RAVC 2020

# 视觉与语言的现状与趋势
# My two cents


金琴
中国人民大学信息学院
08/29/2020

# Vision and Language

- Advancements in deep learning + large-scale datasets → great progress in CV & NLP.
- Building upon these advances, interest in solving challenges that combine linguistic and visual information.
  - ‘Turing test for vision,’ i.e., the ability of a machine to use vision to answer a large and flexible set of queries about objects and agents in an image in a human-like manner.

# Vision and Language Tasks

- V→L
  - Image/Video Captioning
  - Visual Question Generation
- L→V
  - Language Guided Image/Video Generation
- V+L→V/L
  - Visual Question Answering
  - Visual Dialog
  - Referring Expression (Visual Grounding)
  - Visual Navigation
  - Multimodal Machine Translation

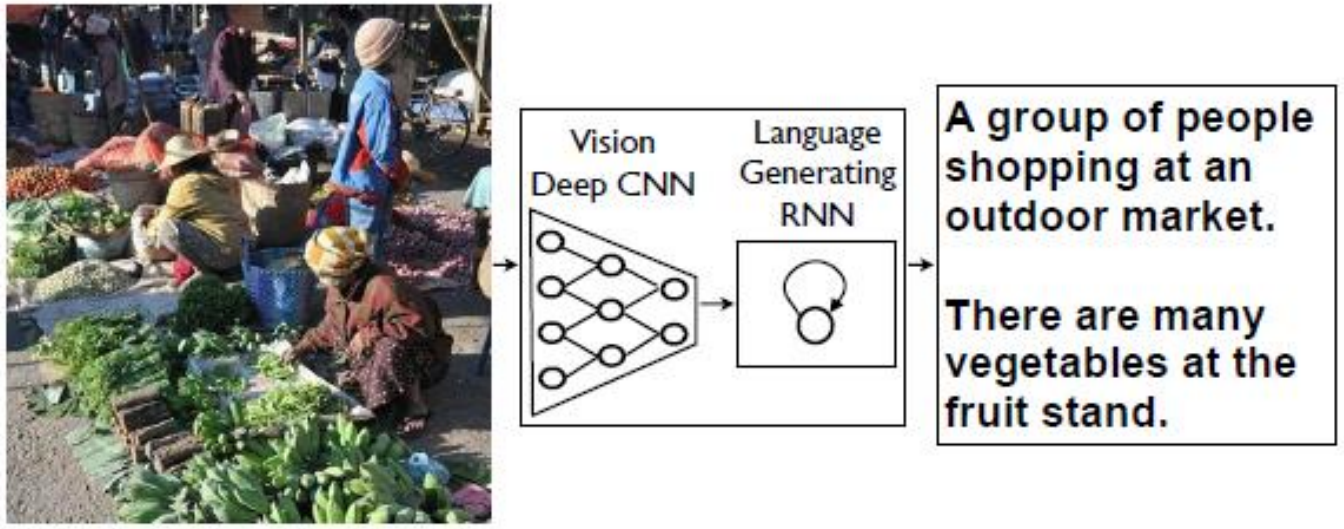
# Image Captioning

- **Automatically describe images with words**

*A giraffe is bowing and eating grass*

# Image Captioning

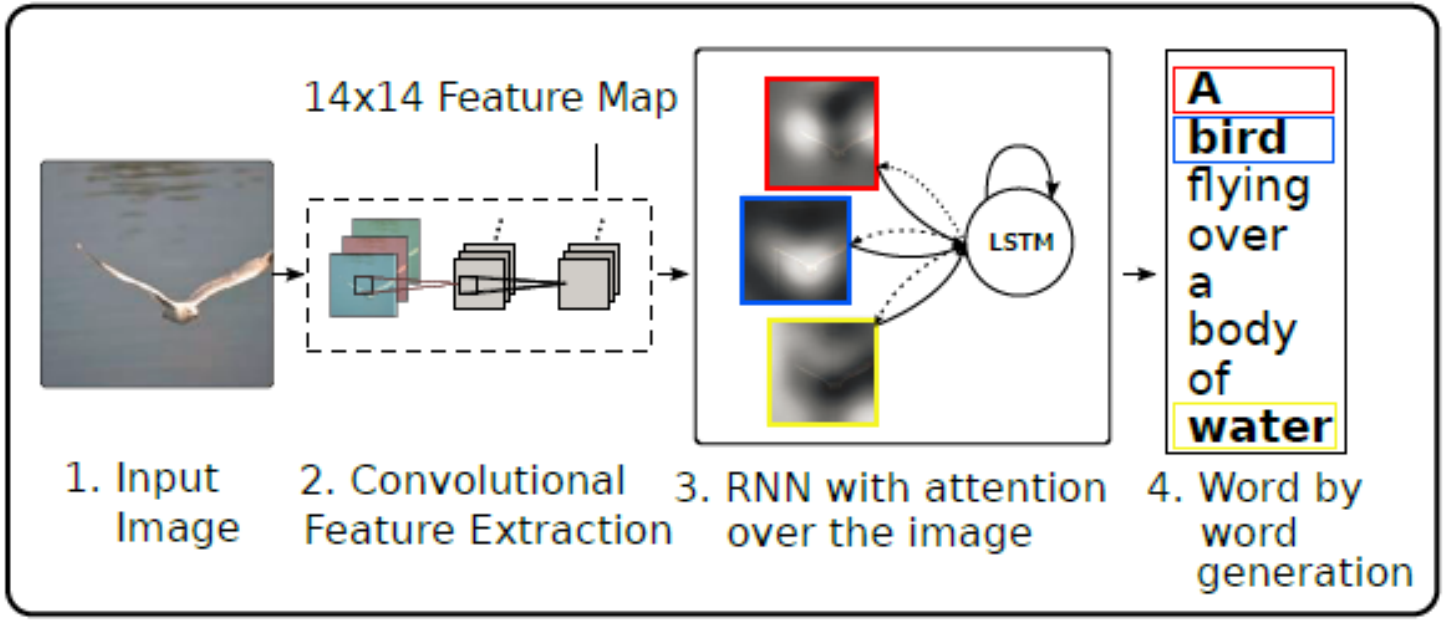
- **Encoder-decoder (CNN-RNN)**




Show and Tell: A Neural Image Caption Generator, Vinyals et al. from Google
http://googleresearch.blogspot.com/2014/11/a-picture-is-worth-thousand-coherent.html

