Image Captioning

Danna Gurari
The University of Texas at Austin
Fall 2019

https://www.ischool.utexas.edu/~dannag/Courses/CrowdsourcingForCV/CourseContent.html

Review

- Last week
  - Object detection applications
  - Object detection evaluation
  - Crowdsourcing object detections
- Assignments (Class Website & Canvas)
  - Lab assignment 2 due next week
  - Project pre-proposal due next week
- Questions?

Today's Topics

- Guest Speaker: Colleen Lyon on image collection
- Image captioning applications
- Image caption evaluation
- Crowdsourcing captions
- Lab: retrieving results from AMT and submitting batches

Today's Topics

- Guest Speaker: Colleen Lyon on image collection
- Image captioning applications
- Image caption evaluation
- Crowdsourcing captions
- Lab: retrieving results from AMT and submitting batches

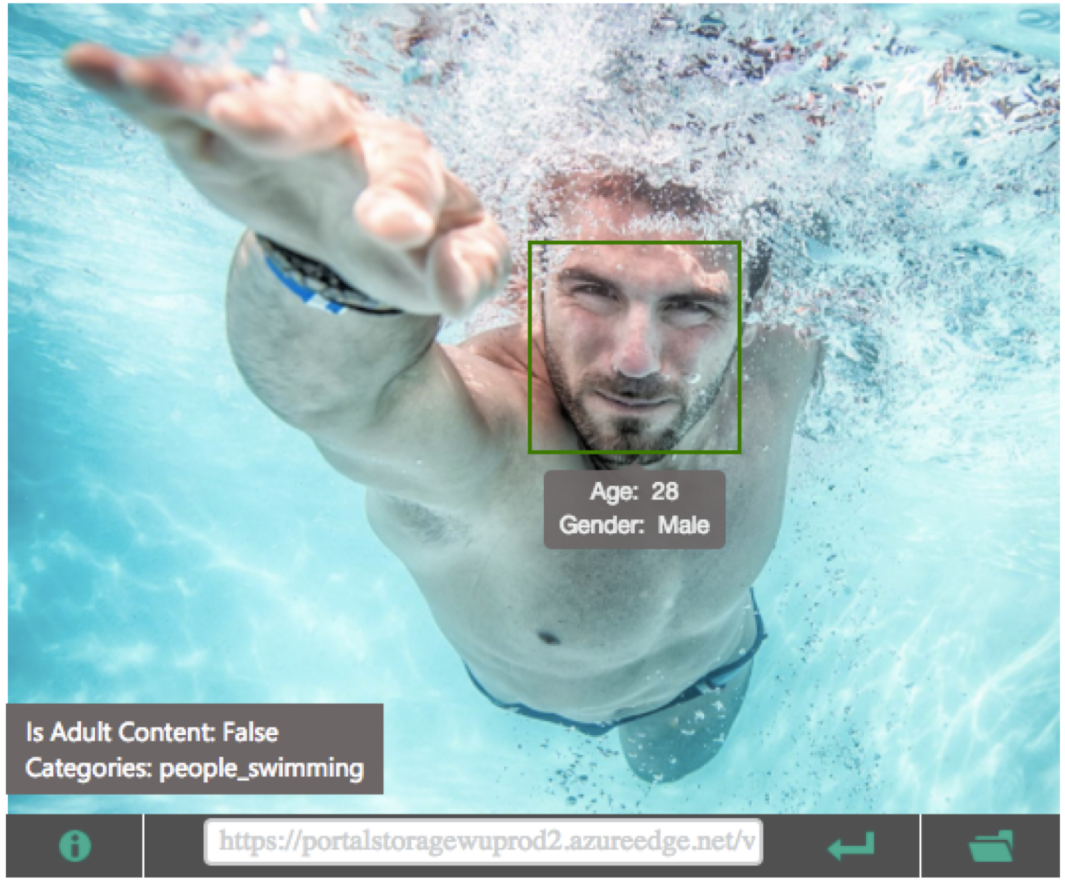
UT Austin Head of Scholarly Communications

Expertise: copyright, open access, open educational resources, and scholarly publishing issues

Today's Topics

- Guest Speaker: Colleen Lyon on image collection
- Image captioning applications
- Image caption evaluation
- Crowdsourcing captions
- Lab: retrieving results from AMT and submitting batches

More “Human-Like” Description



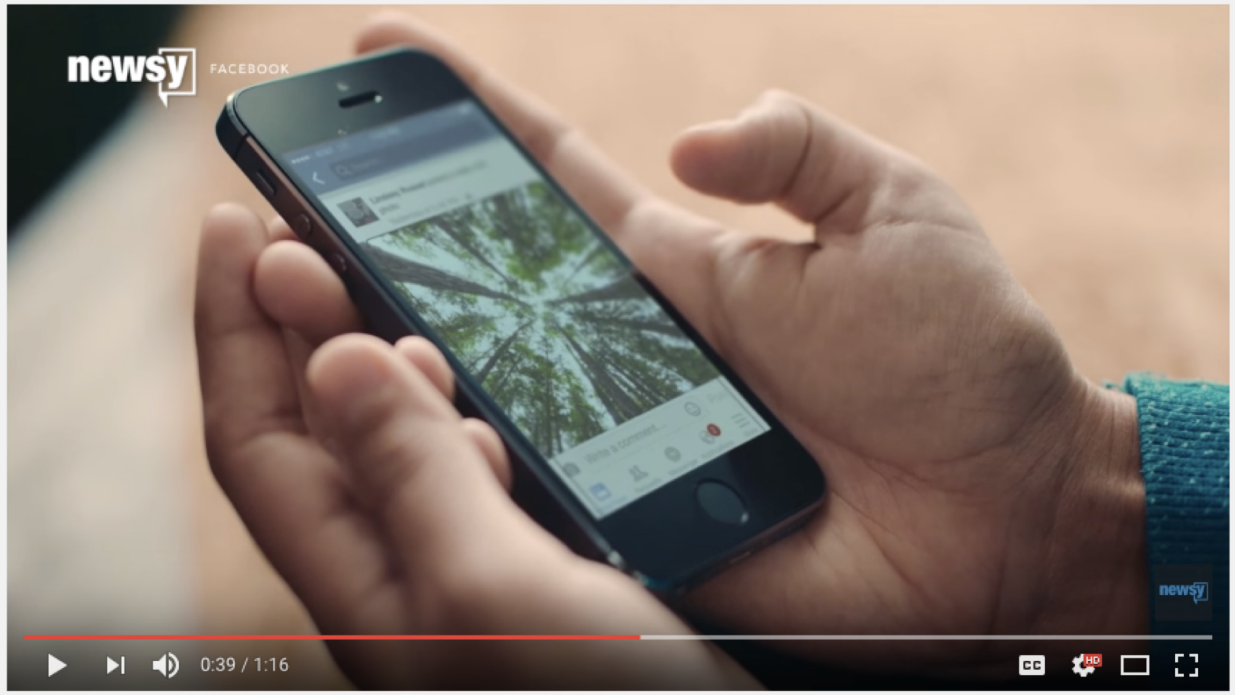
Features:

