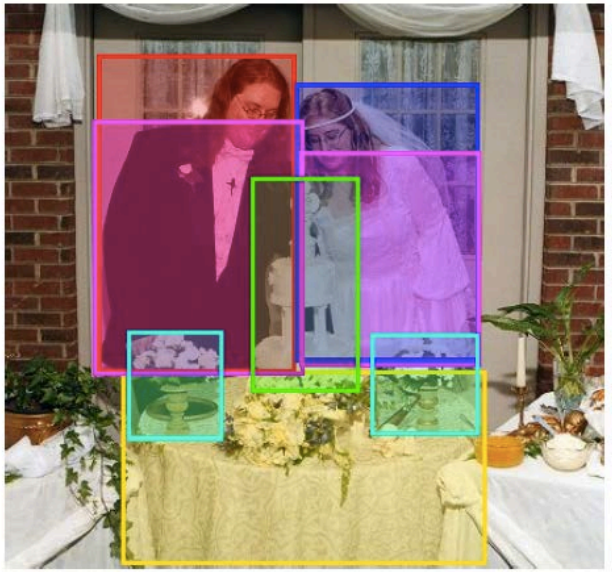
Towards Joint Understanding of Images and Language

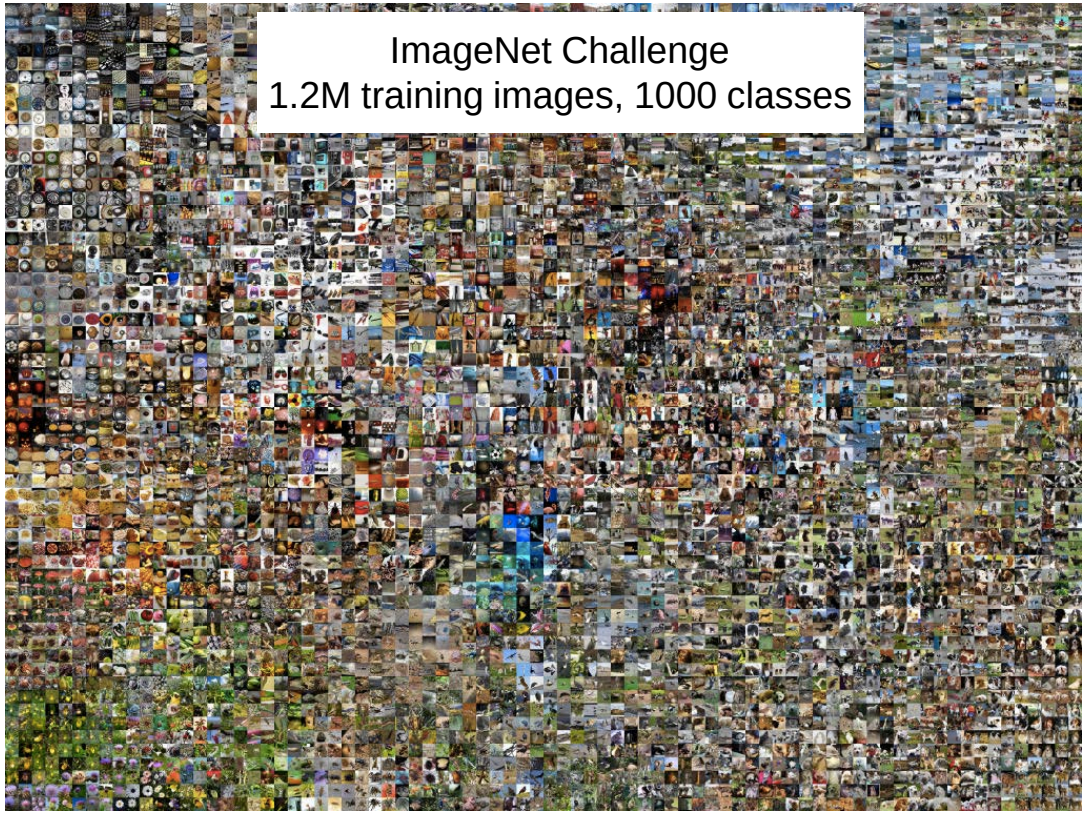
Svetlana Lazebnik

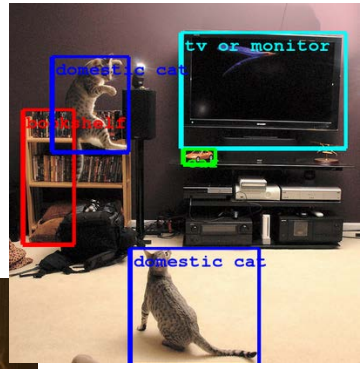
Joint work with J. Hockenmaier,
B. Plummer, L. Wang,
C. Cervantes, J. Caicedo,
Y. Gong, M. Hodosh

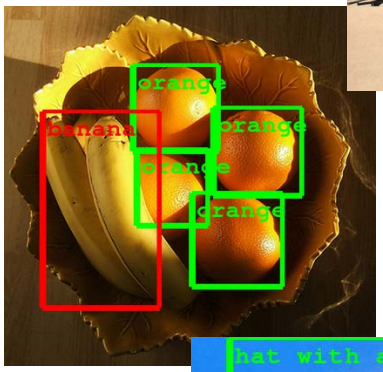


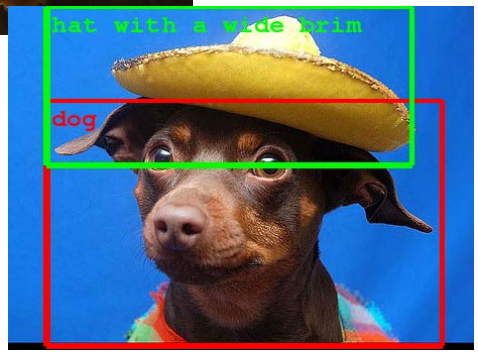
A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Big data and deep learning “solved” image classification









[Computer Eyesight Gets a Lot More Accurate](#)
NY Times Bits blog, August 18, 2014

Next frontier: Image description

A group of young people playing a game of Frisbee

A person riding a motorcycle on a dirt road

Vinyals et al., CVPR 2015

http://www.nytimes.com/2014/11/18/science/researchers-announce-breakthrough-in-content-recognition-software.html

Datasets for image description

- **Flickr30K** (Young et al., 2014): 32K images, five captions per image
- **MSCOCO** (Lin et al., 2014): 100K images, five captions per image

A goalie in a hockey game dives to catch a puck as the opposing team charges towards the goal.
The white team hits the puck, but the goalie from the purple team makes the save.
Picture of hockey team while goal is being scored.
Two teams of hockey players playing a game.
A hockey game is going on.

A group of people are getting fountain drinks at a convenience store.
Several adults are filling their cups and a drink machine.
Two guys getting a drink at a store counter.
Two boys in front of a soda machine.
People get their slushies.

Evaluating image description as ranking

Two boys are playing football.

People in a line holding lit roman candles..

A little girl is enjoying the swings.

A motorbike is racing around a track.

A boy in a yellow uniform.

An elephant is being washed.

Image-to-sentence search: Given a pool of images and captions, rank the captions for each image

[Hodosh, Young, Hockenmaier, 2013]

Evaluating image description as ranking

Two boys are playing football.

People in a line holding lit roman candles..

A little girl is enjoying the swings.

A motorbike is racing around a track.