# Image Captioning

- **Encoder-decoder (CNN-RNN)**
- **Attention**




Show, Attend, and Tell  Neural Image Caption Generation with Visual Attention. Kelvin Xu, Jimmy Ba, Jamie Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio ICML 2015

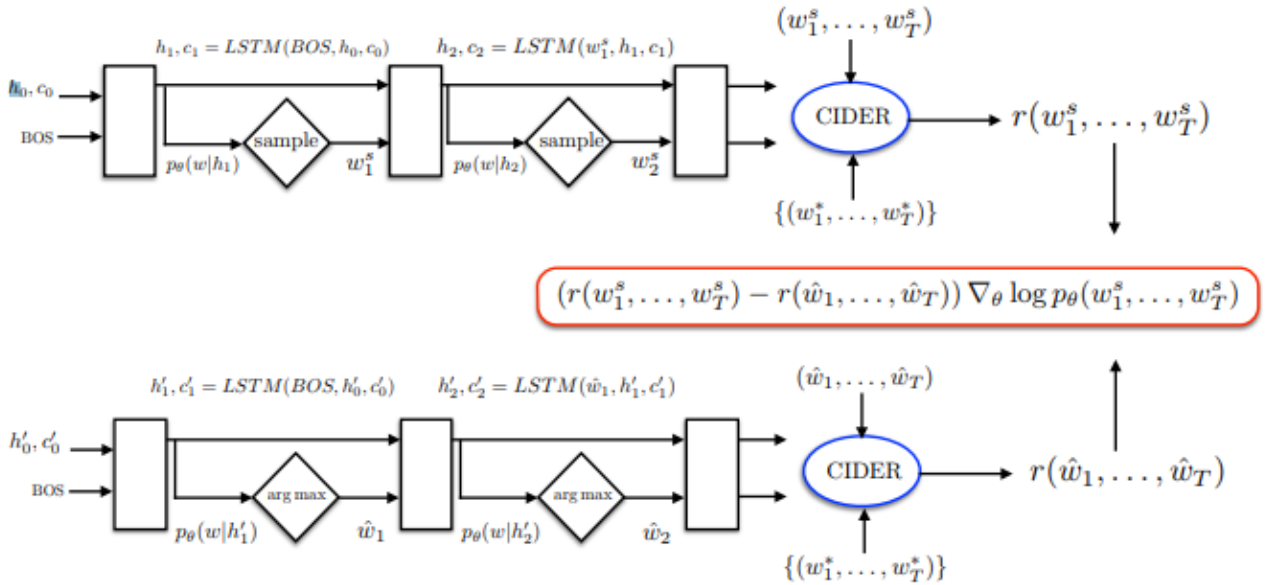
# Image Captioning

- **Encoder-decoder (CNN-RNN)**
- **Attention**

- Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. CVPR 2017.
- SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. CVPR 2017.
- Attention is All You Need. NIPS 2017.
- Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR 2018.
- Deliberate Attention Networks for Image Captioning. AAAI 2019.
- Attention on Attention for Image Captioning. ICCV 2019.
- X-Linear Attention Networks for Image Captioning. CVPR 2020.

# Image Captioning

- Encoder-decoder (CNN-RNN)
- Attention
- Reinforcement Learning




Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, Vaibhava Goel. Self-critical Sequence Training for Image Captioning. CVPR 2017.

# Image Captioning

- **Flickr8k dataset (2013)**
- **Flickr30k dataset (2014)**
- **MSCOCO Captions (2015)**

| Split | Images | Captions per Image | Total Captions |
|---|---|---|---|
| Training | 6,000 | 5 | 30,000 |
| Validation | 1,000 | 5 | 5,000 |
| Test | 1,000 | 5 | 5,000 |
| Total | 8,000 | 5 | 40,000 |

| Split | Images | Captions per Image | Total Captions |
|---|---|---|---|
| Training | 29,000 | 5 | 145,000 |
| Validation | 1,014 | 5 | 5,070 |
| Test | 1,000 | 5 | 5,000 |
| Total | 31,014 | 5 | 155,070 |

| Split | Images | Captions per Image | Total Captions |
|---|---|---|---|
| Training | 113,287 | 5 | 566,435 |
| Validation | 5,000 | 5 | 25,000 |
| Test | 5,000 | 5 | 25,000 |
| Total | 123,287 | 5 | 616,435 |

# Image Captioning

- **Flickr8k dataset (2013)**
- **Flickr30k dataset (2014)**
- **MSCOCO Captions (2015)**

- **Visual Genome (2017)**
- **Conceptual Captions (CC) (2018)**

| Split | Images | Captions |
|---|---|---|
| Training | 3,318,333 | 3,318,333 |
| Validation | 15,840 | 15,840 |
| Test | 22,530 | 22,530 |

| Split | Images | Captions per Image | Total Captions |
|---|---|---|---|
| Training | 6,000 | 5 | 30,000 |
| Validation | 1,000 | 5 | 5,000 |
| Test | 1,000 | 5 | 5,000 |
| Total | 8,000 | 5 | 40,000 |

| Split | Images | Captions per Image | Total Captions |
|---|---|---|---|
| Training | 29,000 | 5 | 145,000 |
| Validation | 1,014 | 5 | 5,070 |
| Test | 1,000 | 5 | 5,000 |
| Total | 31,014 | 5 | 155,070 |

| Split | Images | Captions per Image | Total Captions |
|---|---|---|---|
| Training | 113,287 | 5 | 566,435 |
| Validation | 5,000 | 5 | 25,000 |
| Test | 5,000 | 5 | 25,000 |
| Total | 123,287 | 5 | 616,435 |


- *Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 2017.*
- *Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. ACL 2018.*

# Image Captioning

- **Cross-lingual/Multi-lingual Captioning**

| Split | Images | Language of the Captions | | | |
|---|---|---|---|---|---|
| | | Czech | English | French | German |
| Training | 29,000 | 145,000 | 145,000 | 145,000 | 145,000 |
| Validation | 1,014 | 5,070 | 5,070 | 5,070 | 5,070 |
| Testing | 1,071 | 5,355 | 5,355 | 5,355 | 5,355 |

