Show and Tell: A Neural Image Caption Generator

Vinyals et al. (Google)

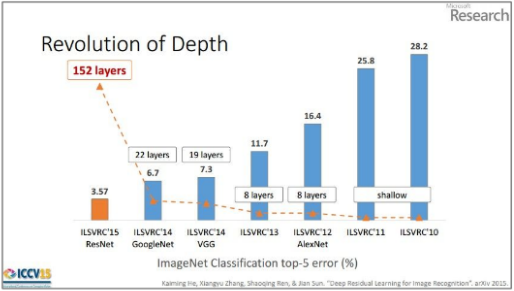
The IEEE Conference on Computer Vision and Pattern Recognition, 2015

The Problem

- **Image Caption Generation**
- Automatically describe content of an image
- Image $\rightarrow$ Natural Language
- Computer Vision + NLP
- Much more difficult than image classification/recognition

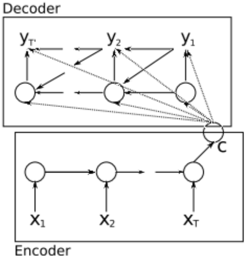
Background

- Success in image classification/recognition
- Close to human level performance
- Deep CNN's, Big Datasets
- Image to fixed length vector



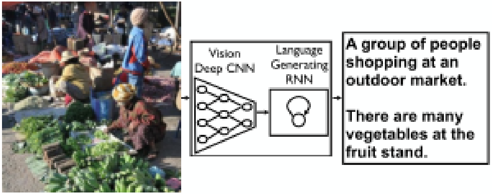
Background

- Machine Translation
- Language generating RNN's
- Decoder-Encoder framework
- Maximize likelihood of target sentence



Idea

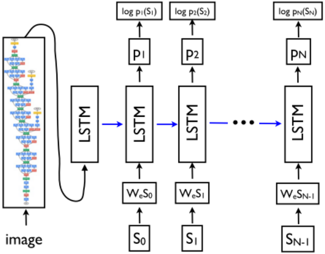
- Combine Vision CNN with Language RNN
- Deep CNN as encoder
- Language Generating RNN as decoder
- End to end model $I \to S$
- Maximize $p(S|I)$



The Model

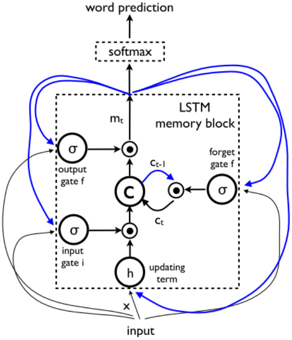
Neural Image Caption (NIC)

- CNN: 22 layer GoogleNet
- LSTM for modeling
  $\log \mathrm{p}(S|I) = \sum_{t=0}^{N} \log \mathrm{p}(S_t|I, S_0, \ldots, S_{t-1})$
- Word embedding W_e



Language LSTM

- Predicts next word in sentence
- Memory cell for longer memory
- S_t one-hot vectors + START/END token
- $x_{-1} = \text{CNN}(I)$, $x_t = W_e S_t$, $p_{t+1} = \text{LSTM}(x_t)$



(a)

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\
m_t &= o_t \odot c_t \\
p_{t+1} &= \text{Softmax}(m_t)
\end{aligned}
$$

(b)

Training

- Loss function $\mathrm{L}(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$
- CNN pre-trained on ImageNet
- Minimize w.r.t. LSTM parameters, W_e and CNN top layer
- SGD on mini-batches
- Dropout and ensembling
- 512 dimensional embedding

Generation

- Give $x_{-1} = \text{CNN}(\text{I})$
- $x_0 = W_e S_0$, S_0 START token
- Sample word S_1
- Feed $W_e S_1$ to LSTM
- BeamSearch, beam size 20

Results

- MSCOCO dataset: 80k train, 40k eval and test
- 5 human made captions per image
- M1-M5 human judgements

	M1	M2	M3	M4	M5
Human[5]	0.638	0.675	4.836	3.428	0.352
Google[4]	0.273	0.317	4.107	2.742	0.233
MSR[11]	0.268	0.322	4.137	2.662	0.234

Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Table 1. Scores on the MSCOCO development set.

Results


A person riding a motorcycle on a dirt road.
Two dogs play in the grass.
A skateboarder does a trick on a ramp.
A dog is jumping to catch a frisbee.
A group of young people playing a game of frisbee.
Two hockey players are fighting over the puck.
A little girl in a pink hat is blowing bubbles.
A refrigerator filled with lots of food and drinks.
A herd of elephants walking across a dry grass field.
A close up of a cat laying on a couch.
A red motorcycle parked on the side of the road.
A yellow school bus parked in a parking lot.
Describes without errors
Describes with minor errors
Somewhat related to the image
Unrelated to the image

Results

- Improved Flickr8k, Flickr30k, PASCAL BLEU scores
- Need better evaluation metrics
- 80% of top-1 in training set
- 50% of top-15 in training set
- Similiar diversity as human captions

A man throwing a frisbee in a park. **A man holding a frisbee in his hand.** **A man standing in the grass with a frisbee.**
A close up of a sandwich on a plate. A close up of a plate of food with french fries. A white plate topped with a cut in half sandwich.
A display case filled with lots of donuts. **A display case filled with lots of cakes.** **A bakery display case filled with lots of donuts.**

Results

- Trained word embeddings W_e

Word	Neighbors
car	van, cab, suv, vehicule, jeep
boy	toddler, gentleman, daughter, son
street	road, streets, highway, freeway
horse	pony, donkey, pig, goat, mule
computer	computers, pc, crt, chip, compute

- Captures semantics from the language data
- Independent of vocubulary size

Summary

- End-to-end model (Encoder-Decoder)
- Vision CNN + Language generating RNN
- Maximize likelihood of S given I
- State of the art results on major datasets
- Datasets are limiting: Unsupervised approaches?