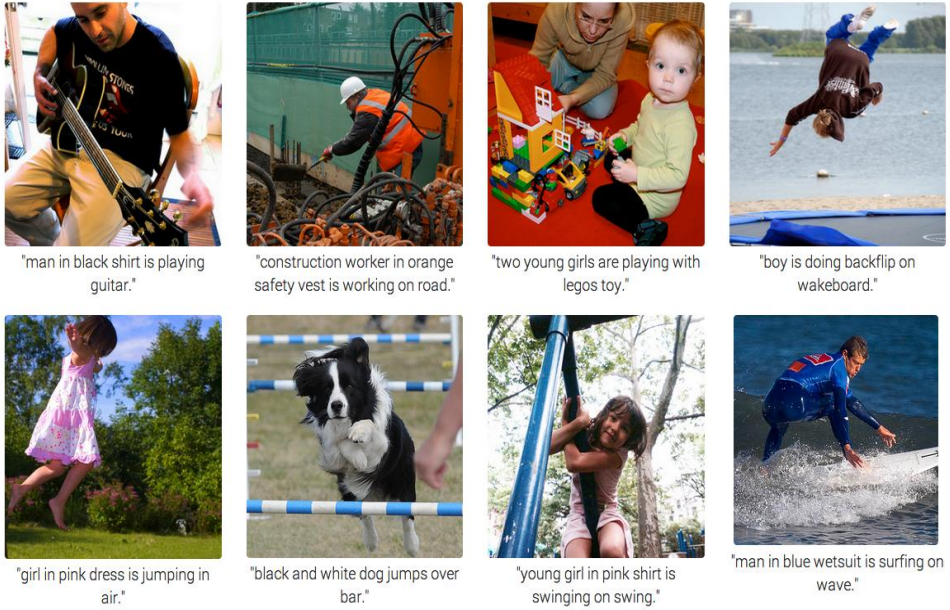
Neural network architectures for image captioning

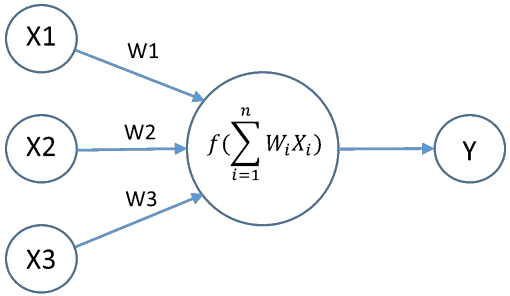
By Emily Kern

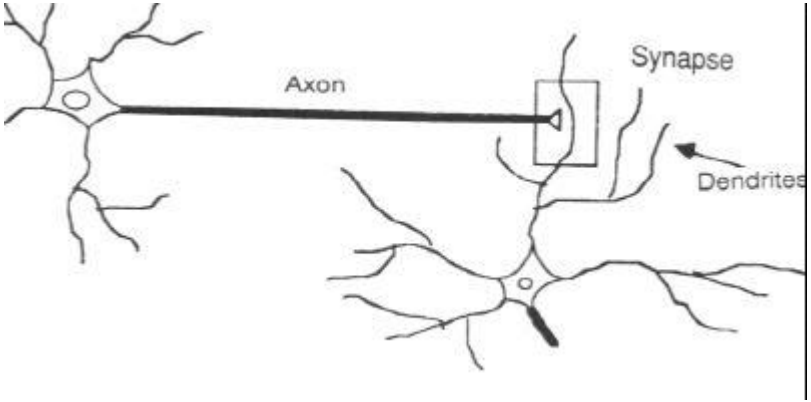
Given a set of images and accompanying human-generated captions, can we train a neural network to predict captions for new images?

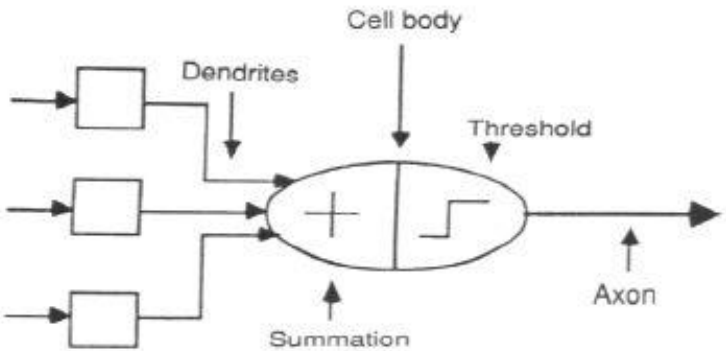


What is a neural network?