**Multi30K-CLID**

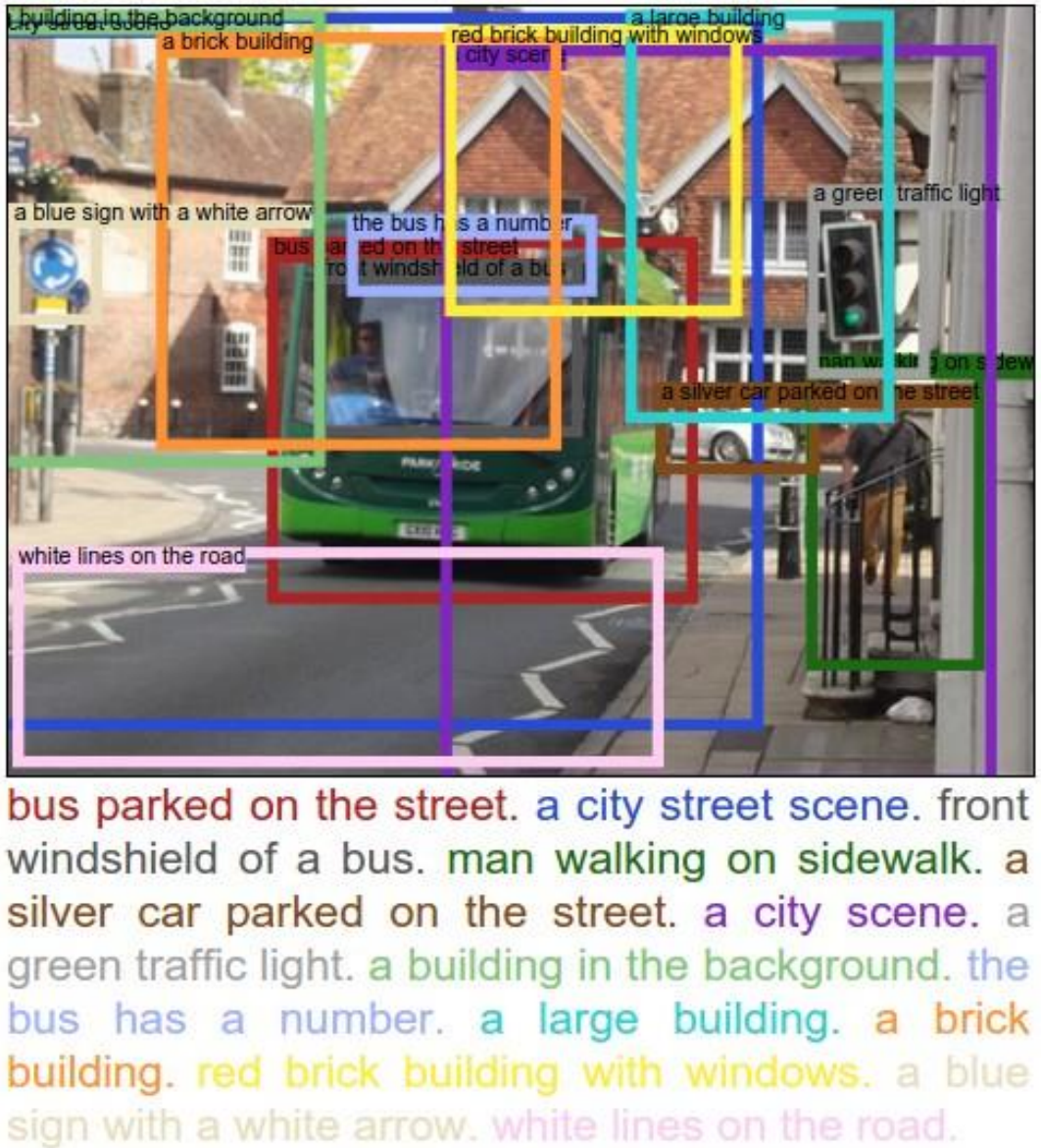
Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards. ACM Multimedia 2019.

# Image Captioning

- **Multiple sentences**
  - **Dense captioning**
  - **Paragraph captioning**

# Image Captioning

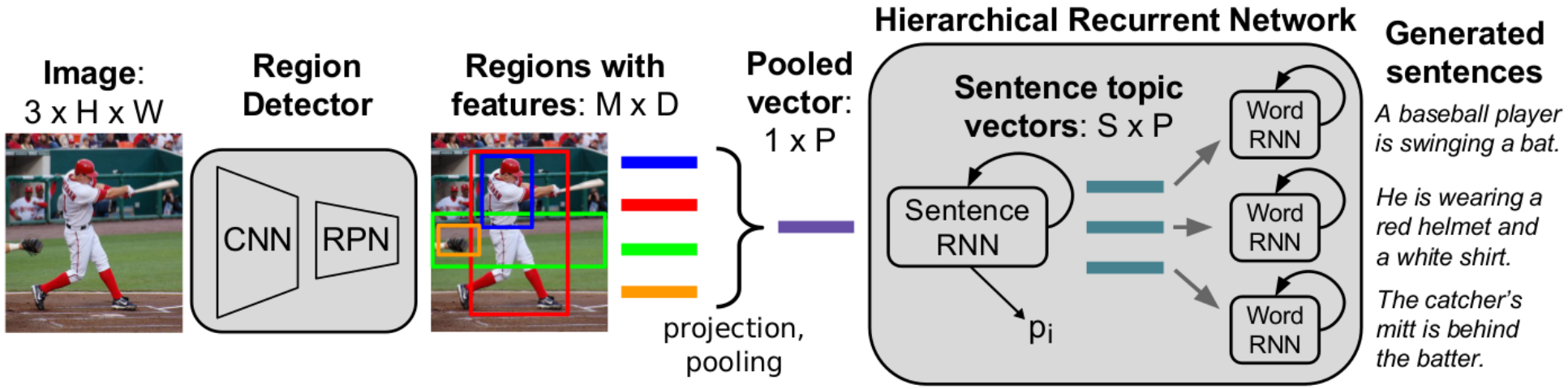
- **Dense captioning**
- requires a computer vision system to both **localize and describe salient regions** in images in natural language




- DenseCap: Fully Convolutional Localization Networks for Dense Captioning. ICCV 2016.
- Learning Object Context for Dense Captioning. AAAI 2019.
- Context and Attribute Grounded Dense Captioning. CVPR 2019.

