A picture is worth 13.6 words (on average)
An on-paper experiment

Write a caption for this image, one sentence in length.

(In English.)
People write weird captions

Another dream car to add to the list, this one spotted in Hanbury St.
People write weird captions

Another dream car to add to the list, this one spotted in Hanbury St.

Shot out my car window while stuck in traffic because people in Cincinatti can't drive in the rain
People write weird captions

1. A distorted photo of a man cutting up a large cut of meat in a garage.

2. A man smiling at the camera while carving up meat.

3. A man smiling while he cuts up a piece of meat.

4. A smiling man is standing next to a table dressing a piece of venison.

5. The man is smiling into the camera as he cuts meat.
What I used to think vision did...
What I used to think vision did...
What I used to think vision did...
What I used to think vision did...
Now I know better....
Detecting on a large scale...

bird

boat

bottle

bowl
What do people describe?

1) Given an image
What do people describe?

1) Given an image

Predict what people will describe

“two women sitting brunette blonde on bench reading magazine”
What do people describe?

1) Given an image

“two women sitting brunette blonde on bench reading magazine”

Predict what people will describe

- women
- bench
- magazine
- grass
- skirt
...
Predicting what will be described

What’s in this image?
Predicting what will be described

What’s in this image?

- man
- baby
- sling
- ladder
- fridge
- table
- watermelon
- chair
- boxes
- cups
- water bottle
- wall
- pacifier
- beard
- glasses
- shirt
- ...

A picture is worth 13.6 words
Predicting what will be described

What’s in this image?
- man
- baby
- sling
- ladder
- fridge
- table
- watermelon
- chair
- boxes
- cups
- water bottle
- wall
- pacifier
- beard
- glasses
- shirt
- ...

What do people describe?
“A bearded man is holding a child in a sling.”
Predicting what will be described

What’s in this image?
- man
- baby
- sling
- ladder
- fridge
- table
- watermelon
- chair
- boxes
- cups
- water bottle
- wall
- pacifier
- beard
- glasses
- shirt
- ...

What do people describe?
“A bearded man is holding a child in a sling.”
“A bearded man stands while holding a small child in a green sheet.”
“A bearded man with a baby in a sling poses.”
“Man standing in kitchen with little girl in green sack.”
“Man with beard and baby”
Predicting what will be described

What’s in this image?
- man
- baby
- sling
- ladder
- fridge
- table
- watermelon
- chair
- boxes
- cups
- water bottle
- wall
- pacifier
- beard
- glasses
- shirt
- ...

What do people describe?
"A bearded man is holding a child in a sling."
"A bearded man stands while holding a small child in a green sheet."
"A bearded man with a baby in a sling poses."
"Man standing in kitchen with little girl in green sack."
"Man with beard and baby"
Description factors

What factors influence what someone will describe about an image?

Two kinds of factors
  - Compositional
  - Semantic
Compositional factors

“A sail boat on the ocean.”

Size/Saliency

Location
Compositional factors

“Two men standing on beach.”

Size/Saliency

Location
Semantic factors

“girl in the street”
Semantic factors

“kitchen in house”
Semantic factors

“elephant in the beach”
Semantic factors

“A tree in water and a boy with a beard”
Using large corpora to compose natural captions

(why write your own material when you can just “steal” it?)
Composing captions

a) monkey playing in the tree canopy, Monte Verde in the rain forest

b) capuchin monkey in front of my window

c) monkey spotted in Apenheul Netherlands under the tree

d) a white-faced or capuchin in the tree in the garden

e) the monkey sitting in a tree, posing for his picture
Composing captions

a) monkey playing in the tree canopy, Monte Verde in the rain forest

b) capuchin monkey in front of my window

c) monkey spotted in Apenheul Netherlands under the tree

d) a white-faced or capuchin in the tree in the garden

e) the monkey sitting in a tree, posing for his picture
Captioning with (some) evidence

Caption images where:

We assume some evidence for 1 object

&

Object detector is confident
Captioning with (some) evidence

Caption images where:

- We assume some evidence for 1 object

  &

- Object detector is confident

  Tag: “mare”

  Evidence for horse
Captioning with (some) evidence

Caption images where:

We assume some evidence for 1 object

&

Object detector is confident

Tag: “mare”
Evidence for horse

High detection score
Grab phrases based on image similarity between query and captioned data base

Object detection similarity - NPs, VPs

Stuff detection similarity – PPs

Scene similarity - PPs

Mash phrases

Compose descriptions using simple rule based concatenation
Getting NPs – Objects

Detect: fruit
Getting NPs – Objects

Detect: fruit

Find matching fruit detections by **color** similarity
Getting NPs – Objects

Detect: fruit

Find matching fruit detections by **color** similarity

Tray of glace fruit in the market at Nice, France

Fresh fruit in the market

A box of oranges was just catching the sun, bringing out detail in the skin.

