

# AUTOMATED VIDEO DESCRIPTION AND VISUAL STORY ENTAILMENT

Natalie Parde  
[parde@uic.edu](mailto:parde@uic.edu)

CS 594: Language and Vision  
Spring 2019

# What is automated video description?

- The process of automatically describing video content by mapping sequences of image frames to sequences of words.

A dog is dancing next to another dog.



Add basil to the bowl, followed by arugala. Then squeeze lemon on top, and add water. Add olive oil after that, and blend. Remove the lid, and add two garlic cloves along with salt and pepper.



# What is visual story entailment?

The process of automatically predicting contextual stories for sequences of images.

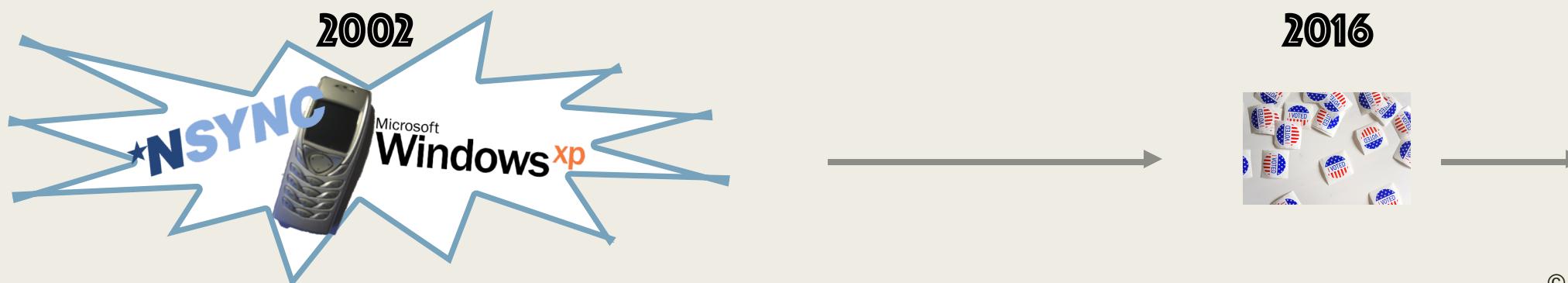
## 1 Story ≠ 5 Captions

	1	2	3	4	5
Desc-in-Isolation					
Desc-in-Sequence	A black frisbee is sitting on top of a roof.	A man is standing in the grass in front of the house kicking a soccer ball.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two soccer balls and a Frisbee are sitting on top of the roof top.
Story-in-Sequence	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

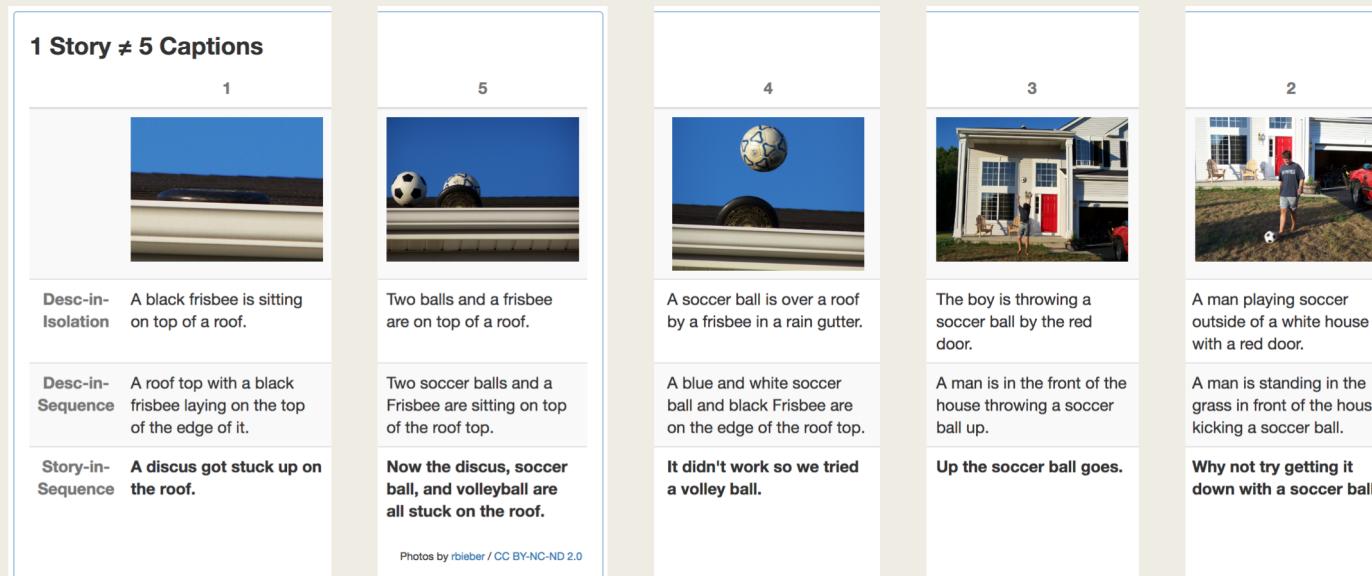
<http://www.visionandlanguage.net/VIST/>