- A computer system modeled after the human brain
- There are many different architectures

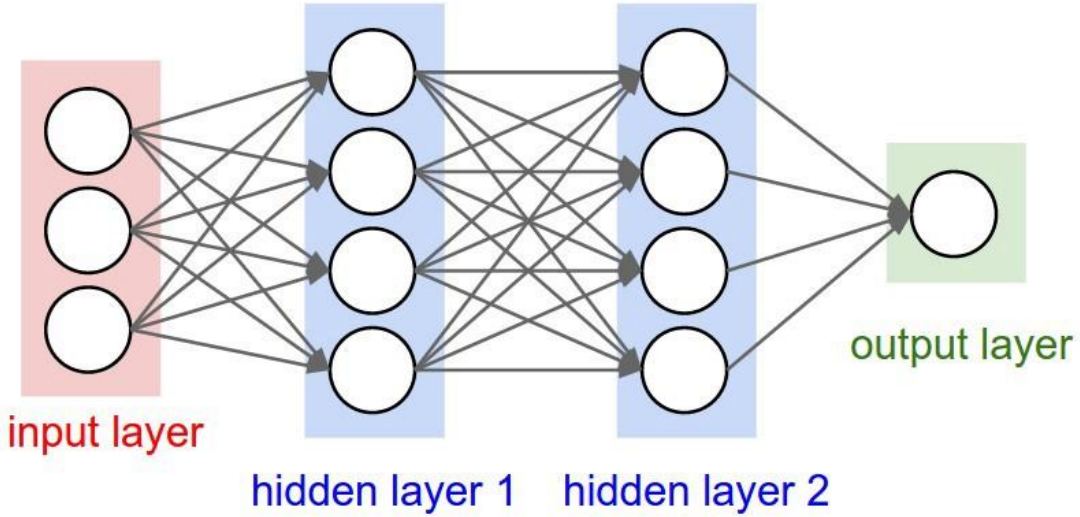






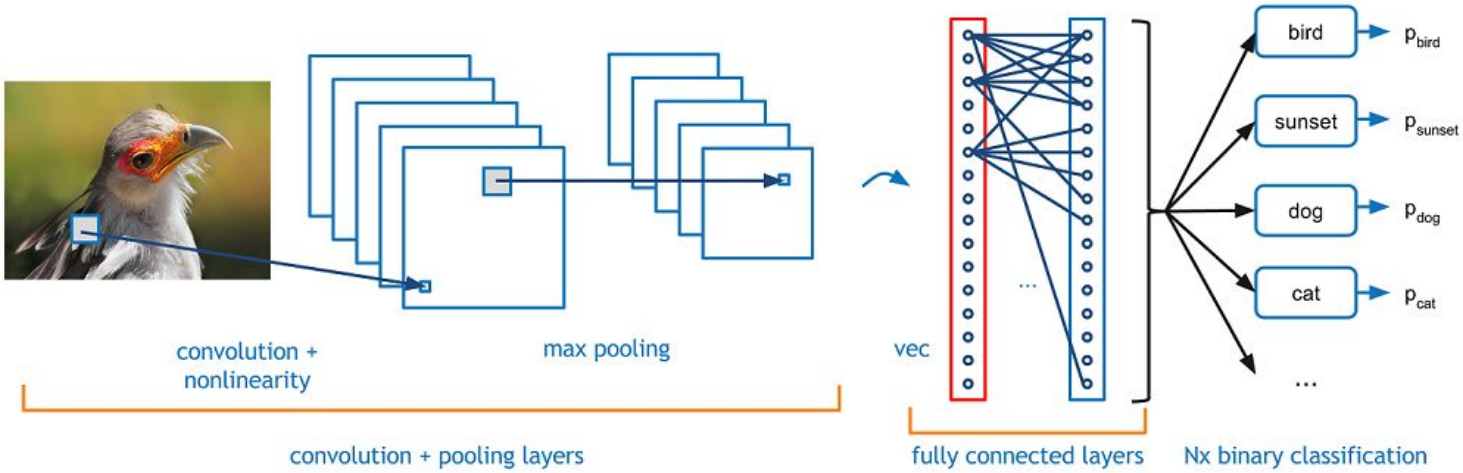
Feed-Forward

- The simplest type of neural network
- Architecture does not include any loops



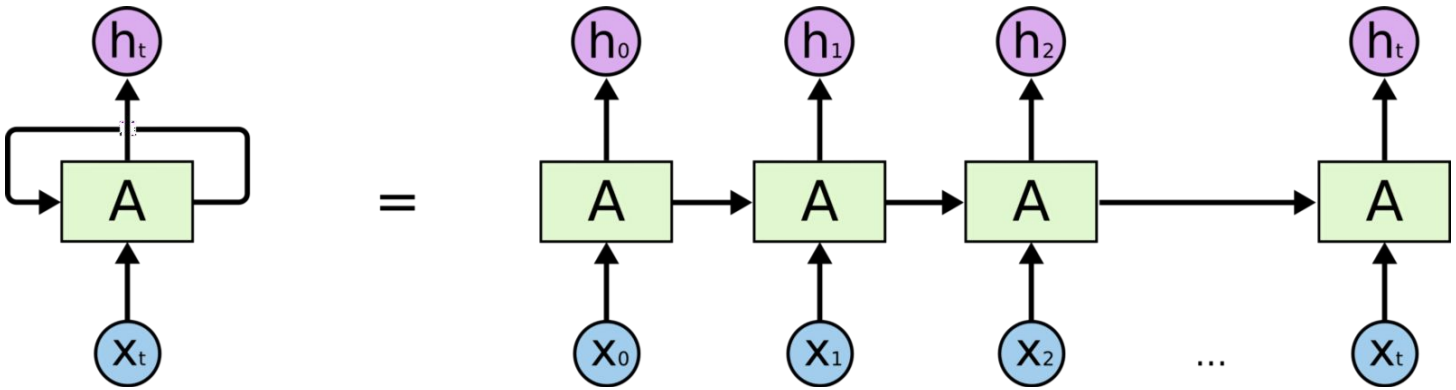
Convolutional (CNN)

- Good at object classification
- Given an image → checks pixel intensity (RGB values)
- Applies filters to understand higher-level features



Recurrent (RNN)

- Good at operating over a sequence of vectors (i.e. sentences, words)
- New state h_t dependent on previous state h_(t-1) and current input x_t
- Short-term memory
- Other implementations (i.e. LSTM, GRU)

