Image Captioning with Show and Tell Model

Gizem Tabak (tabak2)
Safa Messaoud (messaou2)

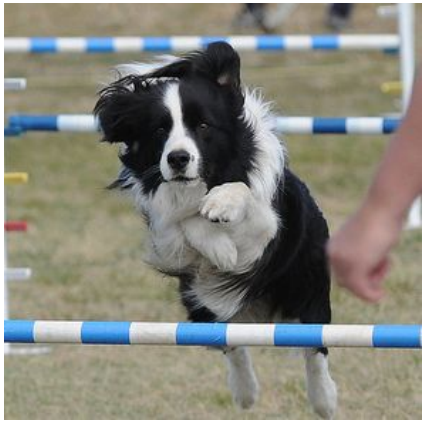
Ankit Rai (rai5)
Tarek Elgamal (telgama2)

Image Captioning

“Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy et al

Image Captioning Examples

“Girl in pink dress is jumping in the air.”



“Black and white dog jumping over the bar.”

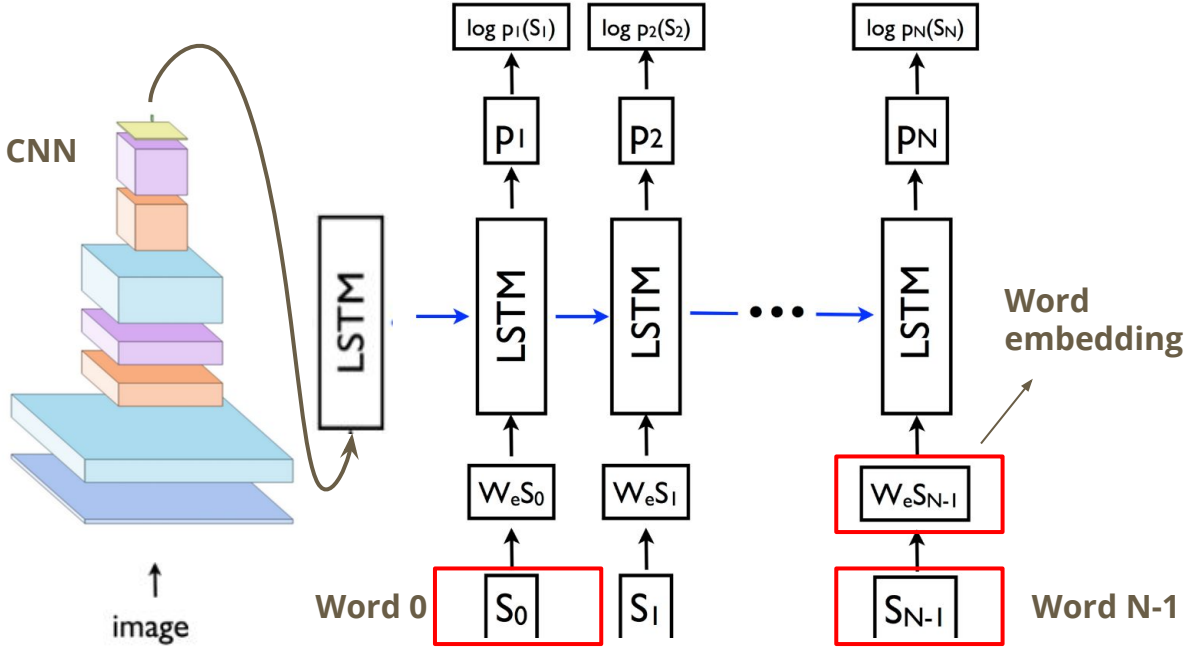
“Man in a blue wetsuit is surfing on a wave.”

Caption typically conveys: Main character, color, activity

“Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy et al.

