Perception
Image Captioning

Yezhou Yang, Ph.D.
Assistant Professor
Zhiyuan Fang, Teaching Assistant
Arizona State University

ASU Ira A. Fulton Schools of Engineering
Arizona State University

Problem Definition

F() = "A furry, grey kitten."

- **Image Captioning:**
  - Describing the image in a open form natural language sentence.

J. Bryner, “USDA Fed Cats and Dogs to Kittens, Alarming Watchdog Report Claims,” LiveScience, 19-Mar-2019. [Online]. Available: https://www.livescience.com/65030-usda-kitten-cannibalism-research.html. [Accessed: 10-May-2019].

Captioning Examples

A dog is running on the grass with a ball.

A young lady leans on a table in the kitchen.

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." Proceedings of the IEEE international conference on computer vision. 2015.

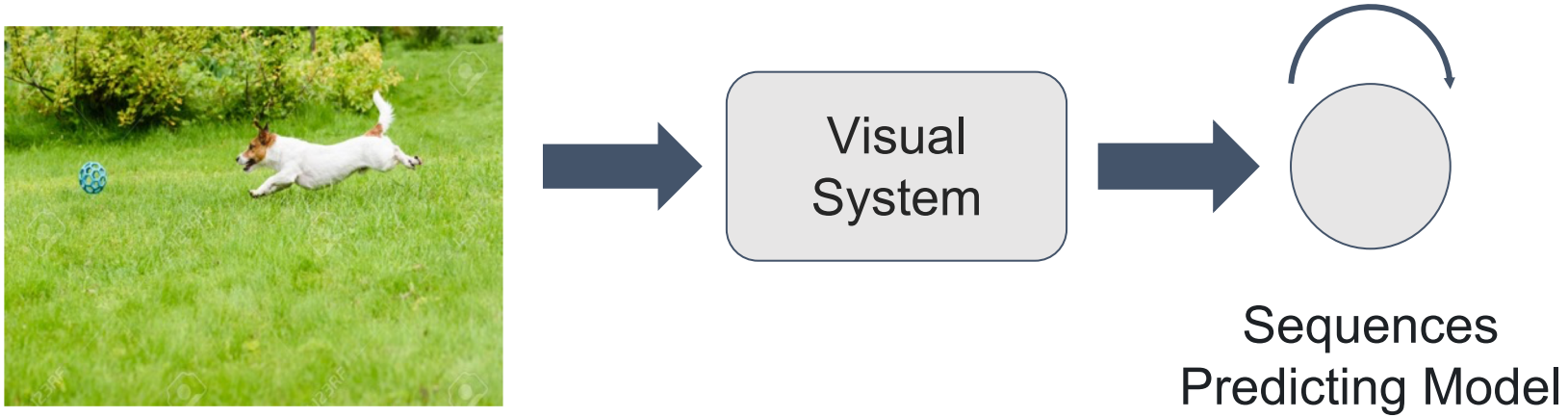
Captioning Examples

The men are playing a game of baseball on the field.

A city bus rides down a street in front of some buildings.

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." Proceedings of the IEEE international conference on computer vision. 2015.

General Workflow

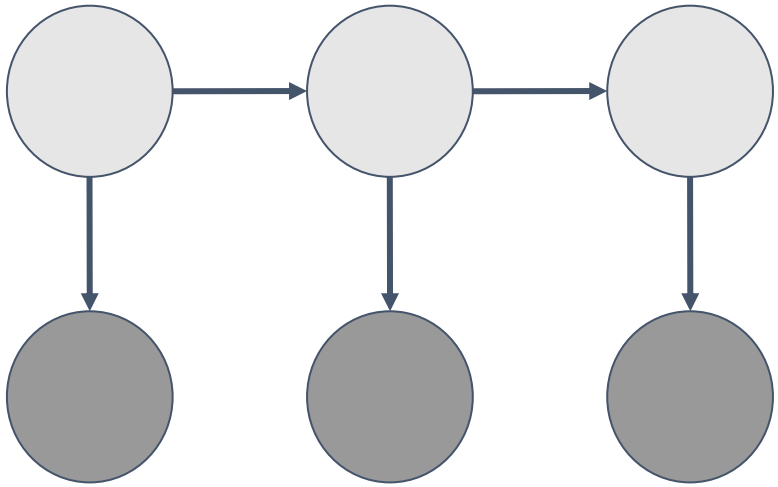


- **Image Captioning:**
  - State-of-the-art method combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image.

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." Proceedings of the IEEE international conference on computer vision. 2015.