Feature Name	Value
Description	{ "type": 0, "captions": [{ "text": "a man swimming in a pool of water", "confidence": 0.7850108693093019 }] }
Tags	[{ "name": "water", "confidence": 0.9996442794799805 }, { "name": "sport", "confidence": 0.9504992365837097 }, { "name": "swimming", "confidence": 0.9062818288803101, "hint": "sport" }, { "name": "pool", "confidence": 0.8787588477134705 }, { "name": "water sport", "confidence": 0.631849467754364, "hint": "sport" }]
Image Format	jpeg
Image Dimensions	1500 x 1155
Clip Art Type	0 Non-clipart
Line Drawing Type	0 Non-LineDrawing
Black & White Image	False

Captions: https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

Visual Assistance for People with Visual Impairments

Facebook



Facebook's New AI Tool Is Helping Blind Users 'See' Photos - Newsy

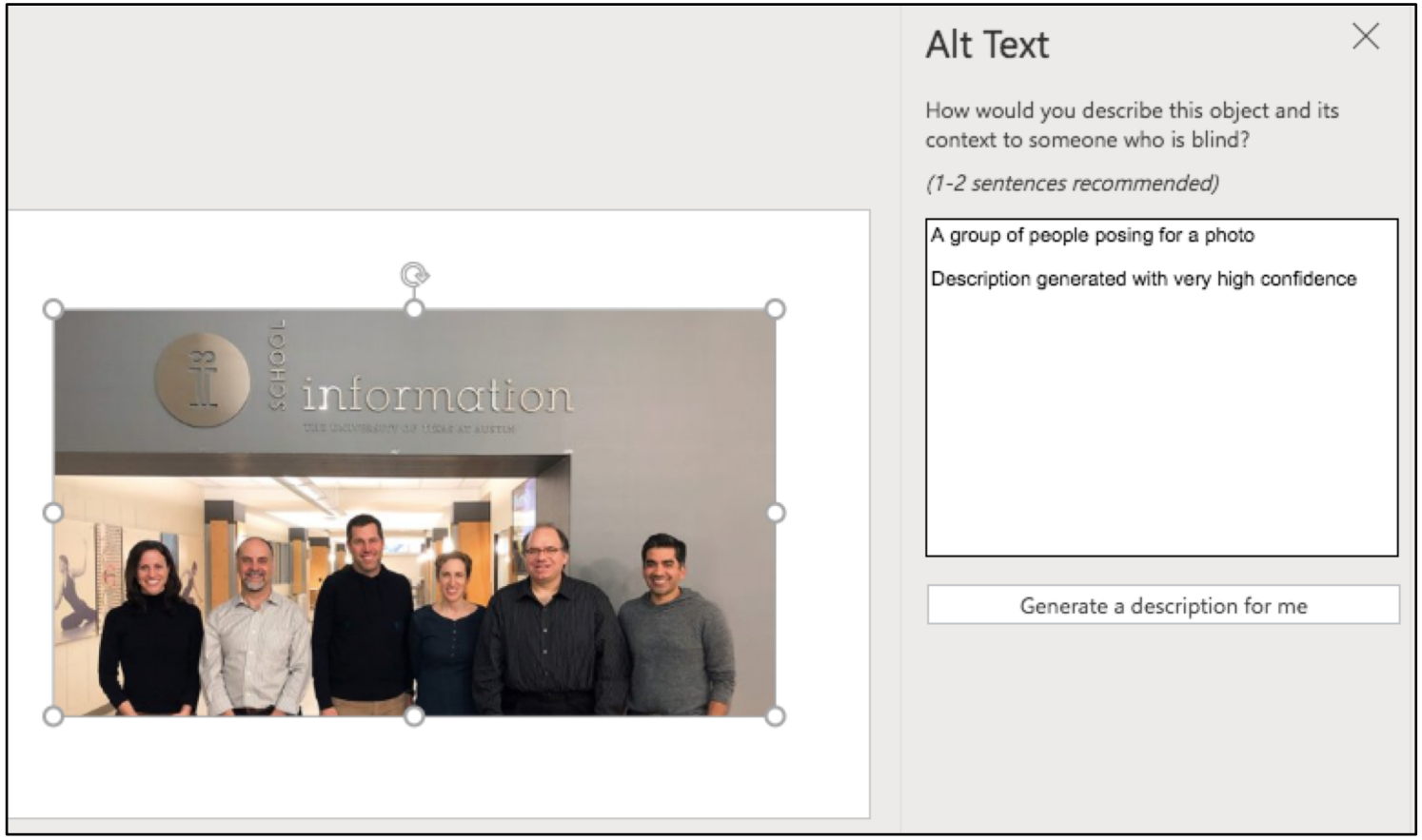
https://www.youtube.com/watch?v=Tjugc8a836Q

Microsoft

Saqib Shaikh : Microsoft Developer Can 'See' Using Artificial Intelligence Headset