# Image Captioning

- **Paragraph captioning**
- Image paragraph captioning involves **generating a detailed, multi-sentence description of the content of an image**.




- Recurrent Topic-Transition GAN for Visual Paragraph Generation. ICCV 2017.
- A Hierarchical Approach for Generating Descriptive Image Paragraphs. CPVR 2017.
- Diverse and Coherent Paragraph Generation from Images. ECCV 2018.
- Convolutional Auto-encoding of Sentence Topics for Image Paragraph Generation. IJCAI 2019.

# Image Captioning

- **Novel Object Captioning**
- To encourage the development of image captioning models that can learn visual concepts from alternative data sources, such as object detection datasets.



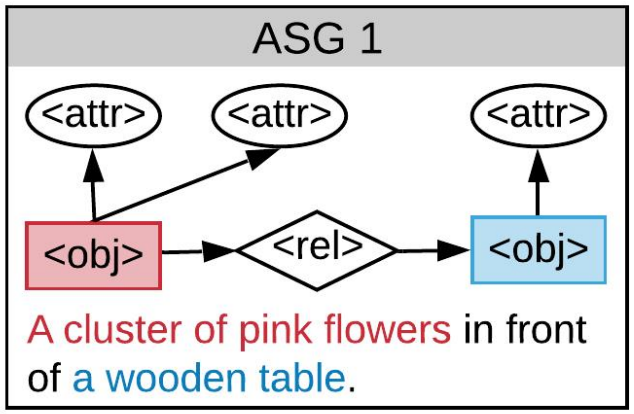

- Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects. CVPR 2017.
- Nueral baby talk. CVPR 2018.
- Partially supervised image captioning. NIPS 2018.
- nocaps: novel object captioning at scale. ICCV 2019.
- Pointing novel objects in image captioning. CVPR 2019.

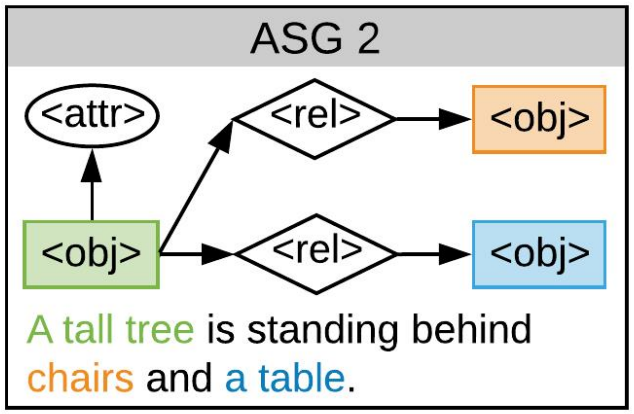
# Image Captioning

- **Controllable Captioning**
  - Style
  - Fine-grained level

Intention-agnostic Captions

- A couple of chairs sitting next to a table with flowers.
- A couple of chairs sitting next to each other on a table.
- A couple of white chairs sitting on top of a wooden table.




- Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. CVPR 2019.
- Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs. CVPR 2020.
- Length-Controllable Image Captioning. ECCV 2020.

# Image Captioning

- **Informative Captioning**

| | |
|---|---|
| BreakingNews | 2016 |
| GoodNews | 2019 |

***General caption:***
a young girl in a white tank top is holding a stick
***News caption:***
Aisling Donnelly holding the snake discovered in a back garden in County Tyrone

***News article:***
A family in County Tyrone have been paid a surprise visit by a snake. Aisling Donnelly of Clonoe just outside Coalisland said her brother and his friend came across the large snake in their back garden…Aisling Donnelly told BBC Radio Ulster that the family were not sure about how to handle or deal with the surprise visitor... However Aisling's sister was brave enough to lift the snake and then the whole family held it after the initial fear disappeared. The snake has not yet been named, although Aisling said the whole country has come round to see it … Although the Donnelly family are not planning on keeping the snake, Aisling added, 'I wouldn't mind one actually'.


- Breakingnews: Article annotation by image and text processing. PAMI 2018.
- Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. CVPR 2019.
- Transform and Tell: Entity-Aware News Image Captioning. CVPR 2020.
- ICECAP: Information Concentrated Entity-aware Image Captioning. ACM Multimedia 2020.

# Image Captioning

- **Informative Captioning**

| | |
|---|---|
| BreakingNews | 2016 |
| GoodNews | 2019 |
| **VizWiz-Captions** | **2018** |

to assist people who are blind to overcome their real daily visual challenges.


A computer screen with a Windows message about Microsoft license terms.


A can of green beans is sitting on a counter in a kitchen.


A photo taken from a residential street in front of some homes with a stormy sky above.


A blue sky with fluffy clouds, taken from a car while driving on the highway.


A hand holds up a can of Coors Light in front of an outdoor scene with a dog on a porch.


A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.


A Winnie The Pooh character high chair with a can of Yoohoo sitting on it in front of a white wall.


A cup holder in a car holding loose change from Canada.

- Captioning Images Taken by People Who Are Blind. ECCV 2020.
- VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People. CVPR 2019.
- VizWiz Grand Challenge: Answering Visual Questions from Blind People. CVPR 2018.