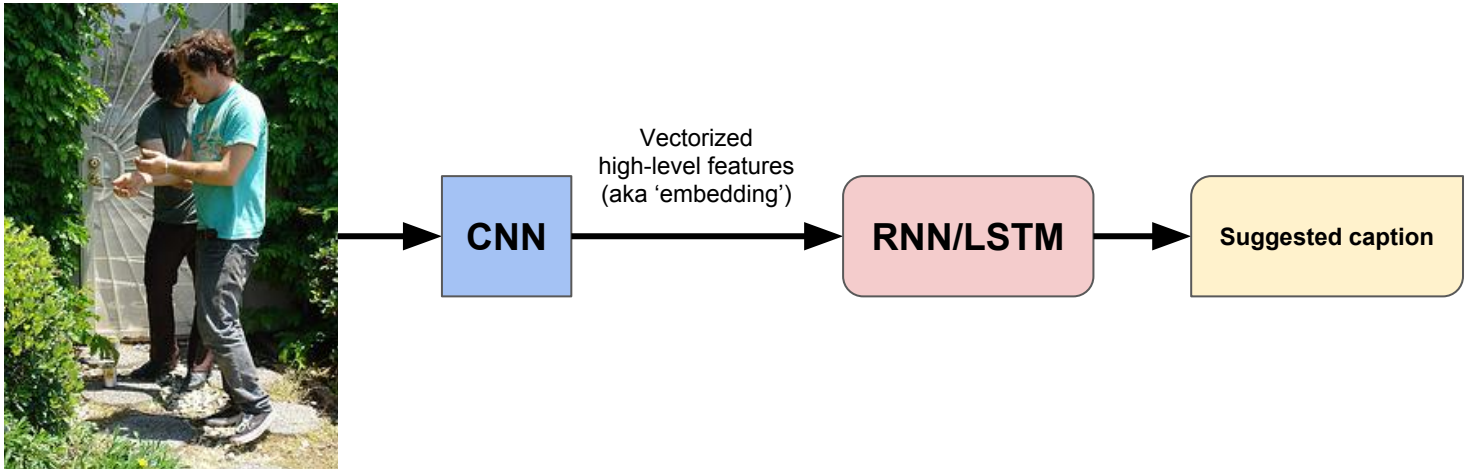


Research

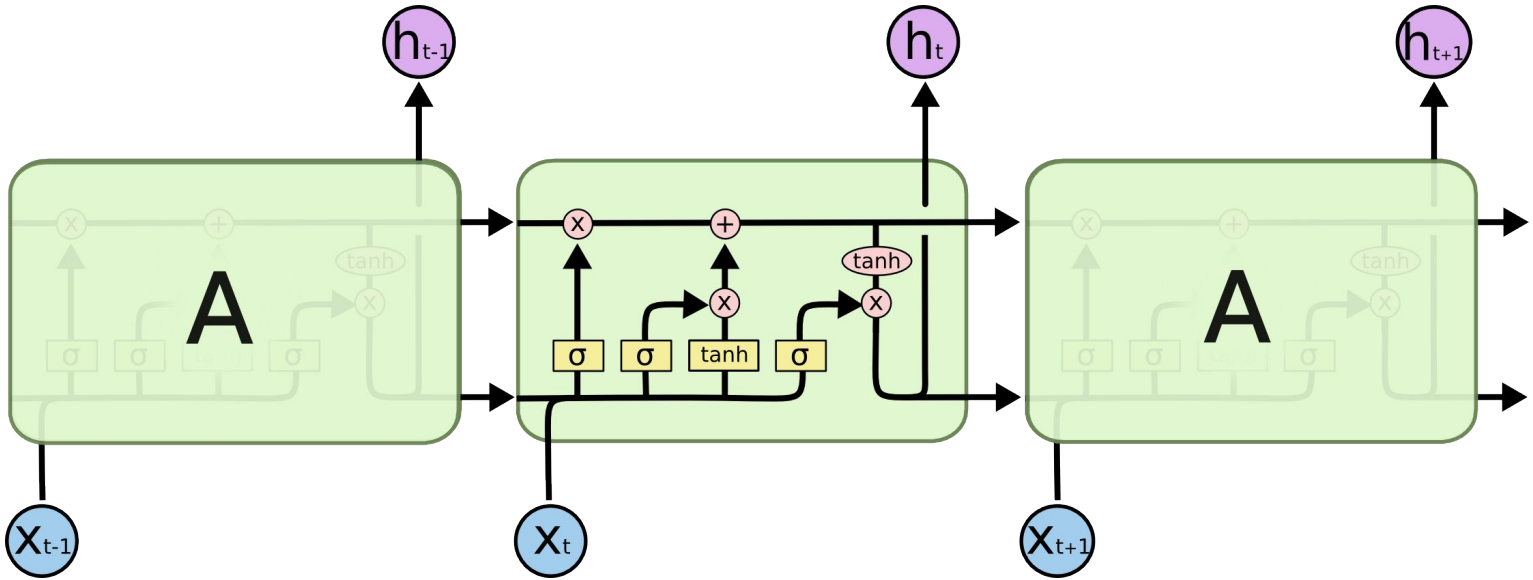
- Vinyals
  - Proposal for image captioning
- Karpathy, Fei-Fei
  - CNN + RNN/LSTM for image captioning
  - Uses Flickr8k and Flickr30 datasets (crowdsourced)

We use the Karpathy and Fei-Fei model as a base

- Encoder-decoder architecture
  - CNN encoder, RNN/LSTM decoder
- Supports flickr8k, and flickr30k



- Keep CNN encoder, use LSTM decoder
- LSTM slow, but better for captions



Flickr Datasets

{‘filename': '1000092795.jpg', 'imgid': 0, 'sentences': [{'tokens': ['two', 'young', 'guys', 'with', 'shaggy', 'hair', 'look', 'at', 'their', 'hands', 'while', 'hanging', 'out', 'in', 'the', 'yard'], 'raw': 'Two young guys with shaggy hair look at their hands while hanging out in the yard.', 'imgid': 0, 'sentid': 0}, {'tokens': ['two', 'young', 'white', 'males', 'are', 'outside', 'near', 'many', 'bushes'], 'raw': 'Two young, White males are outside near many bushes.', 'imgid': 0, 'sentid': 1}, {'tokens': ['two', 'men', 'in', 'green', 'shirts', 'are', 'standing', 'in', 'a', 'yard'], 'raw': 'Two men in green shirts are standing in a yard.', 'imgid': 0, 'sentid': 2}, {'tokens': ['a', 'man', 'in', 'a', 'blue', 'shirt', 'standing', 'in', 'a', 'garden'], 'raw': 'A man in a blue shirt standing in a garden.', 'imgid': 0, 'sentid': 3}, {'tokens': ['two', 'friends', 'enjoy', 'time', 'spent', 'together'], 'raw': 'Two friends enjoy time spent together.', 'imgid': 0, 'sentid': 4}], 'split': 'train', 'sentids': [0, 1, 2, 3, 4]}