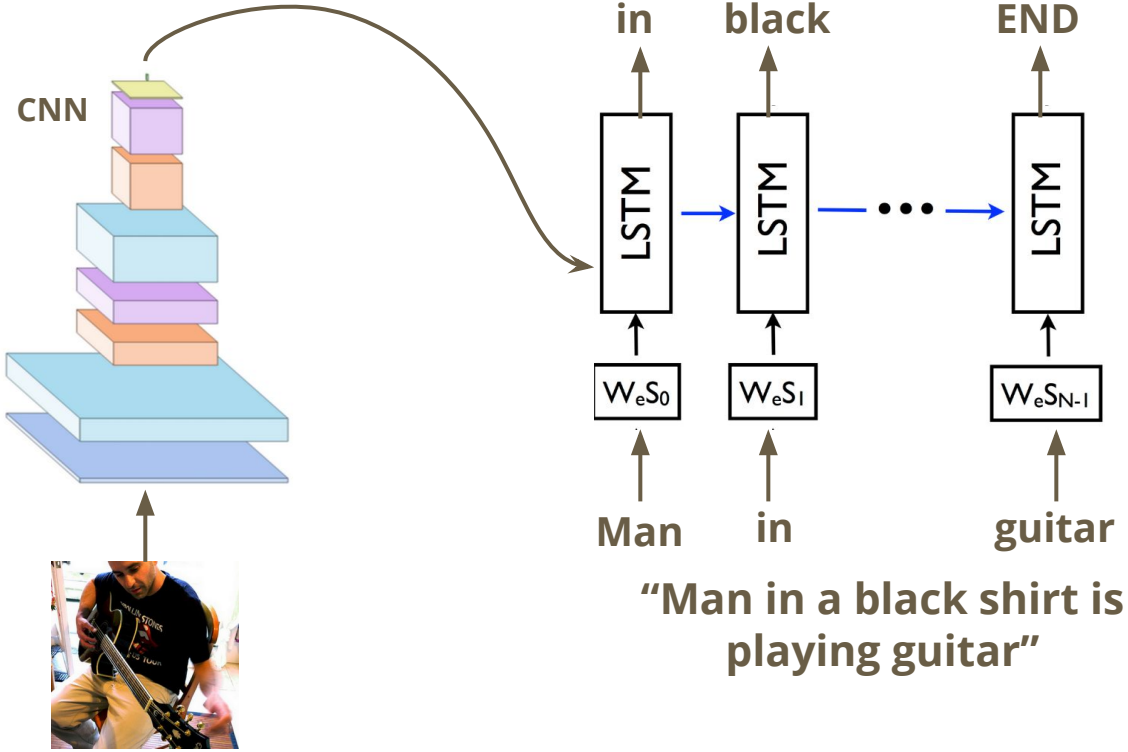
Show and Tell Model

- Pre-trained CNN (Inception) on ImageNet
- One LSTM unit
- Developed by Google
- Winner of MSCoCo image Captioning challenge 2015



Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE Conference on Computer*

Running Example



Implementation Overview

- Show and Tell open-sourced in Tensorflow
- Flickr8K* instead of MSCOCO
- Different RNN units
  - LSTM, GRU, Basic RNN
- Model training parameters
  - Learning rate, optimizer, dropout, decay, gradient clipping, batch size, RNN nodes
- Evaluation on validation set
  - BLEU score
  - Log-probabilities
- Training best model for longer time
- Merging it with txt2im (text to image)

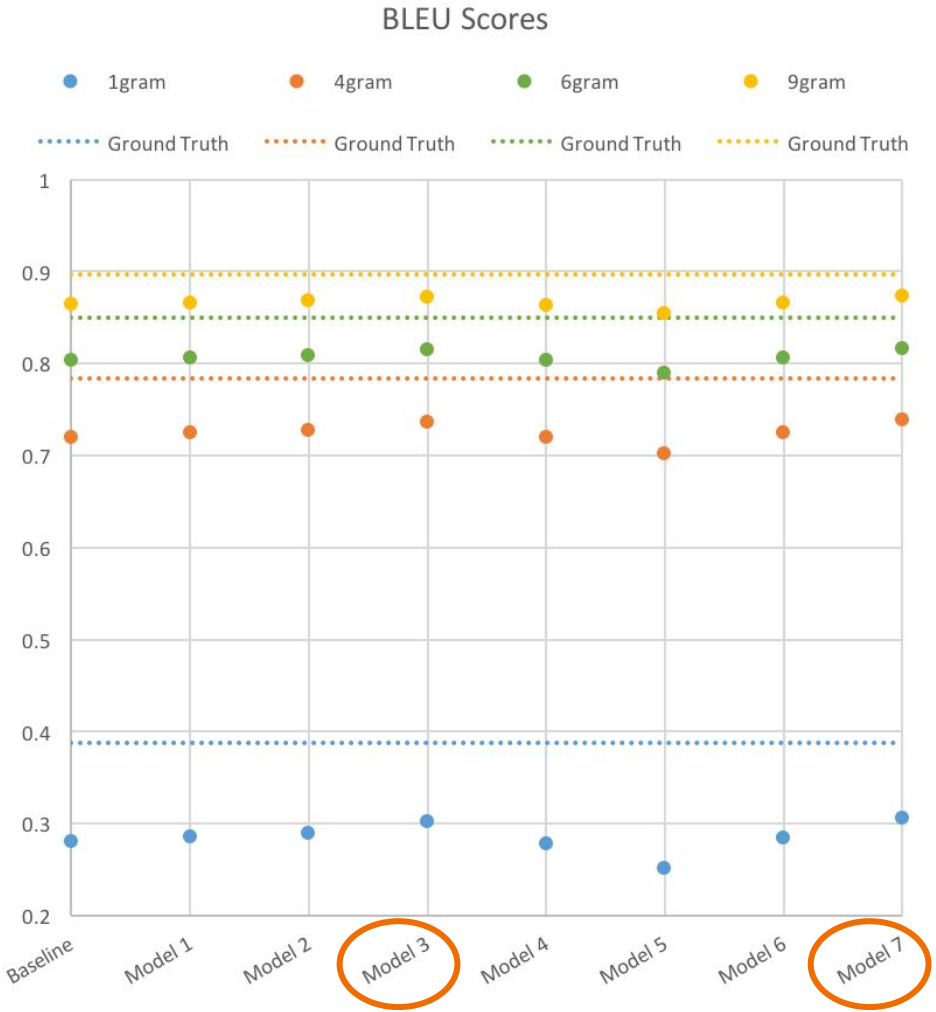
*Source (Flickr8K Dataset)- M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899