# Image Captioning

- **Informative Captioning**

| | |
|---|---|
| BreakingNews | 2016 |
| GoodNews | 2019 |
| VizWiz-Captions | 2018 |
| **TextCaps** | **2020** |

**TextCaps requires models to read and reason about text in images to generate captions about them. Specifically, models need to incorporate a new modality of text present in the images and reason over it and visual content in the image to generate image descriptions.**


a

**Model:** a macdonald ' s sign that is on a brick wall

**Human:** A tile wall with a red circle on it reading Mornington Crescent


b

**Model:** a sign on a table that says 'welcome to the beach'

**Human:** The Cocomaya White Chocolate with London strawberries and cream costs $3.50


c

**Model:** a sign that has the time of 12 : 37 on it

**Human:** A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks


d

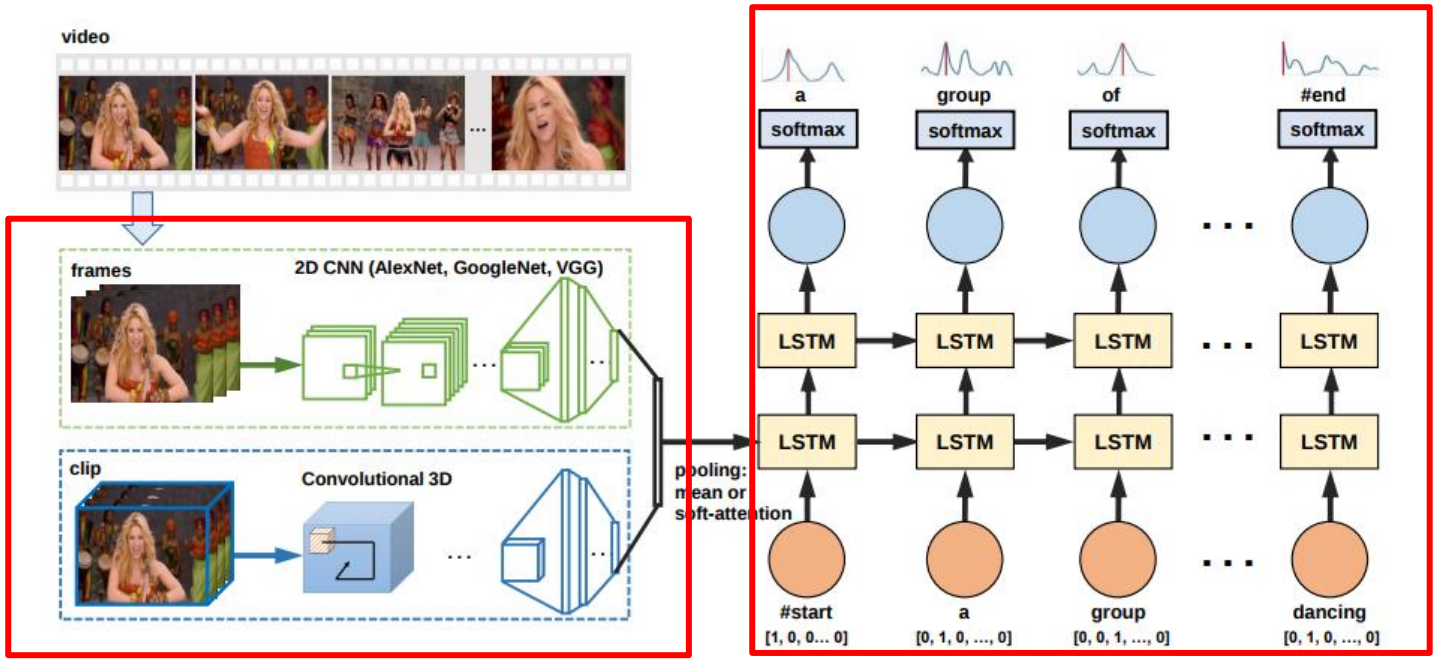
**Model:** a ruler that has the number 2003 on it

**Human:** An old artifact being measured by a ruler that shows it is around 40 millimeters wide

- TextCaps: a Dataset for Image Captioningwith Reading Comprehension. ECCV 2020.

# Video Captioning

A soccer player is kicking a ball into the goal.