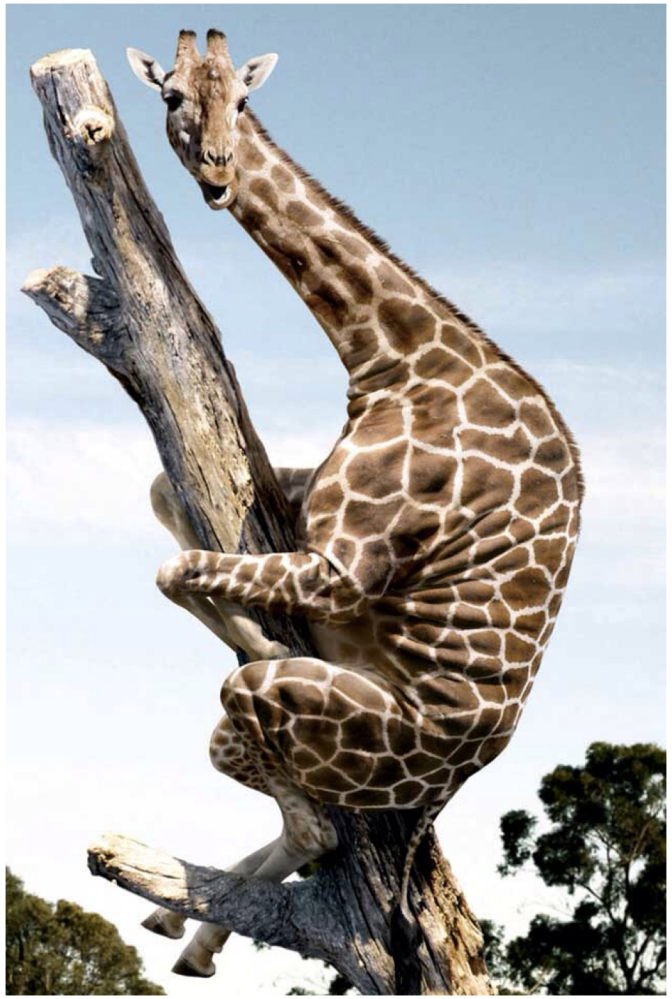
https://www.youtube.com/watch?v=R2mC-NUAmMk

Visual Assistance for People with Visual Impairments



e.g., Microsoft Power Point (Office 365 demo)

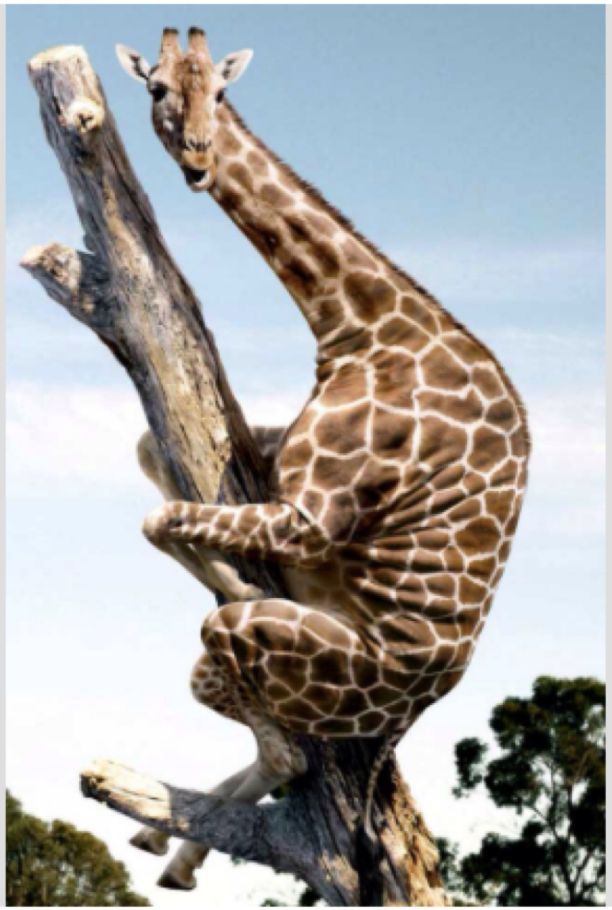
Class Task: Describe Image



Today's Topics

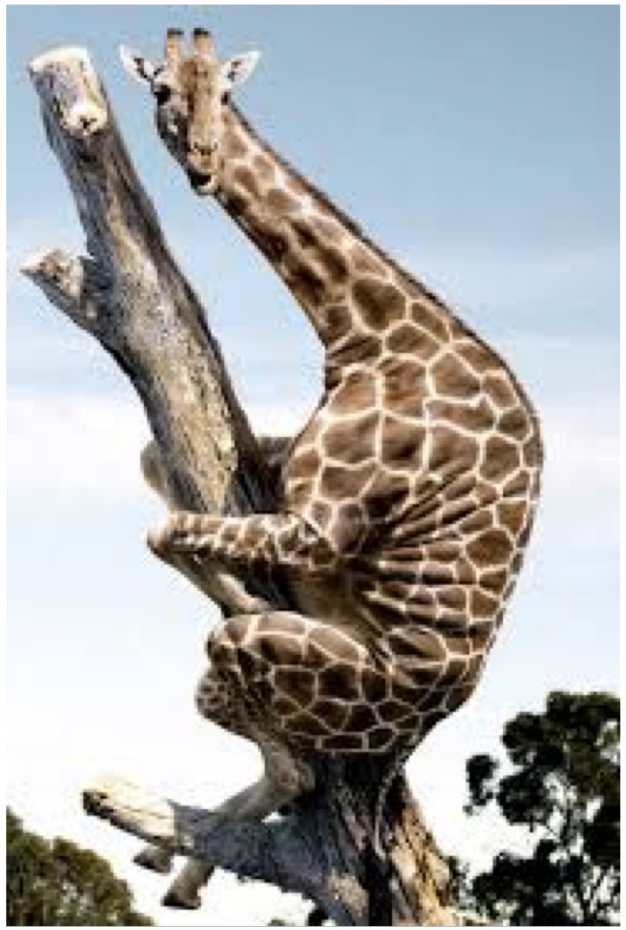
- Guest Speaker: Colleen Lyon on image collection
- Image captioning applications
- Image caption evaluation
- Crowdsourcing captions
- Lab: retrieving results from AMT and submitting batches

Class Task: How to Evaluate Predicted Captions?



FEATURE NAME:	VALUE
Description	{ "tags": ["outdoor", "giraffe", "animal", "mammal", "standing", "field", "top", "branch", "bird", "eating", "head", "grazing", "neck", "water", "large", "man", "grassy", "tall", "group", "dirt", "zoo"], "captions": [{ "text": "a giraffe standing in the dirt", "confidence": 0.982929349 }] }
Tags	[{ "name": "outdoor", "confidence": 0.999545038 }, { "name": "sky", "confidence": 0.999435842 }, { "name": "giraffe", "confidence": 0.99890554 }, { "name": "tree", "confidence": 0.997503936 }, { "name": "animal", "confidence": 0.9092593 }, { "name": "mammal", "confidence": 0.8548364 }, { "name": "day", "confidence": 0.149034753 }]
Image format	"Jpeg"

Class Task: How to Evaluate Predicted Captions?