# More recent than some other tasks we've discussed....

- Earliest work on automated video description:
  - Kojima, Atsuhiro, Takeshi Tamura, and Kunio Fukunaga. 2002. "Natural language description of human activities from video images based on concept hierarchy of actions." *International Journal of Computer Vision* 50.2, pages 171-184.
- Earliest work on visual storytelling:
  - Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233-1239. Association for Computational Linguistics.



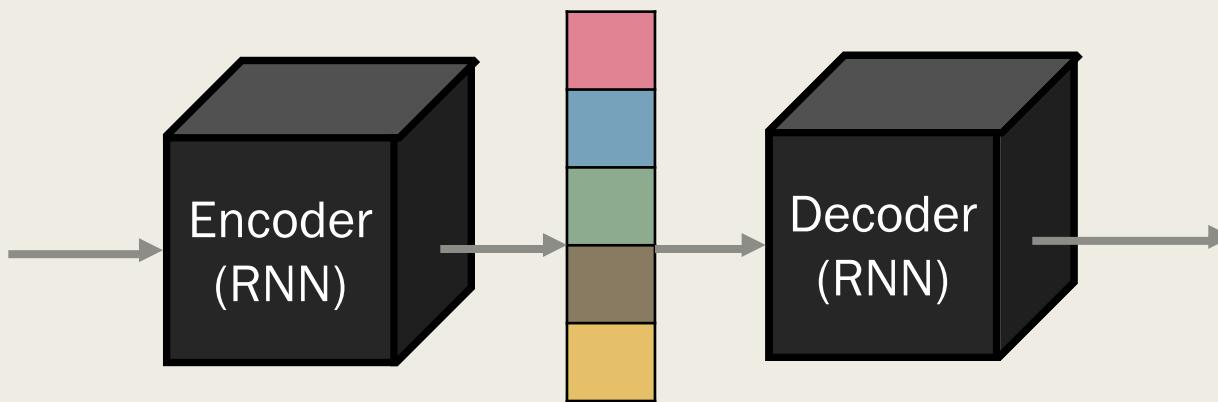
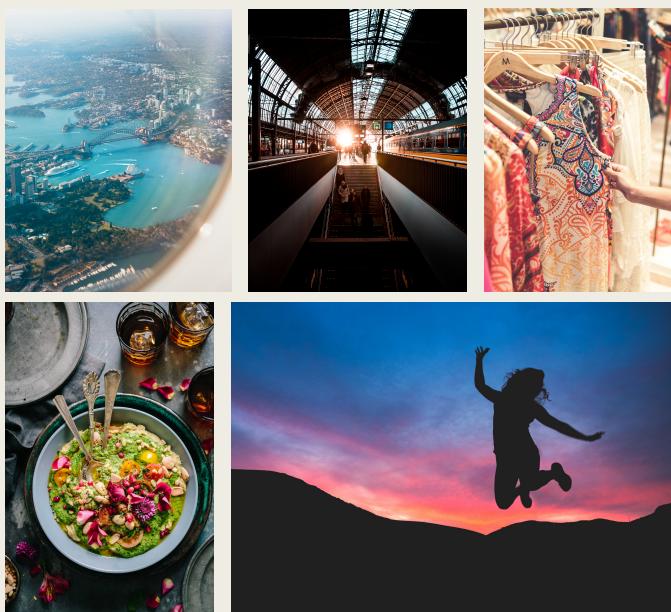
# Unifying theme: temporal sequence matters!