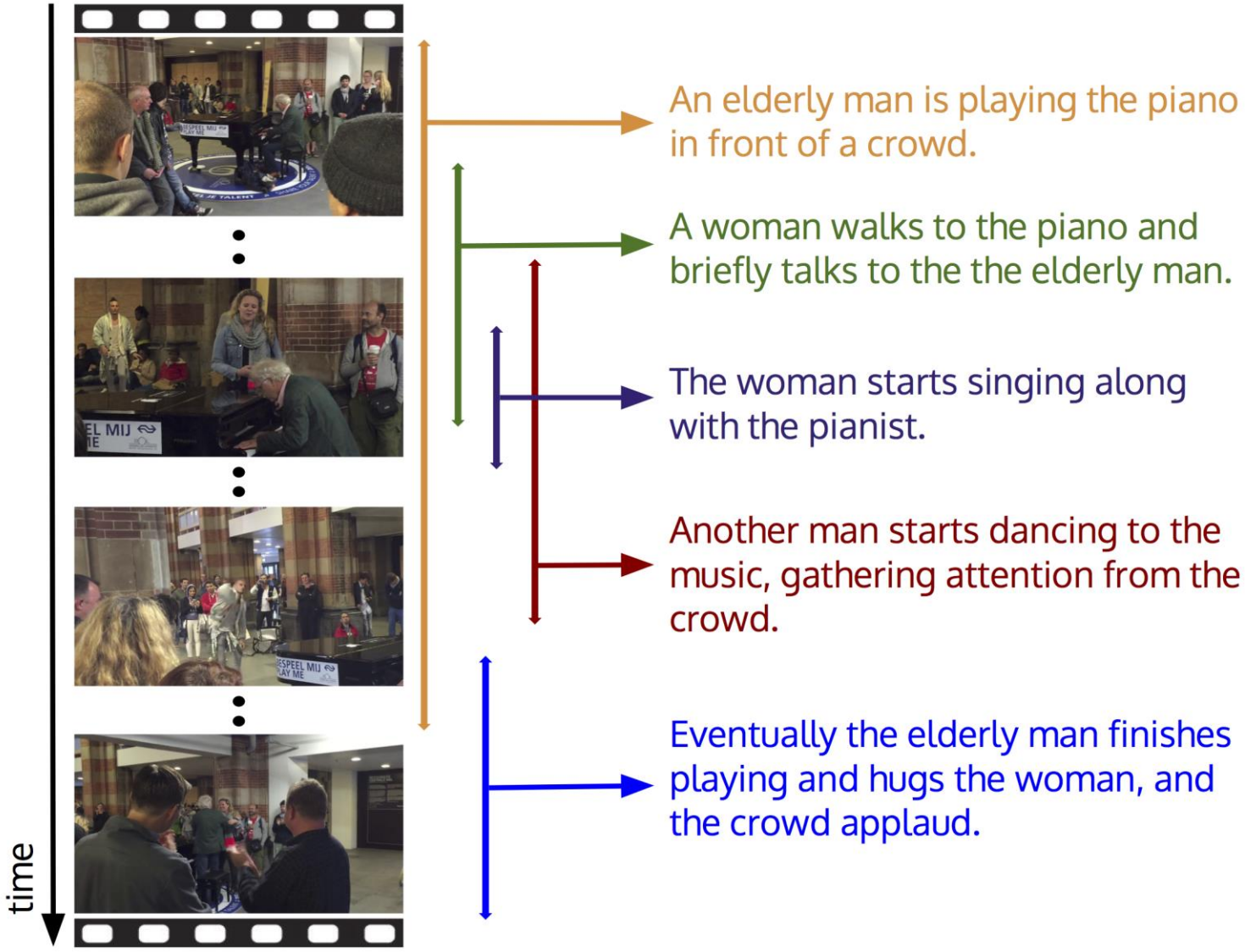


# Video Captioning

- MSVD (2011)
- MSR-VTT (2016, 2017)
- TGIF (2016)
- Vatex (2019)

# Video Captioning

- MSVD (2011)
- MSR-VTT (2016, 2017)
- TGIF (2016)
- Vatex (2019)
- ActivityNet (2015)



# Video Captioning

- MSVD (2011)
- MSR-VTT (2016, 2017)
- TGIF (2016)
- Vatex (2019)
- ActivityNet (2015)
- **TACoS (2014)**
- **MPII-MD (2015)**
- **M-VAD (2015)**
- **YouMakeup (2020)**

**Detailed**: A man took a cutting board and knife from the drawer. He took out an orange from the refrigerator. Then, he took a knife from the drawer. He juiced one half of the orange. Next, he opened the refrigerator. He cut the orange with the knife. The man threw away the skin. He got a glass from the cabinet. Then, he poured the juice into the glass. Finally, he placed the orange in the sink.
**Short**: A man juiced the orange. Next, he cut the orange in half. Finally, he poured the juice into a glass.
**One sentence**: A man juiced the orange.

# Video Captioning

- MSVD (2011)
- MSR-VTT (2016, 2017)
- TGIF (2016)
- Vatex (2019)
- ActivityNet (2015)
- **TACoS (2014)**
- **MPII-MD (2015)**
- **M-VAD (2015)**
- **YouMakeup (2020)**

**AD**: Abby gets in the basket.

**Script**: After a moment a frazzled Abby pops up in his place.

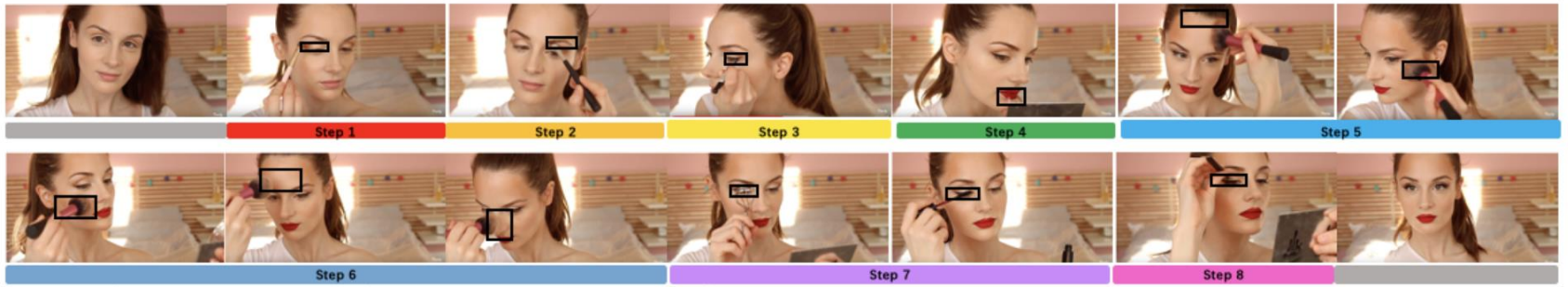
Mike leans over and sees how high they are.

Mike looks down to see – they are now fifteen feet above the ground.

Abby clasps her hands around his face and kisses him passionately. For the first time in her life, she stops thinking and grabs Mike and kisses the hell out of him.

# Video Captioning

- MSVD (2011)
- MSR-VTT (2016, 2017)
- TGIF (2016)
- Vatex (2019)
- ActivityNet (2015)
- **TACoS (2014)**
- **MPII-MD (2015)**
- **M-VAD (2015)**
- **YouMakeup (2020)**




| Step | Temporal range | Spatial Face Area | Step Description |
|---|---|---|---|
| Step 1 | 00:02:10~00:02:39 | eyelid | Apply eyeshadow over eyelid with brush |
| Step 2 | 00:02:40~00:02:54 | brow | Draw eyebrow using the brow shadow with brush |
| Step 3 | 00:02:56~00:03:10 | lashline | Draw winged eyelinear with eyelinear pen |
| Step 4 | 00:03:11~00:03:24 | lip | Apply red lipgloss on lips |
| Step 5 | 00:03:25~00:03:28 | cheek, hailine | Apply bronzer on the cheeks and hairline with brush |
| Step 6 | 00:03:29~00:03:34 | cheekbone, forehead, nose | Apply highlighter on the cheekbones, forehead and nose with brush |
| Step 7 | 00:03:35~00:03:47 | lash | Curl the lashes and apply mascara on the lashes |
| Step 8 | 00:03:54~00:04:15 | lash | Apply false lashes on the lashes |