FEATURE NAME:	VALUE
Description	{ "tags": ["outdoor", "giraffe", "animal", "mammal", "standing", "field", "top", "branch", "bird", "eating", "head", "grazing", "neck", "water", "large", "man", "grassy", "tall", "group", "dirt", "zoo"], "captions": [{ "text": "a giraffe standing in the dirt", "confidence": 0.982929349 }] }

Evaluation: Human Judgments

Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6

- The description accurately describes the image (Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013; Hodosh et al., 2013).
- The description is grammatically correct (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013).
- The description has no incorrect information (Mitchell et al., 2012).
- The description is relevant for this image (Li et al., 2011; Yang et al., 2011).
- The description is creatively constructed (Li et al., 2011).
- The description is human-like (Mitchell et al., 2012).

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic Description from Images: A Survey of Models, Datasets, and Measures. JAIR 2016.

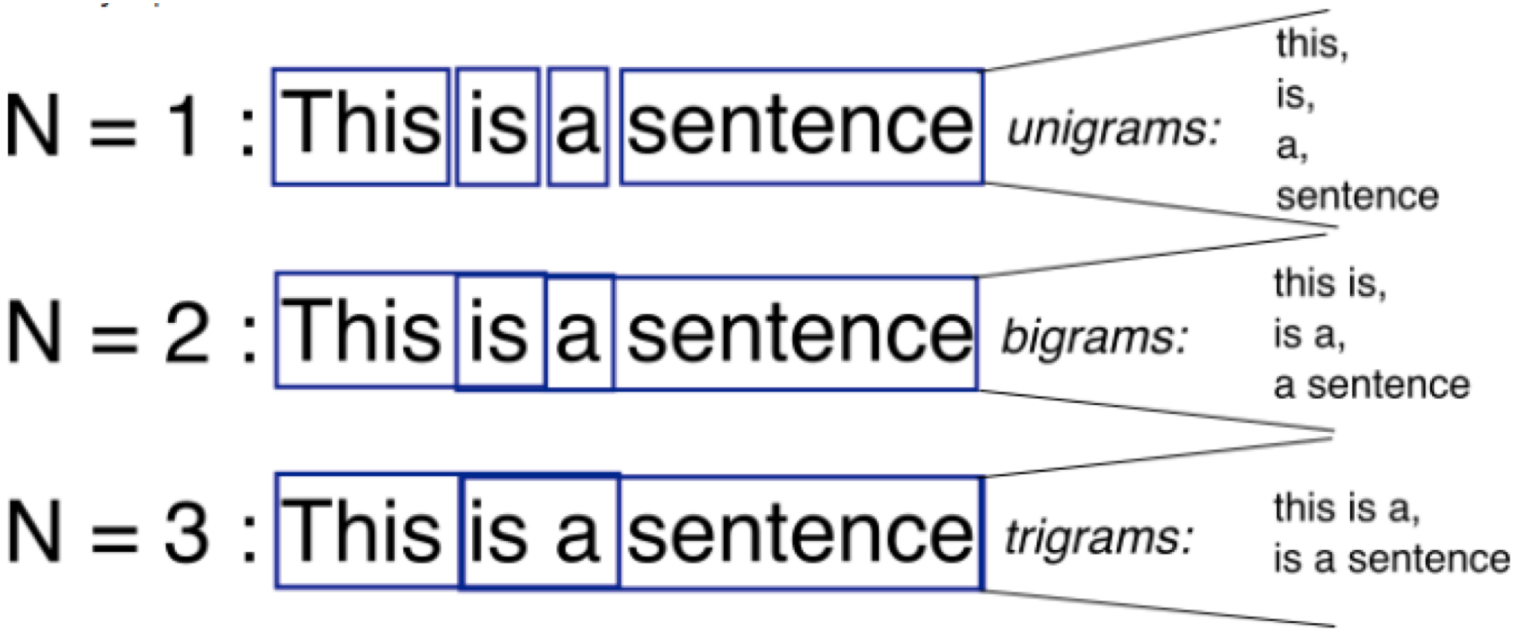
Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: compute similarities of n-grams between a predicted caption and each ground truth caption



http://recognize-speech.com/language-model/n-gram-model/comparison

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: measure similarity of a predicted caption to how most people describe an image based on *n*-grams unique to the image

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.

R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

What content do most people describe in this image?

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.

R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Problem: n-gram methods scores these as very similar

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Problem: n-gram methods scores these as very different

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: compare scene graph of prediction to scene graph of ground truth

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

What is the meaningful semantic content in these captions?

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Meaningful semantic content in these captions:

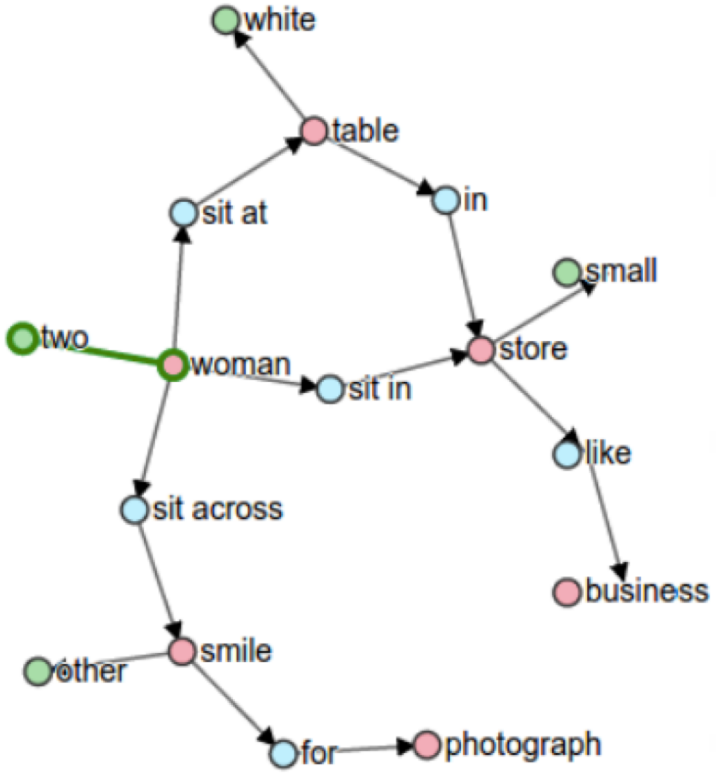
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"



