialog.org/challenge/2018



Aishwarya Agrawal
Georgia Tech



Stanislaw Antol
Traptic, Inc.



Jiasen Lu
Georgia Tech



Tejas Khot
CMU



Peng Zhang
Amazon



Akrit Mohapatra
Virginia Tech



Douglas Summers-Stay
Army Research Lab



Meg Mitchell
Google Research



Larry Zitnick
FAIR

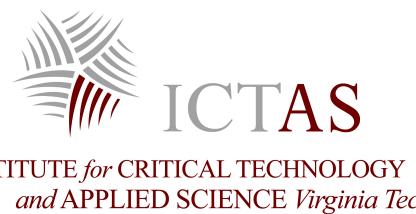


Dhruv Batra
Georgia Tech / FAIR



Devi Parikh
Georgia Tech / FAIR

Sponsors



Thanks!

Questions?

Email: visualqa@gmail.com