# Visual Question Answering

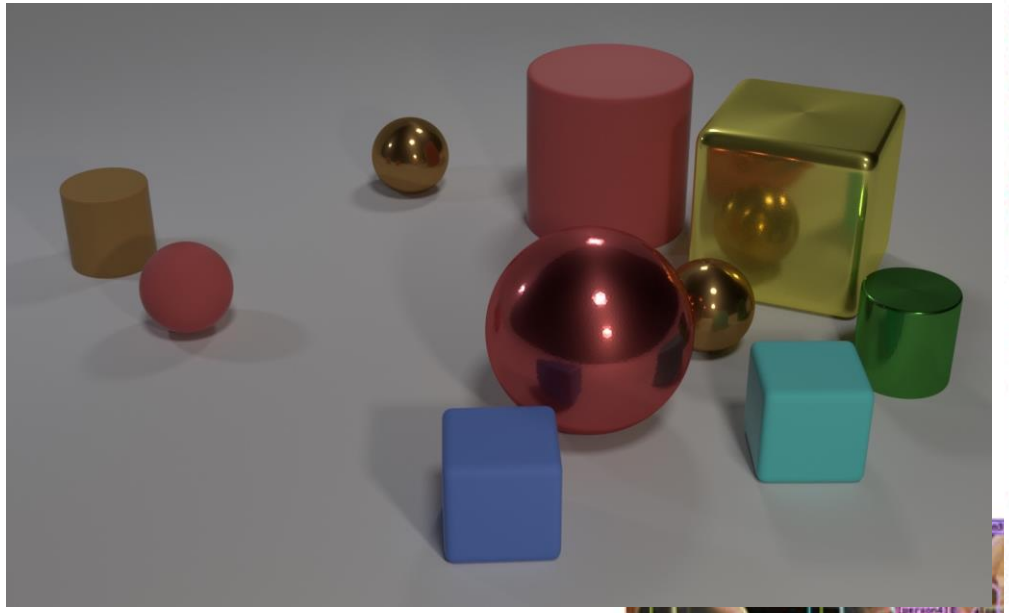
| | |
|---|---|
| • VQA v1 | 2015 |
| • VQA v2 | 2017 |
| • CLEVR | 2017 |
| • GQA | 2019 |
| • Visual Genome | 2017 |
| • TextVQA | 2019 |
| • ST-VQA | 2019 |
| • OCR-VQA-200k | 2019 |
| • EST-VQA | 2020 |
| • TVQA | 2018 |
| • YouMakeUp | 2019 |
| • TVQA+ | 2020 |

# Visual Question Answering

- VQA v1 2015
- VQA v2 2017
- CLEVR 2017
- GQA 2019
- Visual Genome 2017
- TextVQA 2019
- ST-VQA 2019
- OCR-VQA-200k 2019
- EST-VQA 2020
- TVQA 2018
- YouMakeUp 2019
- TVQA+ 2020

- **Joint embedding approaches**
- **Attention mechanisms**
- **Compositional Models**
- **Models using external knowledge base**

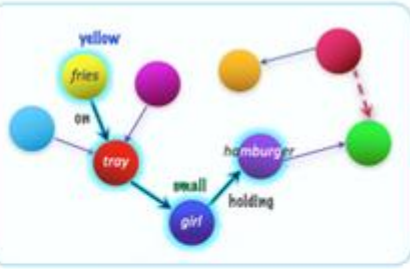
# Visual Question Answering



Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: Select: <dobject> → Choose <type>:<attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → Relate: girl,holding → Filter size: small → Relate: object, left → Filter color: red → Relate: food,on → Choose color: yellow | brown



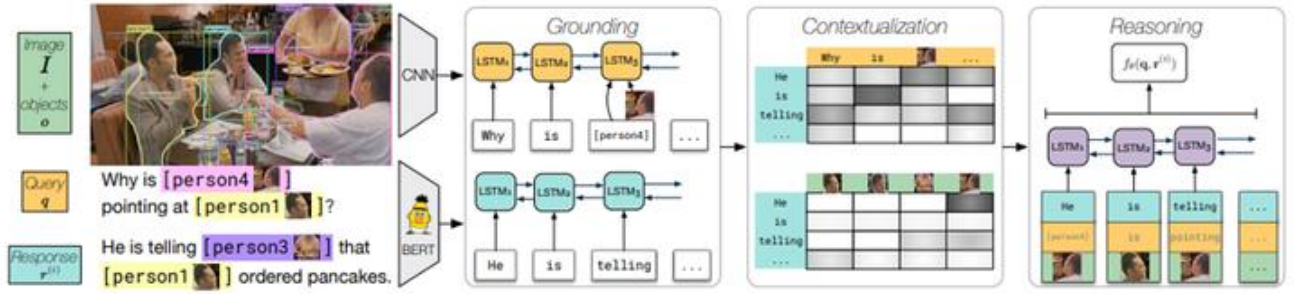

| Graph Normalization | Question Generation | Sampling and Balancing | Entailments Relations | New Metrics |
|---|---|---|---|---|
| • Ontology construction<br>• Edge Pruning<br>• Object Augmentation<br>• Global Properties | • Patterns Collection<br>• Compositional References<br>• Decoys Selection<br>• Probabilistic Generation | • Distribution Balancing<br>• Type-Based Sampling<br>• Deduplication | • Functional Programs<br>• Entailment Relations<br>• Recursive Reachability | • Consistency<br>• Validity & Plausibility<br>• Distribution<br>• Grounding |

Hudson, Drew A., and Christopher D. Manning. "GQA: a new dataset for compositional question answering over real-world images." CVPR 2019