P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Today's Topics

- Guest Speaker: Colleen Lyon on image collection
- Image captioning applications
- Image caption evaluation
- Crowdsourcing captions
- Lab: retrieving results from AMT and submitting batches

Image Captioning Datasets

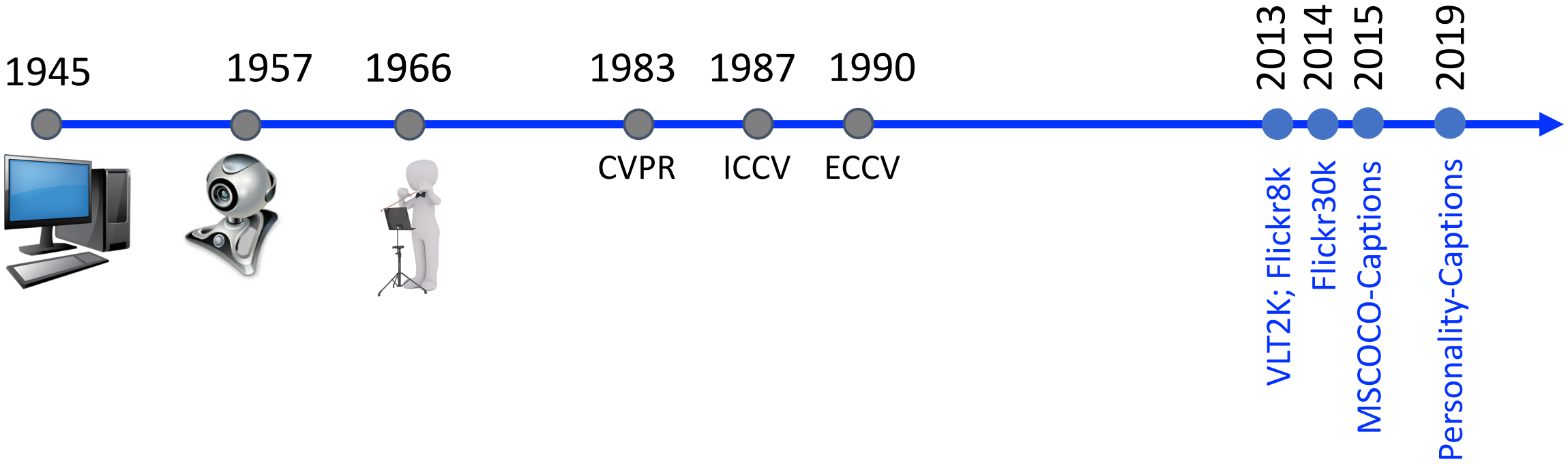
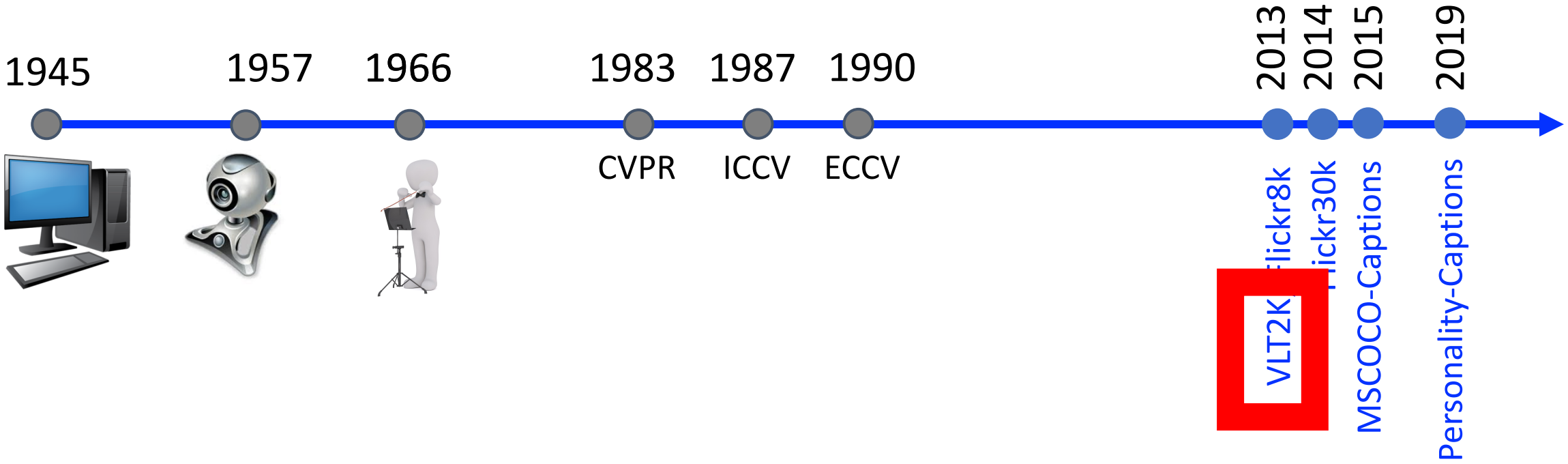


Image Captioning Datasets



VLT2K: Crowdsourcing Task

Guidelines and Examples:

Read these guidelines carefully. You must write exactly two sentences.

1. Describe the action being performed and mention the person performing the action and all objects involved in the action.
2. Describe any objects in the image that are not directly involved in the action.