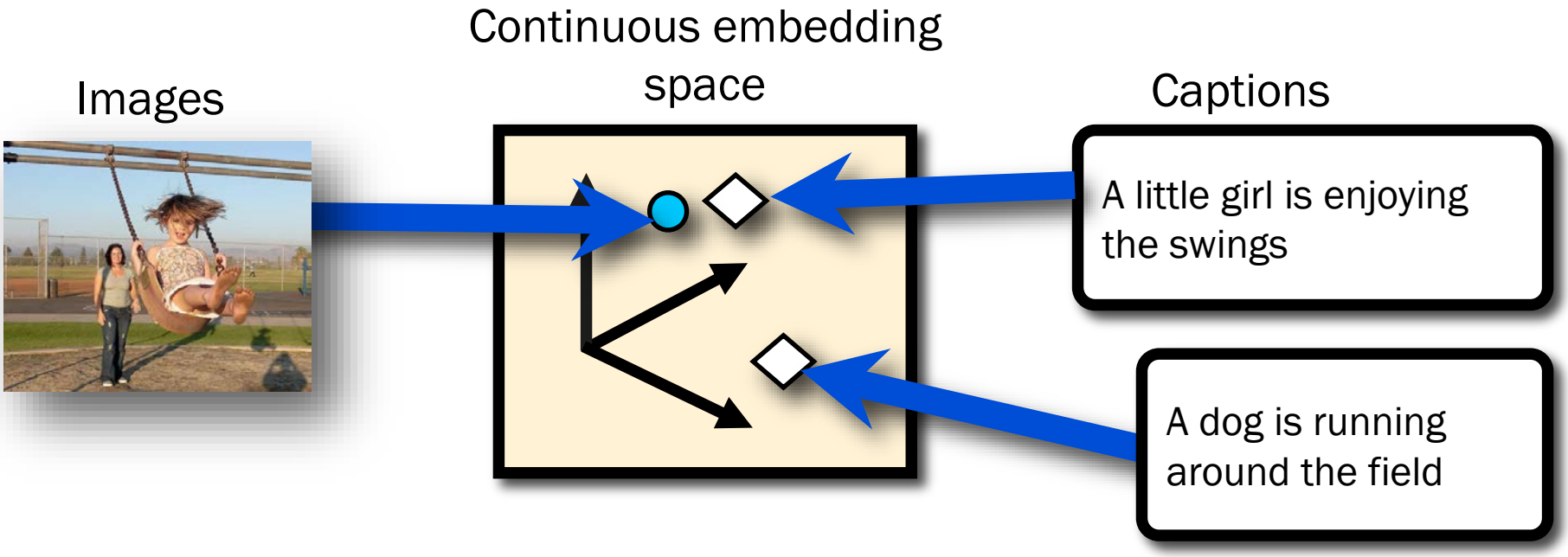
A boy in a yellow uniform.

An elephant is being washed.

Sentence-to-image search: Given a pool of images and captions, rank the captions for each image

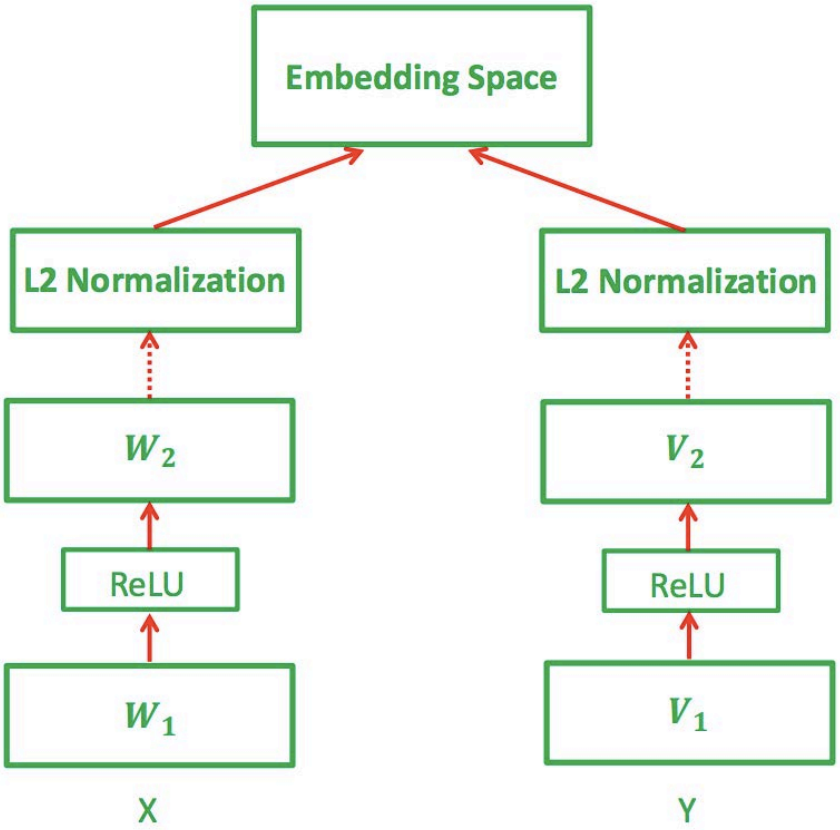
[Hodosh, Young, Hockenmaier, 2013]