<http://www.visionandlanguage.net/VIST/>



# Typically framed as a sequence-to-sequence task.



I arrived at my destination late in the morning. Then I took a train downtown! I went shopping for awhile. After that I went out to dinner. I love traveling!

street lights in the background

a crowd of people

most people are facing a stage

people on the stage are holding musical instruments



a concert

outdoor venue at night

people are having fun

probably not electronic music

a performance is taking place

Key differences between static image captioning and visual story entailment?

- Greater reliance on *world knowledge* and temporal context
- Increased use of terms that are:
  - *Abstract*
  - *Dynamic*
- Oftentimes more *subjective*

# Common Evaluation Metrics

Similar to image captioning

METEOR, BLEU, CIDEr

Human Evaluation

- *Turing Test: Which sample was generated by a human/machine?*
- *Ratings for individual characteristics*
  - Focus
  - Structure and coherence
  - Level of detail

Real errors from recent visual storytelling models (things to watch out for when generating sequential stories and video descriptions)!

### Singular/Plural Disagreement

- *The resort was beautiful. The beach was nice. The beaches were amazing. The water was so calm. The food was delicious.*

### Absurdity

- *The kitchen was a lot of work. Here is a picture of a box. I had to take a picture of my work. We had to take a picture of the menu. I had a great time.*

### Point-of-View Inconsistency

- *I was so excited to be graduating today. He was very proud of his graduation. Graduation day is always a success. He was very proud of his accomplishments. He was very proud of his accomplishments.*

Real errors from recent visual storytelling models (things to watch out for when generating sequential stories and video descriptions)!

### Contradictions

- We went to the art gallery. We saw a lot of people there. **The streets were empty. The streets were full of people.** This is a picture of a woman.

### Repetitions

- *It was a beautiful day for a trip to the beach. We took a trip to the beach. We went to the beach.* The beach was beautiful. As *the sun went down, the sun went down.*

### Standard Image Captions

- **This is a picture of a street.** It was a long drive. There was a lot of damage to the side of the road. **This is a picture of a man.** After that we found a trail that was in the middle of the forest.

# Video Description Resources

- Datasets

- *Microsoft Video Description Corpus:* <https://www.microsoft.com/en-us/download/details.aspx?id=52422&from=https%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fdownloads%2F38cf15fd-b8df-477e-a4e4-a4680caa75af%2F>
- *Large-Scale Movie Description and Understanding Challenge Dataset:* <https://sites.google.com/site/describingmovies/lsmdc-2016/download>
- *ActivityNet Captions Dataset:* <https://cs.stanford.edu/people/ranjaykrishna/densevid/>

- Lecture

- *Generating Natural-Language Video Descriptions Using Text-Mined Knowledge, by Ray Mooney:* <https://youtu.be/0mlIN7K2sgU>

# Visual Story Entailment Resources

- Visual Storytelling Dataset:  
<http://www.visionandlanguage.net/VIST/>
- COMICS Dataset: <https://github.com/miyyer/comics>
- Visual Storytelling Challenge:  
<https://evalai.cloudcv.org/web/challenges/challenge-page/76/overview>

# Wrapping up....

- Overview of automated video description
- Overview of visual story entailment
- Early video description and visual story entailment work
- Differences with image captioning
- Common evaluation metrics
- Open problems in sequential vision + language work
- Resources