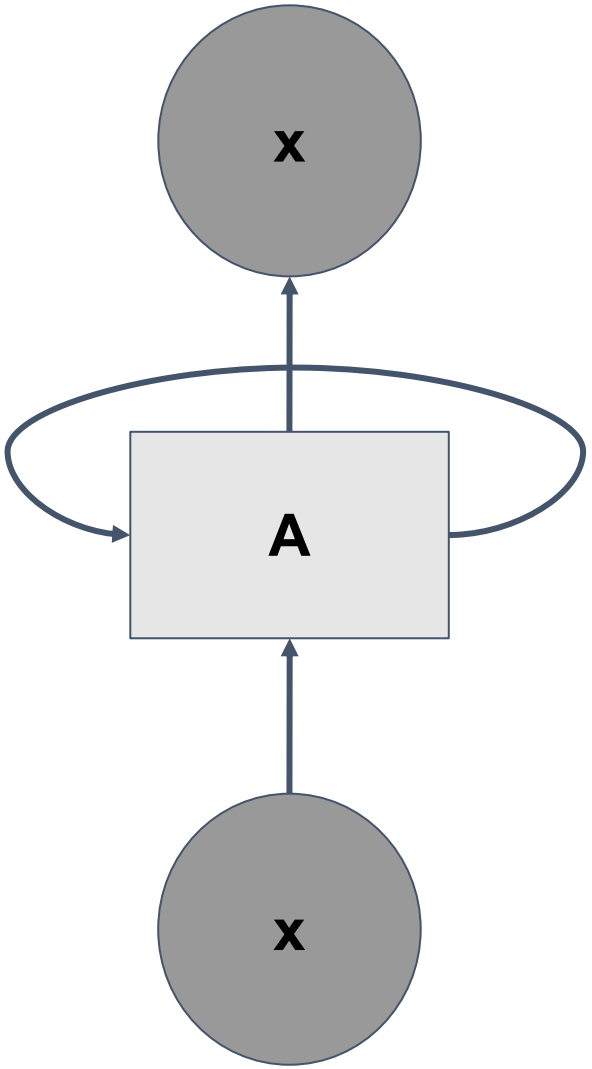
Relevant Methods

- To generate the languages, we are training a sequence predicting model:
  - P(next word | previous words)
- Hidden Markov Model (HMM):
  - Is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states.



Relevant Methods

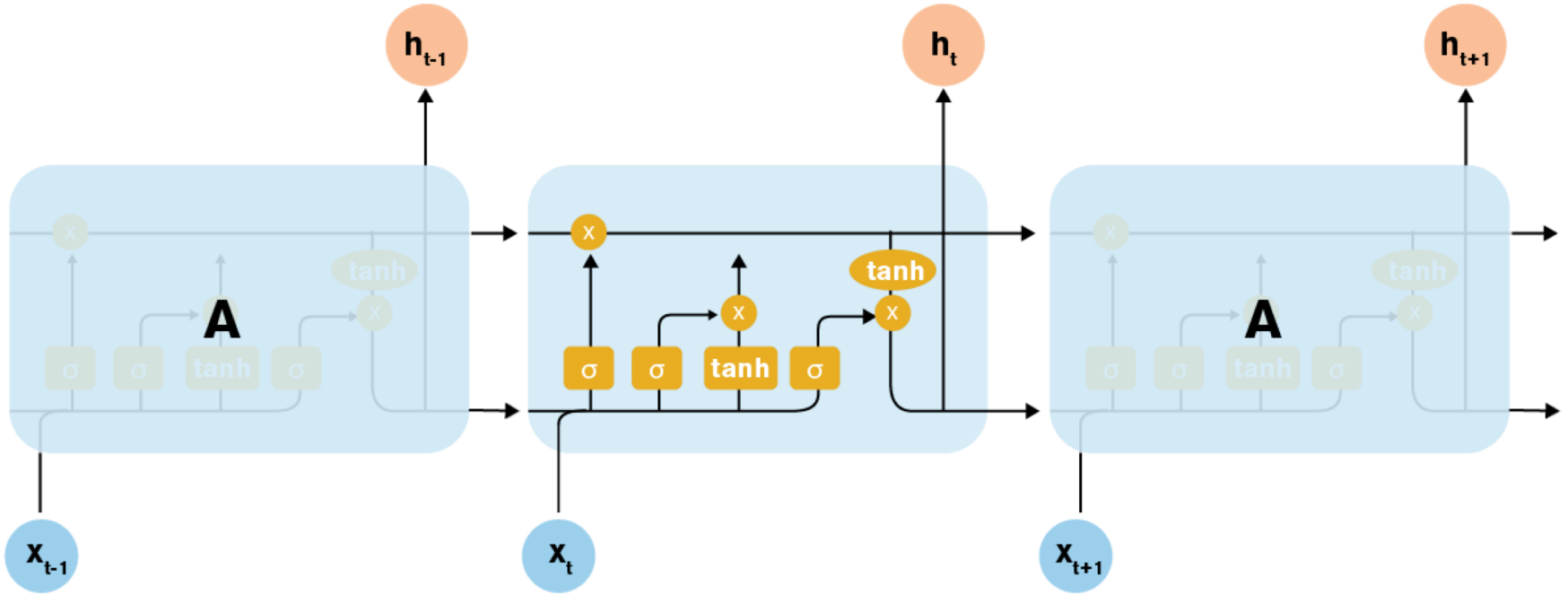
- To generate the languages, we are training a sequence predicting model:
  - P(next word | previous words)
- Recurrent Neural Network (RNN):
  - Is a class of artificial neural network where connections between nodes form a directed graph along a temporal sequence.