A joint embedding space for images and text



- Use Canonical Correlation Analysis (CCA) to project images and text to a joint latent space (Hodosh, Young, and Hockenmaier, 2013; Gong, Ke, Isard, and Lazebnik, 2014)

Deep image-text embeddings



Images

Text

Wang, Li and Lazebnik, CVPR 16

Deep image-text embeddings

	Image-to-sentence			Sentence-to-image		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
Karpathy & Fei-Fei 2015 AlexNet + BRNN	22.2	48.2	61.4	15.2	37.7	50.5
Mao et al. 2015 VGGNet + mRNN	35.4	63.8	73.7	22.8	50.7	63.1
Klein et al. 2015 VGGNet + CCA	35.0	62.0	73.8	25.0	52.7	66.0
Wang et al. 2015 VGGNet + deep embed.	**40.3**	**68.9**	**79.9**	**29.7**	**60.1**	**72.1**

Wang, Li and Lazebnik, CVPR 16

Beyond global representations

- Flickr30K Entities dataset (Plummer, Wang, Cervantes, Caicedo, Hockenmaier, Lazebnik, ICCV 2015)

A man with pierced ears is wearing glasses and an orange hat.

A man with glasses is wearing a beer can crocheted hat.

A man with gauges and glasses is wearing a Blitz hat.

A man in an orange hat starring at something.

A man wears an orange hat and glasses.

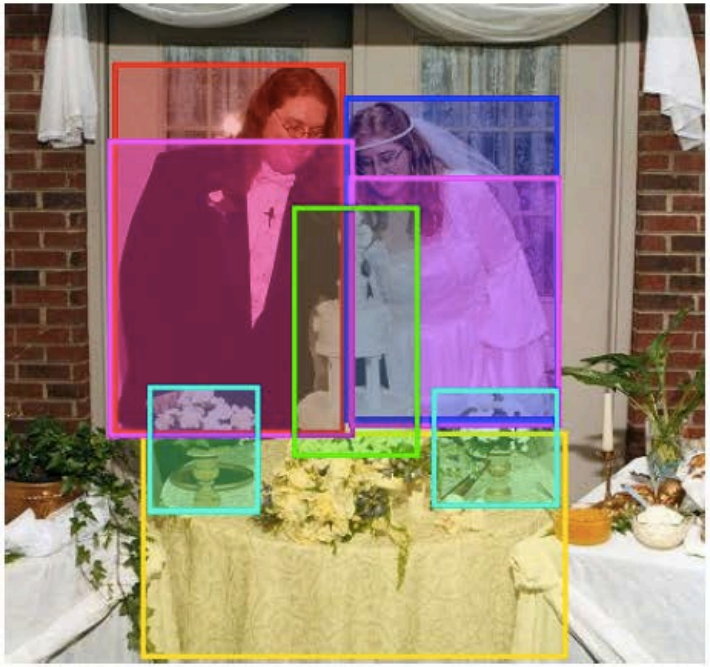
Bounding boxes for all mentioned entities