Early iteration on flickr8k

Pretty good

a skateboarder is doing a trick on a ramp
logprob: -8.79

a man in a blue shirt is running on a track
logprob: -13.88

a man is riding a bicycle on a ramp
logprob: -12.86

a boy in a red shirt is jumping off a swing set
logprob: -15.39

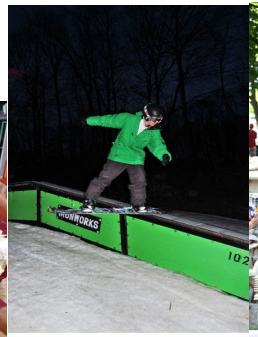
a brown dog is running through the snow
logprob: -5.08

a football player in a red uniform is running in the field
logprob: -12.09

a man in a wetsuit surfing
logprob: -7.06

Not so good

a basketball player in a red uniform is running with a football
logprob: -13.34



a man in a red shirt is riding a bike on a dirt road
logprob: -14.31

a man in a red shirt is standing on a rock with a black dog
logprob: -18.43

a man in a red uniform is running in a field
logprob: -14.21

a man in a yellow shirt is surfing on a wave
logprob: -12.23

a man in a black shirt is standing in front of a white building with a black and white flag
logprob: -20.82

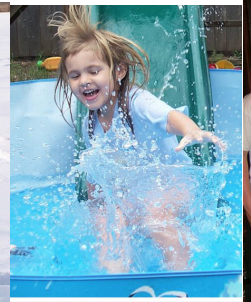
a man in a black shirt is standing in front of a large building
logprob: -15.89

Late iteration on flickr8k

Lots of men in red shirts, benches, and snow

a dog is running through the snow
logprob: -7.90

a dog is running through the snow
logprob: -7.93



a man in a red shirt is jumping up a rock wall
logprob: -18.65

a man in a black shirt is sitting on a bench
logprob: -14.41

a man in a red shirt is standing on a bench
logprob: -15.95

a man in a red shirt is standing on a bench
logprob: -15.96

a man in a red shirt is sitting on a bench
logprob: -14.85

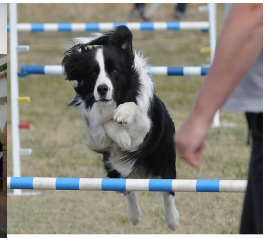
a man in a red shirt is sitting on a bench
logprob: -14.45

Early iteration on flickr30k

Pretty good

a football player in a red jersey is running
logprob: -11.09

a group of people are walking down a street
logprob: -8.57



a black and white dog is running through a field
logprob: -7.69

a man in a red shirt is riding a bike on a dirt path
logprob: -12.24

a group of people are in the water
logprob: -8.78

Not so good

a man in a blue shirt is standing in a river with a large fish
logprob: -21.63

a woman in a black shirt is holding a baby
logprob: -13.76

a young girl in a pink shirt is holding a toothbrush
logprob: -14.47

a man in a black shirt is jumping off a rock into a pool
logprob: -17.92

a woman in a red shirt is holding a red umbrella
logprob: -14.23

a young boy in a red shirt is looking at a toy
logprob: -16.70

Late iteration on flickr30k

Pretty good

a man in a blue shirt is playing a game
`logprob: -14.45`

a man in a white shirt is playing a guitar
`logprob: -11.25`

a man in a blue shirt is riding a bicycle
`logprob: -12.69`

a group of people are playing in a field
`logprob: -11.52`

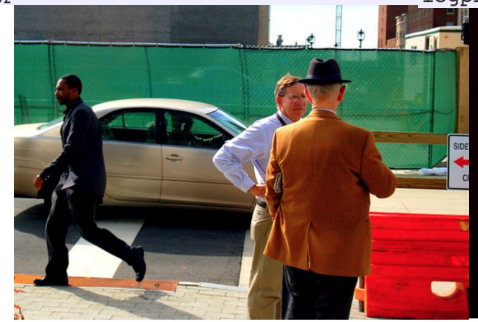
a baseball player is playing the ball
`logprob: -10.49`