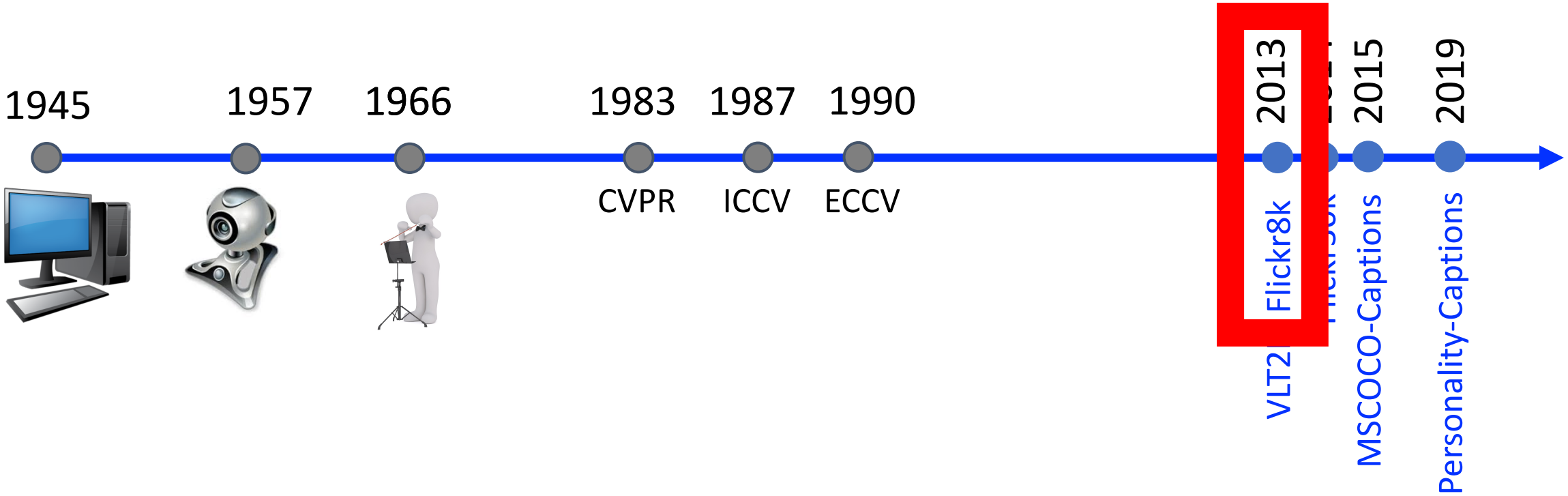
A man is reading a newspaper.
It is cloudy and there are
skyscrapers in the background.

A boy is typing on a laptop.
There is a brown bookshelf
behind him and a bright window.

A man is talking on the telephone.
There is a red lampshade and
three red chairs in the background.

D. Elliott and F. Keller. Image description using visual dependency representations. EMNLP 2013.

Image Captioning Datasets



Flickr8K: Crowdsourcing Task

Guidelines:

- You must describe each of the following five images with one sentence.
- Please provide an accurate description of the activities, people, animals and objects you see depicted in the image
- Each description must be a single sentence under 100 characters. Try to be concise.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description for all five images, leaving nothing blank.

Examples of good and bad descriptions.

(1) The dog is wearing a red sombrero.
Very Good: This describes the two main objects concisely and accurately.

(2) White dog wearing a red hat.
Good: Incomplete sentences like this are fine.

(3) The white dog is wearing a pink collar.
Okay: This describes the dog, but it ignores the hat.

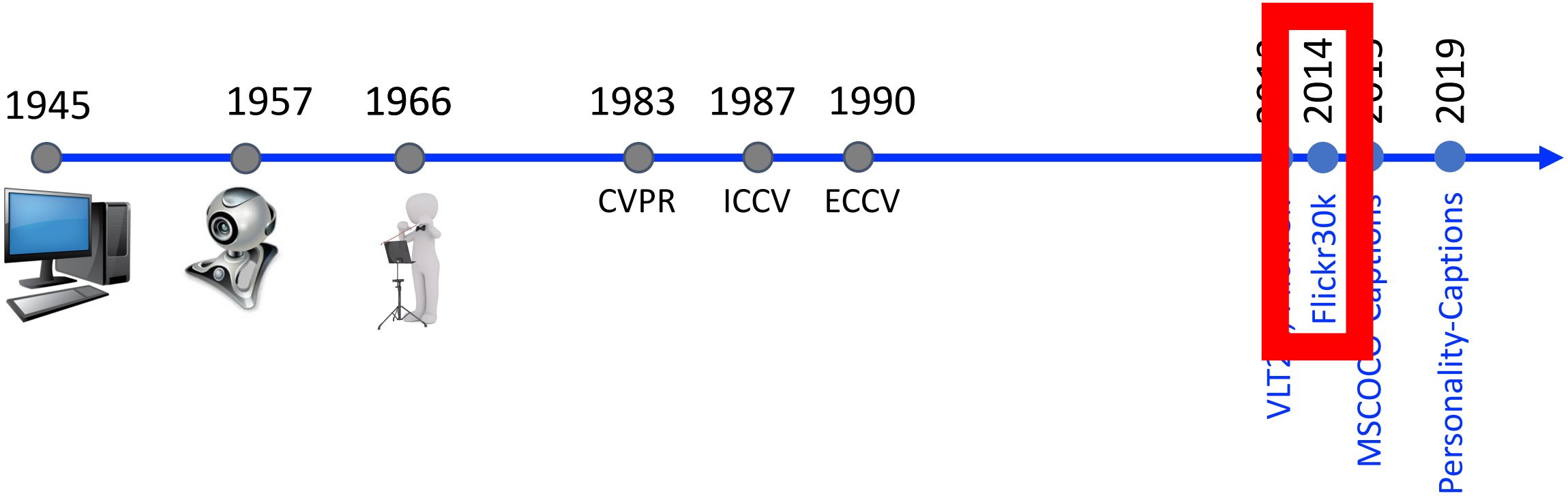
(4) The red hat is adorned with gold sequins.
Bad: This ignores the dog.

(5) The dog is angry because he is hungry.
Bad: This is speculation.

(6) The dog.
Very Bad: This could describe any image of any dog.

M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models, and evaluation metrics. JAIR 2013.

Image Captioning Datasets



Flickr30K: Crowdsourcing Task (same as Flickr8K)

Guidelines:

- You must describe each of the following five images with one sentence.
- Please provide an accurate description of the activities, people, animals and objects you see depicted in the image
- Each description must be a single sentence under 100 characters. Try to be concise.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description for all five images, leaving nothing blank.

Examples of good and bad descriptions.

(1) The dog is wearing a red sombrero.
Very Good: This describes the two main objects concisely and accurately.