c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

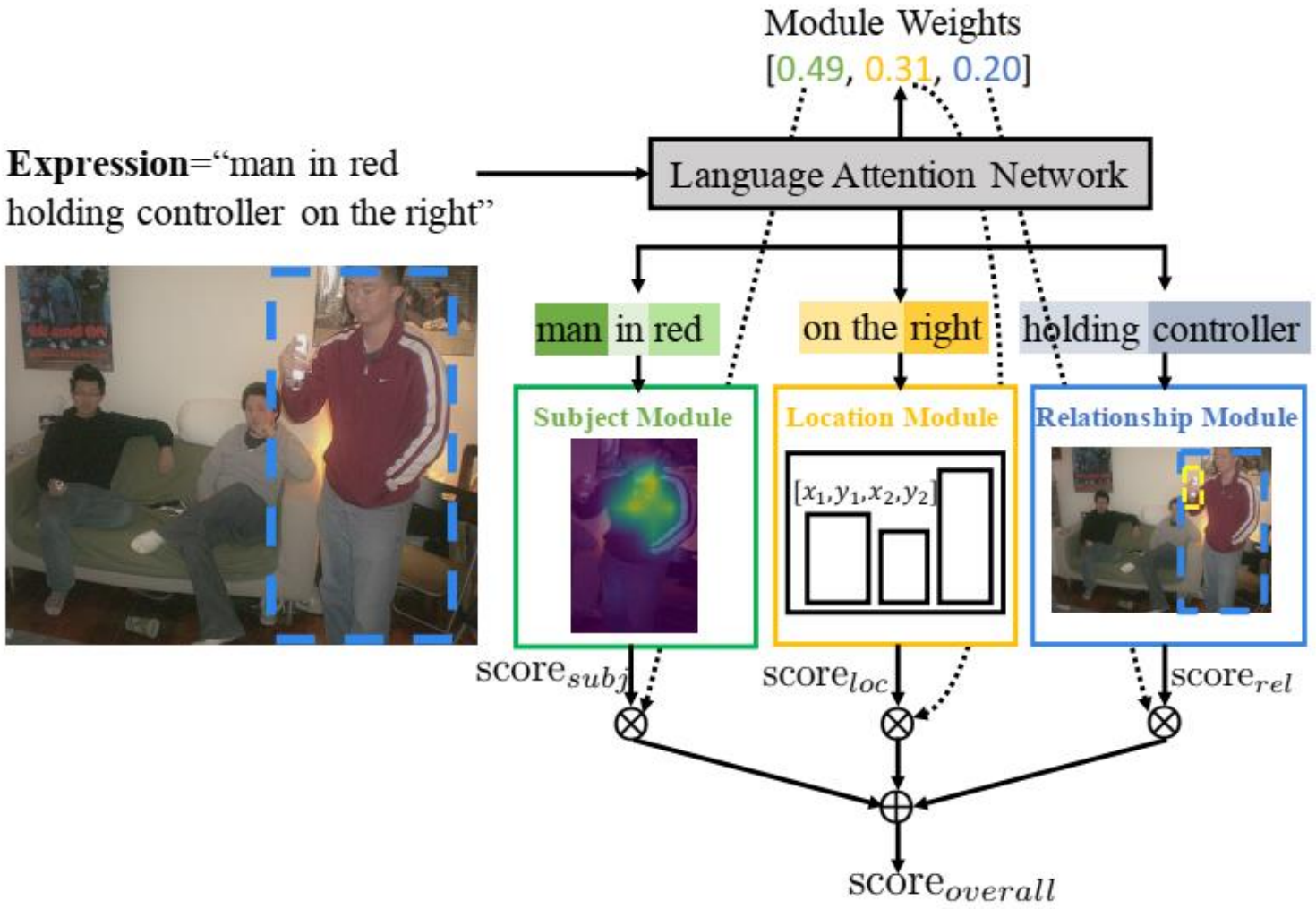
I chose a) because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
**d) [person3] is delivering food to the table, and she might not know whose order is whose.**




Zellers, Rowan, Yonatan Bisk, Ali Farhadi, and Yejin Choi. "From Recognition to Cognition: Visual Commonsense Reasoning." CVPR 2019

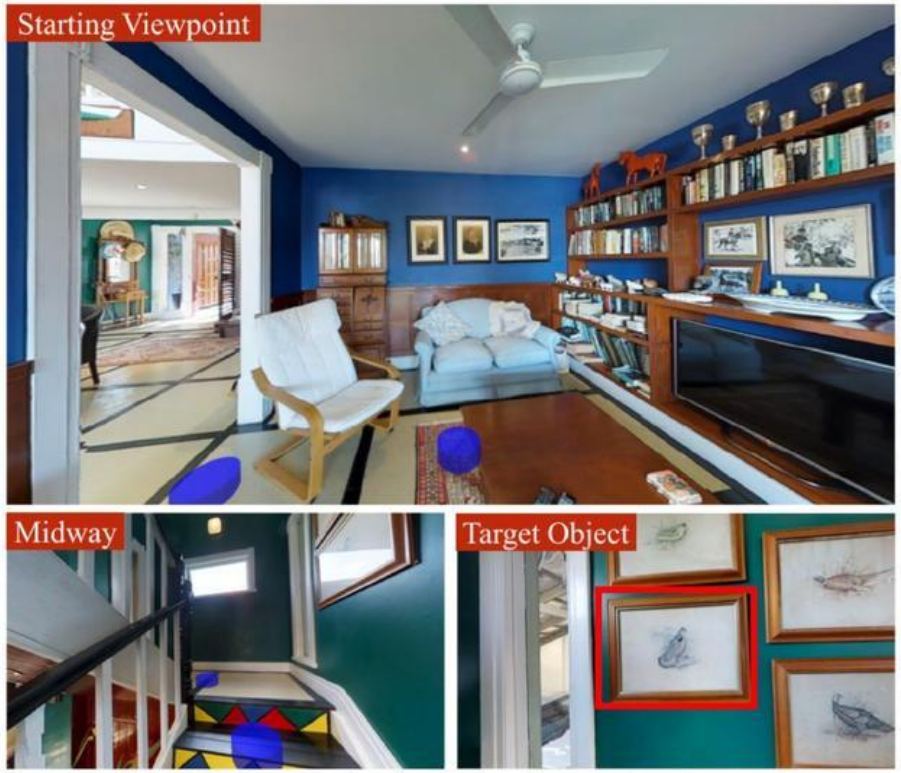
# Referring Expression




- RefCOCO 2014
- GuessWhat?! 2017
- CLEVR-Ref 2018
- Cityscapes-Ref 2018
- Talk2Car 2019
- CLEVR-Ref+ 2019

# Visual Navigation

- Vision-and-Language Navigation (VLN)
- Embodied Question Answering (EQA)
- REVERIE




Instruction: Bring me the bottom picture that is next to the top of stairs on level one.


- Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR 2018.
- Embodied question answering. CVPR 2018.
- Mattnet: Modular attention network for referring expression comprehension. CVPR 2018.
- Tactical rewind: Self-correction via backtracking in vision-and-language navigation. CVPR 2019.
- REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. CVPR 2020.

# Multimodal Machine Translation

A bird flies
over the water

Model

Ein Vogel fliegt
über das Wasser

# Aspects to be further addressed

- Evaluation
  - Evaluation metrics
  - Evaluation dataset
- Multimodal Fusion/Representation
  - interaction
- Unsupervised learning
  - generalization
- Reasoning/Interpretability

# Thank You!