The street market in Santanyi, Mallorca is a must for the oranges and local crafts.

mandarin oranges in glass bowl

An orange tree in the backyard of the house.
Getting NPs – Objects

The muddy elephant
An elephant
small elephant
A very large and seemingly old elephant
musk male elephant
African elephant
the temple elephant

Fushia flower
a flower
a pink zinna
flower
This beautiful
flower
a roman pink
flower
a tiny pink flower
pink bursting
flowers
a perfectly pink
gerbera daisy

a lonesome duck
a native new zealand duck
The duck
male Mallard duck
several other ducks
a so-called navigation duck
this duck
a duck
duck
mandarin duck
Getting VPs – objects

Detect: cow

Find matching cow detections by shape/pose similarity

theses cows live in the field behind my house

The cow was more interested in eating than looking at me with a camera!

A cow eating flowers in the south of the Netherlands.

While cycling north on Tremaine Road near Milton, this cow gazed across the road intently.
Getting PPs – stuff

Detect: grass

Find matching grass detections by color similarity

- green manure in the veg field - Plaw Hatch
- I am happy in a field of green Maryland grass
- Sheep in a field spotted during a coastal drive from Tramore to Dungarvon
- Found on hawthorn in boggy grass field
Extract scene descriptor

Find matching images by scene similarity

Pedestrian street in the Old Lyon with stairs to climb up the hill of fourviere

I'm about to blow the building across the street over with my massive lung power.

View from our B&B in this photo

Only in Paris will you find a bottle of wine on a table outside a bookstore

Getting PPs – scenes
Composing captions
Composing captions

object color
object pose
scene
stuff
Composing captions

**NP:** the sheep

**VP:** meandered along a desolate road

**PP:** in the highlands of Scotland

**PP:** through frozen grass
A picture is worth 13.6 words

Various composition patterns:
NP VP
NP PP_stuff
NP PP_scene
...
NP VP PP_scene PP_stuff
Composing captions

Various composition patterns:
NP VP
NP PP_stuff
NP PP_scene
...
NP VP PP_scene PP_stuff

the sheep meandered along a desolate road in the highlands of Scotland through frozen grass
A duck was having a bath in the harbor at Whitehaven, Cumbria, England in the water near Camley St.

A female Monarch butterfly was visiting the plant in my front yard in Devon 17/10/10.

Stained glass window depicting Christ and numerous saints in Washington National Cathedral in the Eglise.

A double-decker bus under some spreading shade trees.

Her flower girl dress designed by Mainbocher in the house.

A cat enjoys hiding under the tree.

Good results
Not so good results
Not so good results

Language issues

A Moo cow tied up around the city eating grass in various places under the tree at the young tree

male tiger sighting in twelve months of a street
Not so good results

Language issues

A Moo cow tied up around the city eating grass in various places under the tree at the young tree

Vision issues

A girl walking by in a green field in the sun

The silhouetted building and cross stands under water around Loon Mountain

male tiger sighting in twelve months of a street
Not so good results

Language issues

A Moo cow tied up around the city eating grass in various places under the tree at the young tree

male tiger sighting in twelve months of a street

Vision issues

a girl walking by in a green field in the sun

The silhouetted building and cross stands under water around Loon Mountain

Just plain silly

bike was left here by an ancient civilization not as sophisticated as our own in the grass of granite

dogs running pic, this time, racing through the sea at Fraisthorpe near Bridlington of Christmas tree in bed
What about 2nd language learning?

- Obvious problems
  - Assumes knowledge 1st language
  - Assumes knowledge of the world
  - Still don't have a robot...
What about 2nd language learning?

- Obvious problems
  - Assumes knowledge 1st language
  - Assumes knowledge of the world
  - Still don't have a robot...

- But we do have software with exercises for SLA
What about 2nd language learning?

- Obvious problems
  - Assumes knowledge 1st language
  - Assumes knowledge of the world
  - Still don't have a robot...

- But we do have software with exercises for SLA

It's hard for people, too!
What about 2nd language learning?

- Obvious problems
  - Assumes knowledge of the first language
  - Assumes knowledge of the world
  - Still don't have a robot...

- But we do have software with exercises for SLA

It's hard for people, too!
Aspects of computational 2nd LL

• Very specific linguistic variants
  • Number, case, agreement, etc.
  • *Not enough* to get the majority case
Aspects of computational 2\textsuperscript{nd}LL

- Very specific linguistic variants
  - Number, case, agreement, etc.
  - \textit{Not enough} to get the majority case

- Focus on subtle visual differences
Aspects of computational 2\textsuperscript{nd}LL

- AI-style reasoning & one-shot learning
What is needed to solve this?

- Linguistic model over character sequences (words not okay!) w/o any L-specific background
- Pre-trained (?) visual detectors for objects, poses and physical relationships (eg., gaze)
- Ability to reason and generalize from a few examples
Thanks!

Questions?

Yiannis Aloimonos
Tamara Berg
Alex Berg
Jesse Dodge
Amit Goyal
Yejin Choi
Xufeng Han
Alyssa Mensch
Meg Mitchell
Karl Stratos
Ching Lik Teo
Yezhou Yang
Kota Yamaguchi