(2) White dog wearing a red hat.
Good: Incomplete sentences like this are fine.

(3) The white dog is wearing a pink collar.
Okay: This describes the dog, but it ignores the hat.

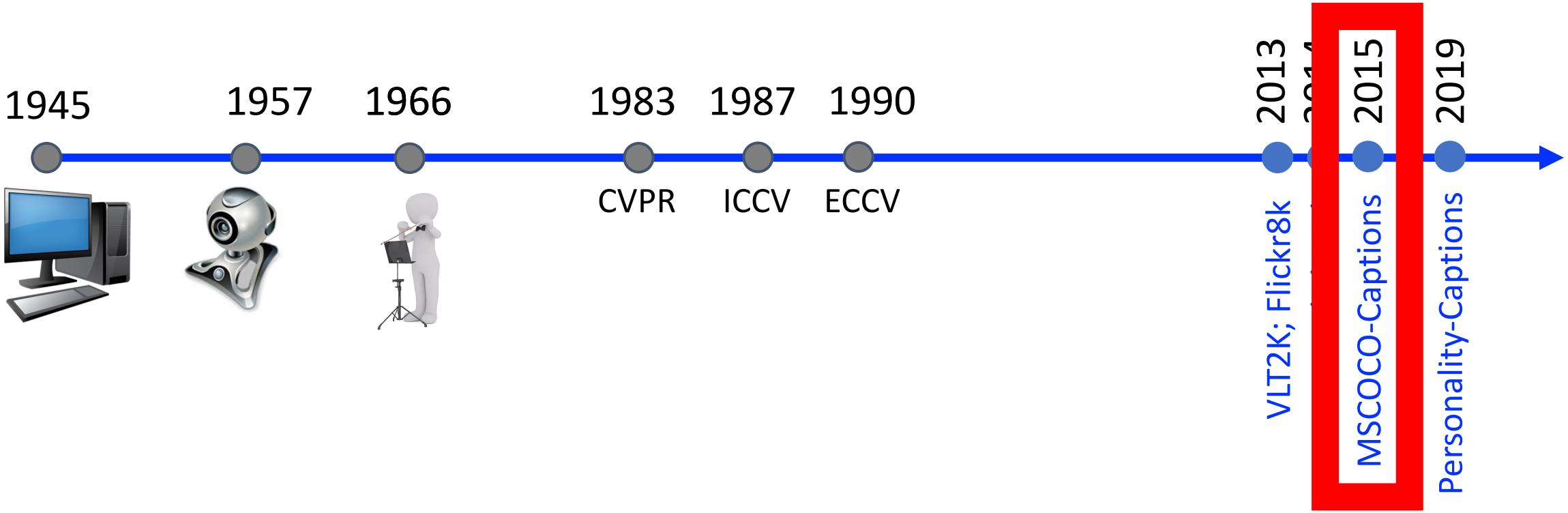
(4) The red hat is adorned with gold sequins.
Bad: This ignores the dog.

(5) The dog is angry because he is hungry.
Bad: This is speculation.

(6) The dog.
Very Bad: This could describe any image of any dog.

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image description to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics 2014.

Image Captioning Datasets



MSCOCO

Instructions:

- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

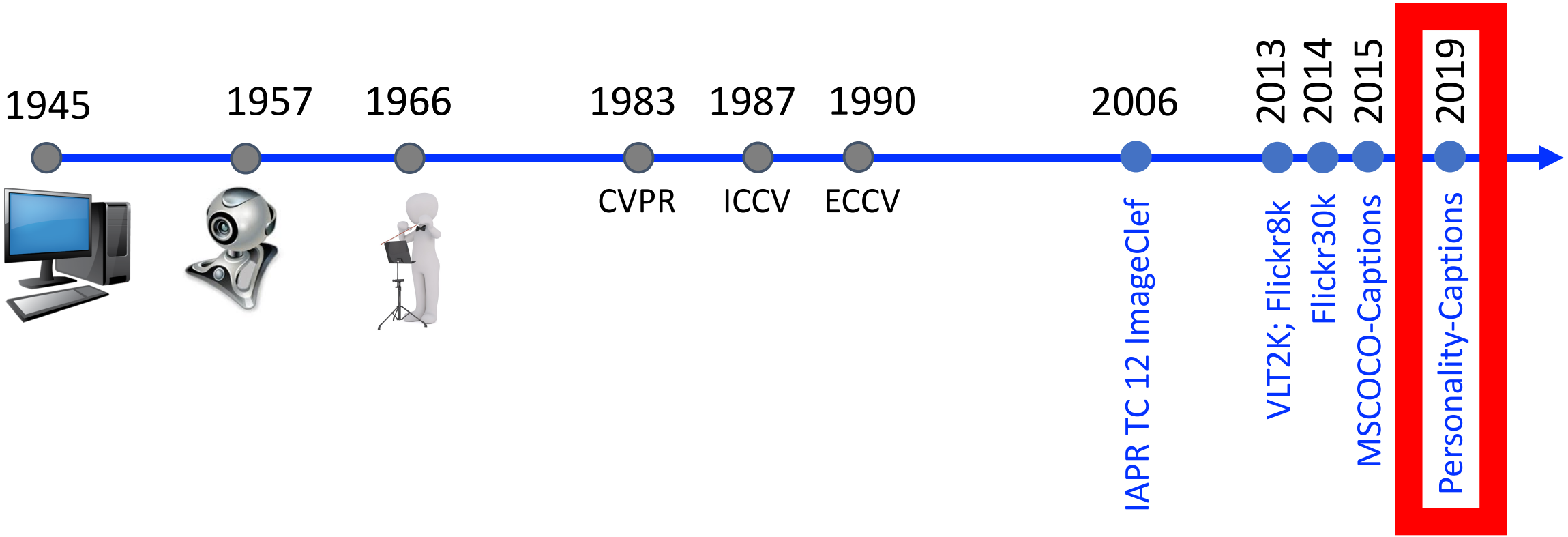
Please describe the image:

Enter description here

prev next

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv 2015.

Image Captioning Datasets



Personality-Captions: Crowdsourcing Task

215 personalities selected from this list: http://ideonomy.mit.edu/essays/traits.html

Comment on an Image

Description

In this task, you will be shown 5 images, and will write a comment about each image. The goal of this task is to write something about an image that someone else would find engaging.