Not so good

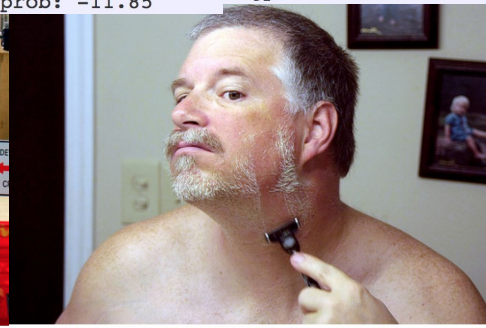
a group of people are sitting on a bench
`logprob: -9.89`

a man in a black shirt is playing a guitar
`logprob: -11.85`

a man in a blue shirt is standing on a rock
`logprob: -16.36`



a man in a blue shirt is playing a guitar
`logprob: -11.69`



a man in a black shirt is sitting on a bench
`logprob: -13.80`

Flickr30k

Early

a man in a blue shirt is jumping over a rock
`logprob: -16.93`

Late

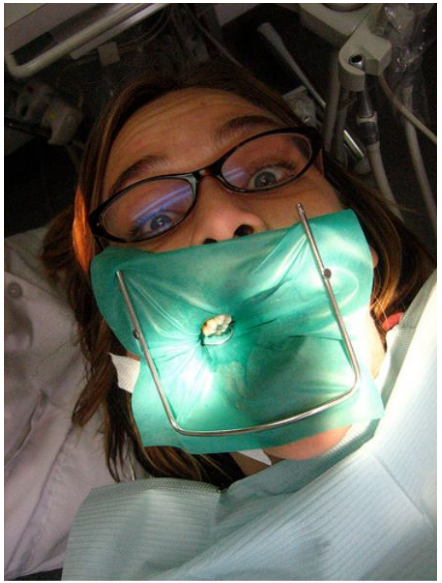
a man in a black shirt is jumping off a diving board
`logprob: -14.60`

Flickr30k

Early

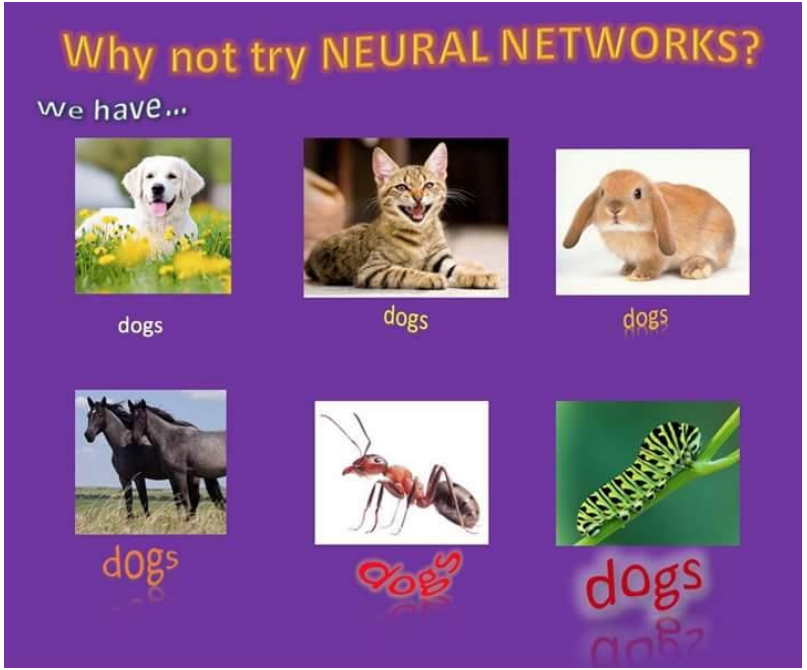
a young boy in a red shirt is looking at a toy
`logprob: -16.70`

Late



a man in a black shirt is sitting on a bench
`logprob: -13.66`

- Later iterations suffered from word biases and repeated captions
- Captions were coherent, if questionable at times
- Captions seemed more accurate/confident when less detailed



- Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
- M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899
- Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, ICC
- Peter Young, Alice Lai, Micah Hodosh and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics, 2(Feb):67-78, 2014.V, 2015.

- **Acknowledgements: Kristina Striegnitz, David Frey, CROCHET**