Relevant Methods

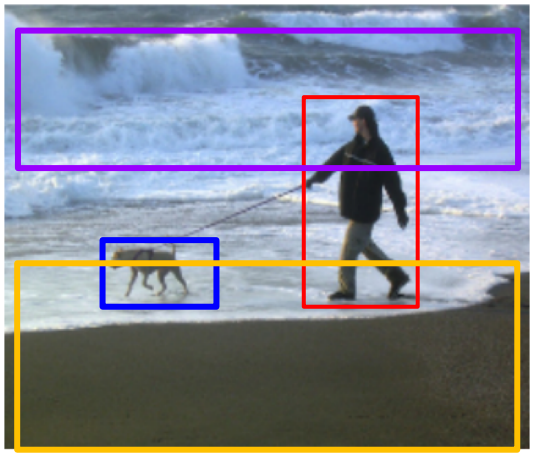
Long Short-term Memory (LSTM):

- Is a variation of RNN to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs.

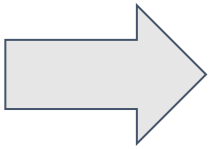


"Understanding LSTM Networks," Understanding LSTM Networks -- colah's blog. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed: 11-May-2019].

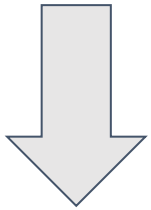
Relevant Methods



Visual Classifier



{ Person, Dog, Coat, Sea }

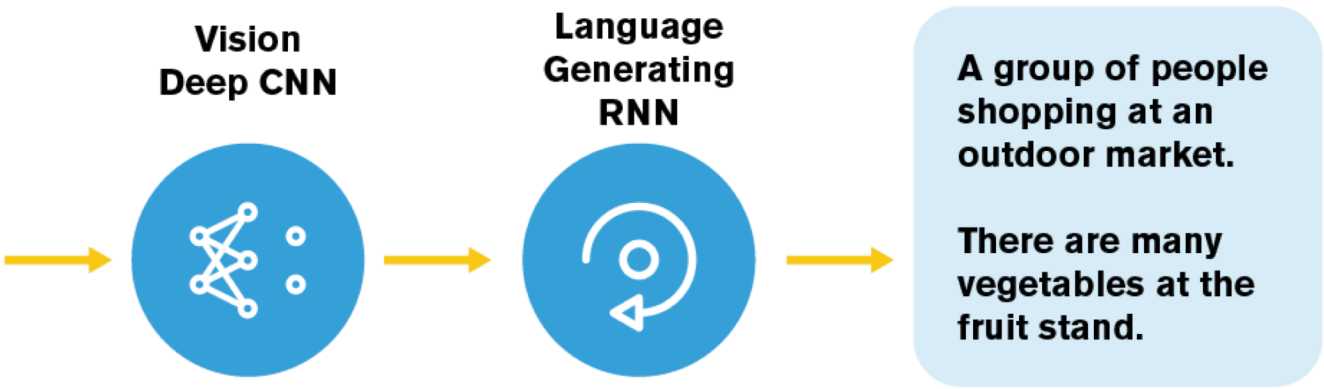


HMM

"A person is walking their dog on the coast."

Yang, Yezhou, et al. "Corpus-guided sentence generation of natural images." EMNLP, 2011.

Relevant Methods



Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Reading Materials

Reading Materials

- Image captioning with semantic attention
- Self-critical sequence training for image captioning
- Neural Networks and Deep Learning:Image Captioning
- Designing, Visualizing and Understanding Deep Neural Networks

You, Quanzeng, et al. "Image captioning with semantic attention." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
Rennie, Steven J., et al. "Self-critical sequence training for image captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
Neural Networks and Deep Learning:Image Captioning, Mike Mozer
Designing, Visualizing and Understanding Deep Neural Networks, John Canny

Reading Materials

Reading Materials

- Automated Image Captioning with ConvNets and Recurrent Nets
- Corpus-guided sentence generation of natural images
- Show and tell: A neural image caption generator
- An Introduction to Conditional Random Fields

Automated Image Captioning with ConvNets and Recurrent Nets, Andrej Karpathy, Fei-Fei Li
Yang, Yezhou, et al. "Corpus-guided sentence generation of natural images." EMNLP, 2011.
Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
An Introduction to Conditional Random Fields, Charles Sutton and Andrew McCallum