Coreference chains for all mentions of the same set of entities

Flickr30K Entities Dataset

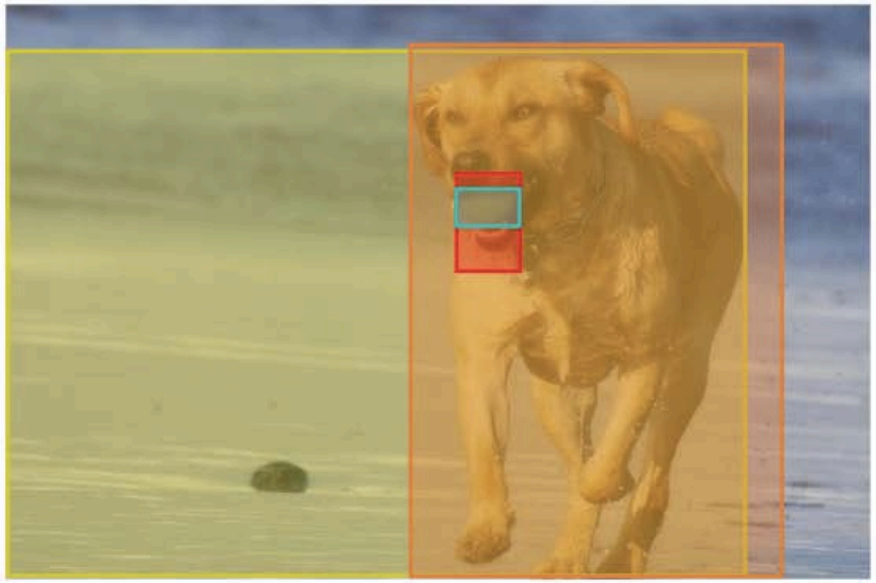
- 244K coreference chains, 267K bounding boxes

During **a gay pride parade** in **an Asian city**, **some people** hold up **rainbow flags** to show their **support**.
A group of youths march down **a street** waving **flags** showing **a color spectrum**.
Oriental people with **rainbow flags** walking down **a city street**.
A group of people walk down **a street** waving **rainbow flags**.
People are **outside** waving **flags** .

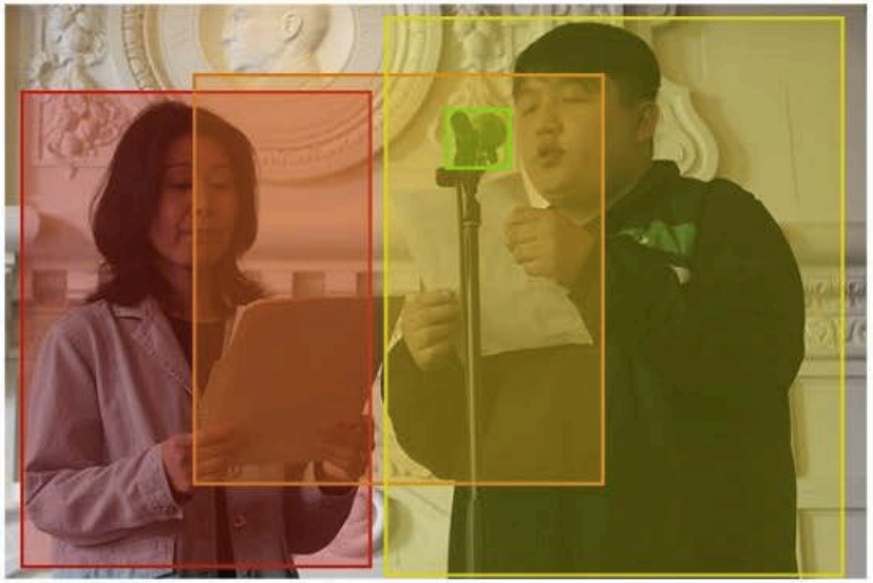


A couple in **their wedding attire** stand behind **a table** with **a wedding cake** and **flowers**.
A bride and **groom** are standing in front of **their wedding cake** at **their reception**.
A bride and **groom** smile as **they** view **their wedding cake** at **a reception**.
A couple stands behind **their wedding cake**.
Man and **woman** cutting **wedding cake**.