Results

	Dropout	Learning Rate	Sequential Model (Nodes)	Optimizer (Batch Size)	Training Loss
Baseline	0.3	1	LSTM (512)	SGD (32)	2.4651
Model 1	0.5	3	LSTM (512)	SGD (32)	2.5575
Model 2	0.5	3	GRU	SGD (32)	2.3606
Model 3	0.3	0.01 Decay 0.7	LSTM (512)	Adam	2.7263
Model 4	**0**	**1**	**LSTM (512)**	**SGD (32)**	**1.6204**
Model 5	0.5	0.5	RNN	SGD (16)	3.2560
Model 6	0.3	1	LSTM (256)	SGD (32)	2.5599
Model 7	0.3	0.01 Decay 0.9	LSTM (512)	Adam	2.5544

Results

	Dropout	Learning Rate	Sequential Model (Nodes)	Optimizer (Batch Size)
Baseline	0.3	1	LSTM (512)	SGD (32)
Model 1	0.5	3	LSTM (512)	SGD (32)
Model 2	0.5	3	GRU	SGD (32)
Model 3	0.3	0.01 Decay 0.7	LSTM	Adam
Model 4	0	1	LSTM	SGD (32)
Model 5	0.5	0.5	RNN	SGD (16)
Model 6	0.3	1	LSTM (256)	SGD (32)
Model 7	0.3	0.01 Decay 0.9	LSTM	Adam



Results

Baseline Model: A skateboarder jumps off a ramp.
Model 1: A man in a blue shirt and jeans is jumping off a ramp.
Model 2: A boy in a blue shirt and jeans is jumping on a skateboard.
Model 3: A young boy in a red shirt is playing in the water.
Model 4: A man does a skateboard trick in midair.
Model 5: A boy is doing a trick on a skateboard.
Model 6: A man on a skateboard jumps over a ramp.
Model 7: A man in a black shirt is standing on a snowy mountain.

Bicyclist is jumping on ramp covered with graffiti.

Results - Model 3

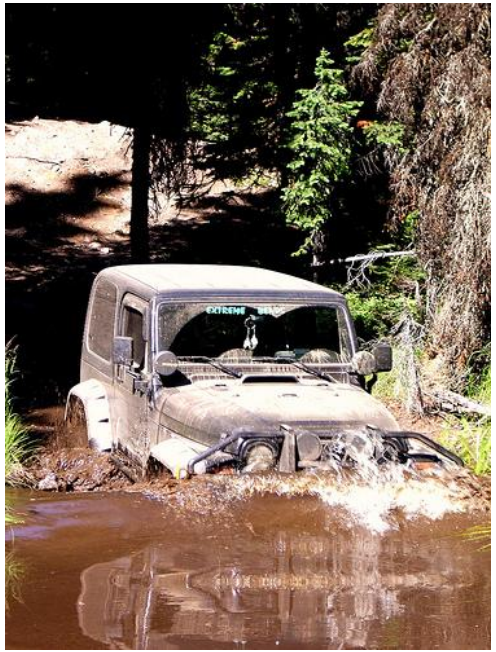
“A young boy in a red shirt is playing in the water.”

Results - Model 3

SAME CAPTION FOR ALL IMAGES

Results - Model 7

“A man in a black shirt is standing on a snowy mountain.”

