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
Two young guys with shaggy hair look at their hands while hanging out in the yard.

Two young, White males are outside near many bushes.

Two men in green shirts are standing in a yard.

A man in a blue shirt standing in a garden.

Two friends enjoy time spent together.
Early iteration on flickr8k

Pretty good

Not so good
Late iteration on flickr8k

Lots of men in red shirts, benches, and snow
Early iteration on flickr30k

Pretty good

Not so good
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

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 group of people are playing in a field
  logprob: -11.53
- A baseball player is playing the ball
  logprob: -10.49
- A man in a blue shirt is playing a guitar
  logprob: -11.69
- A man in a blue shirt is sitting on a bench
  logprob: -13.80
a man in a blue shirt is jumping over a rock
logprob: -16.93

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

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

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

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

Acknowledgements: Kristina Striegnitz, David Frey, CROCHET