A new task: Phrase localization



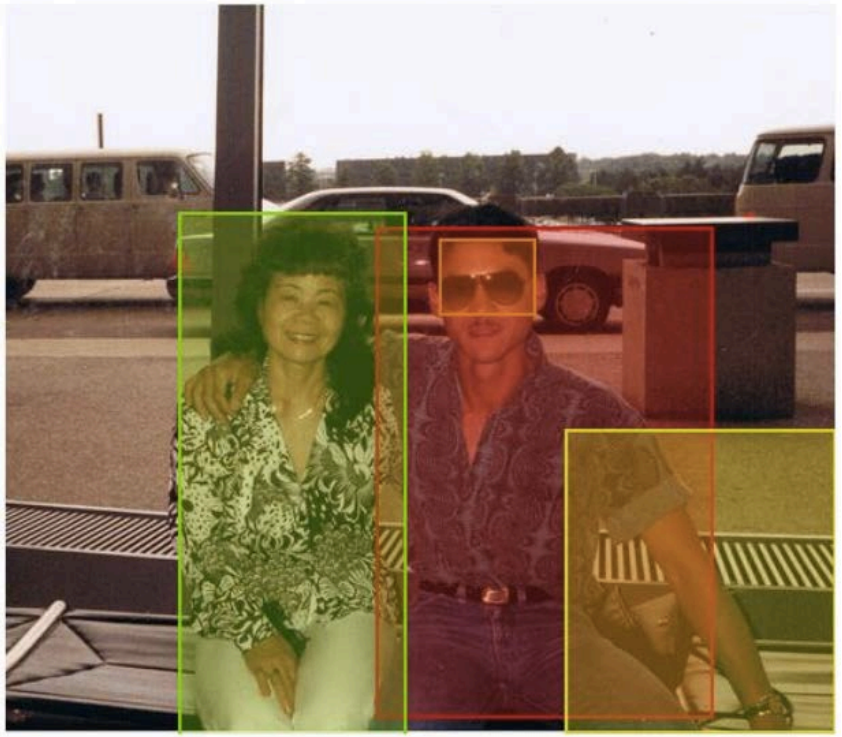
The yellow dog [0.33] walks on the beach [0.74] with a tennis ball [0.66] in its mouth [0.79].



A dark-haired woman [0.40] is looking at papers [0.89] standing next to a dark-haired man [0.39] speaking into a microphone [0.79].

Phrase localization is hard!

A man [0.49] in a gray sweater [0.73] speaks to two women [0.70] and a man [0.49] pushing a shopping cart [0.49] through Walmart [0.79].

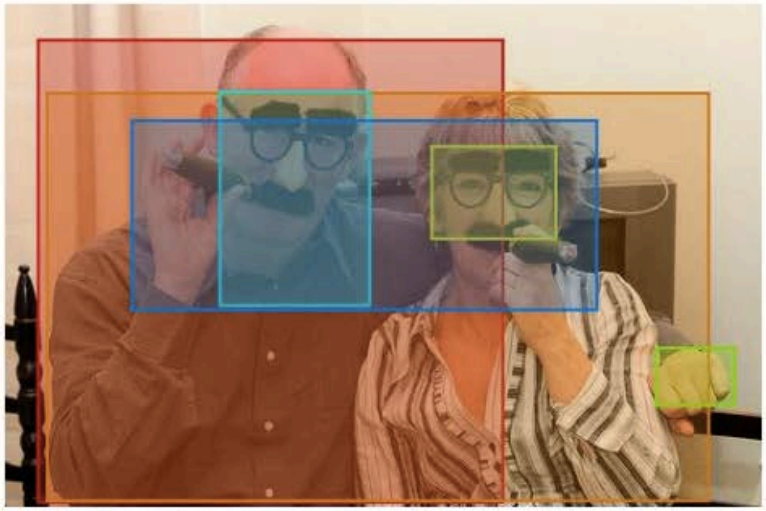


A man [0.39] in sunglasses [0.39] puts his arm [0.85] around a woman [0.38].

Phrase localization is hard!

- Improving image description using phrase localization is even harder

Ground truth sentence



A man [0.38] and a woman [0.39] wearing costume glasses [0.75] (with attached eyebrows [0.79], nose [0.85], and moustache [0.74]) and holding cigars [0.77].

Top retrieved sentence

A man [0.38] in a striped shirt [0.71] and glasses [0.48] speaks into a microphone [0.72].

So, are we done?

- Learning to associate images with simple captions seems to be a much easier task than we might have thought a few years ago.
- But we’re fooling ourselves if we think our systems ‘understand’ images or sentences.
- We need datasets and models that encode a wider variety of visual cues and reveal the compositional nature of images and language.