STEP 1

With each new photo, you will be given a **personality trait** that you will try to emulate in your comment. For example, you might be given "**snarky**" or "**glamorous**". The personality describes ***YOU***, not the picture. It is *you* who is snarky or glamorous, not the contents of the image.

STEP 2

You will then be shown an image, for which you will write a comment *in the context of your given personality trait*. Please make sure your comment has at least **three words**. Note that these are *comments*, not captions.

E.g., you may be shown an image of a tree. If you are "**snarky**", you might write "What a boring tree, I bet it has bad wood;" or, if you were "**glamorous**", you might write "What an absolutely beautiful tree! I would put this in my living room it's so extravagent!"

Image

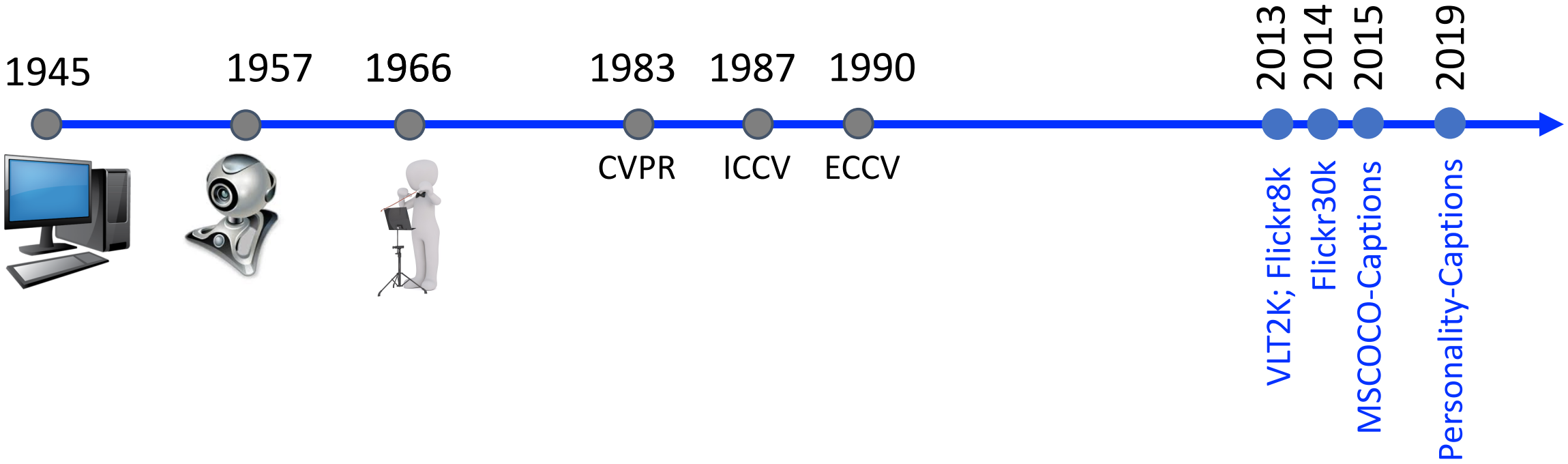
Your assigned personality is:

Adventurous

Reminder - please do not write anything that involves any level of discrimination, racism, sexism and offensive religious/politics comments, otherwise the submission will be rejected.

K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. CVPR 2019.

Image Captioning Datasets



Inter-Human Caption Agreement

- BLEU
- METEOR
- Rouge
- CIDEr

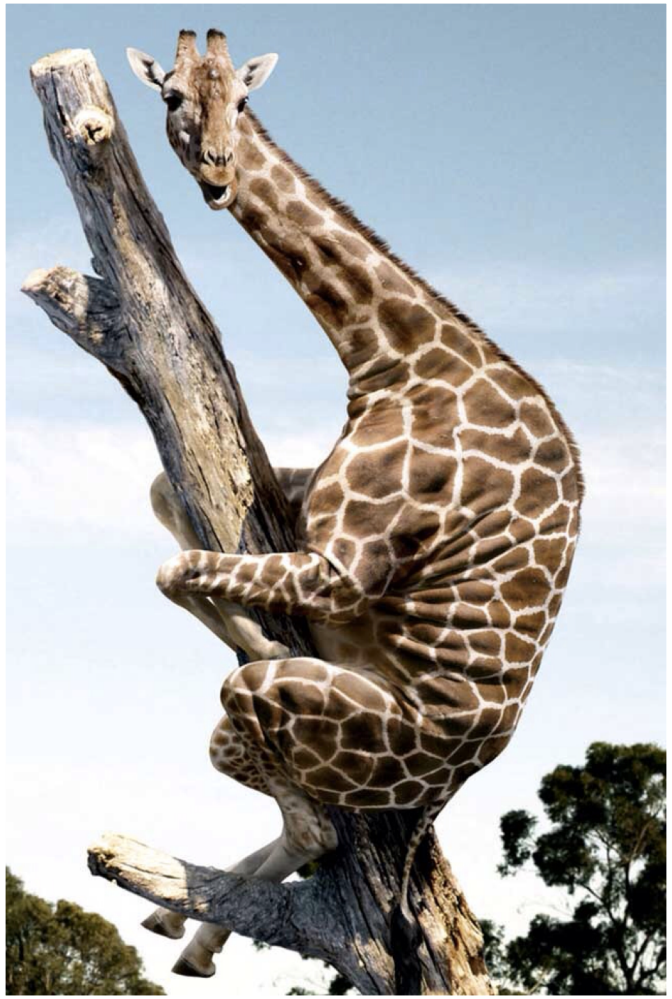
40,775 images:

Metric Name	MS COCO c5	MS COCO c40
BLEU 1	0.663	0.880
BLEU 2	0.469	0.744
BLEU 3	0.321	0.603
BLEU 4	0.217	0.471
METEOR	0.252	0.335
ROUGE_L	0.484	0.626
CIDEr-D	0.854	0.910

Better results when more human opinions are considered!

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv 2015.

Class Task: What Instructions Would You Use to Collect Captions from Human Annotators?



Today's Topics

- Guest Speaker: Colleen Lyon on image collection
- Image captioning applications
- Image caption evaluation
- Crowdsourcing captions
- Lab: retrieving results from AMT and submitting batches