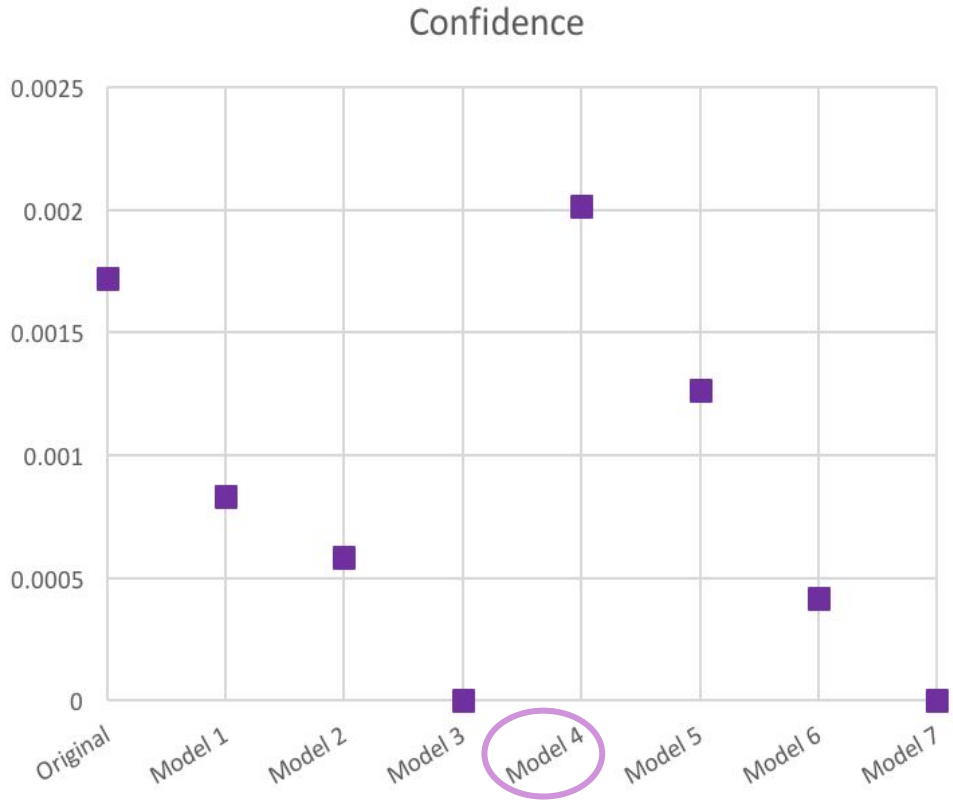


Results

- Even though BLEU metric gave high evaluation score, some of the models are not properly trained.
- BLEU score is not representative for sentence-level comparison.
  - It depends on caption length.
- Log-probabilities of captions (confidence)

Results

	Training Loss	BLEU Score	Probability (Confidence)
Baseline	2.4651	0.280	0.001722
Model 1	2.5575	0.286	0.000833
Model 2	2.3606	0.290	0.000583
Model 3	2.7263	0.303	0
Model 4	**1.6204**	**0.279**	**0.002017**
Model 5	3.2560	0.251	0.001268
Model 6	2.5599	0.285	0.000415
Model 7	2.5544	0.306	0



Results

	Dropout	Learning Rate	Sequential Model (Nodes)	Optimizer (Batch Size)	Training Loss	BLEU Score	Probability (Confidence)
Baseline	0.3	1	LSTM (512)	SGD (32)	2.4651	0.280	0.001722
Model 1	0.5	3	LSTM (512)	SGD (32)	2.5575	0.286	0.000833
Model 2	0.5	3	GRU	SGD (32)	2.3606	0.290	0.000583
Model 3	0.3	0.01 Decay 0.7	LSTM (512)	Adam	2.7263	0.303	0
Model 4	0	1	LSTM (512)	SGD (32)	1.6204	0.279	0.002017
Model 4 (50k)	**0**	**1**	**LSTM (512)**	**SGD (32)**	**2.3384**	**0.288**	**0.002147**
Model 5	0.5	0.5	RNN	SGD (16)	3.2560	0.251	0.001268
Model 6	0.3	1	LSTM (256)	SGD (32)	2.5599	0.285	0.000415
Model 7	0.3	0.01 Decay 0.9	LSTM (512)	Adam	2.5544	0.306	0

All models are trained with 15k iterations except Model 4 (50k).

Results - Model 4

GOOD CAPTIONS

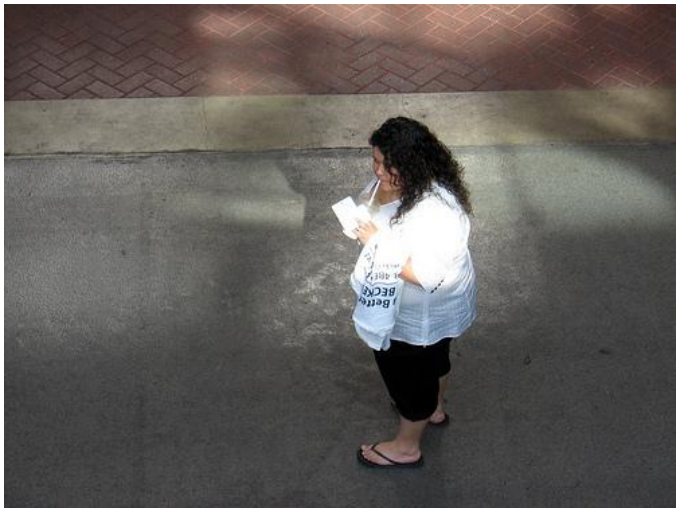
“A dog swimming in water.”
p = 0.019199

“Two men play basketball.”
p=0.050603

“The surfer is riding a wave in the ocean.”
p=0.028798

Results - Model 4

BAD CAPTIONS



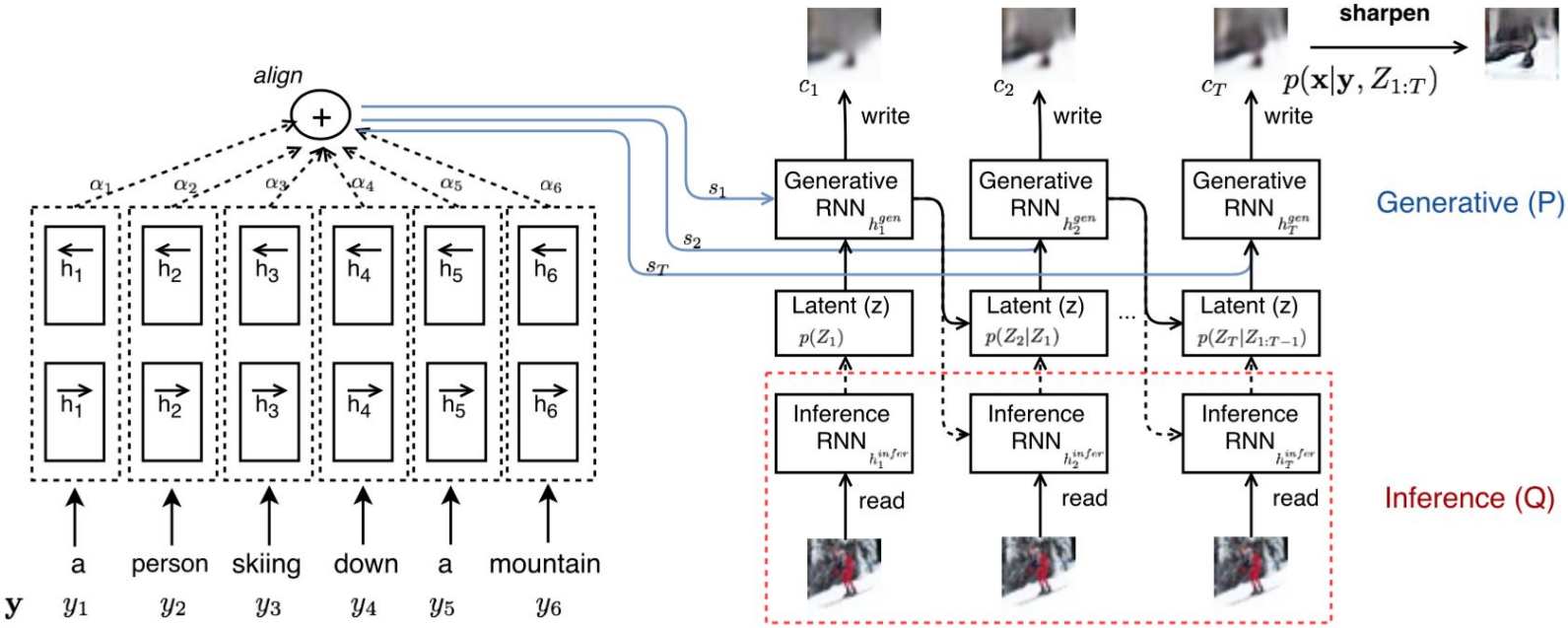
“A little girl in a white shirt and black skirt is holding up a red bag.”
p=0.000000

“A man in a purple shirt is playing a guitar.”
p = 0.000006

“A man in a red shirt is standing in front of a white truck.”
p = 0.000014

txt2im

Based on "Generating Images from Captions with Attention" by Elman Mansimov et al (ICLR 2016)



In **every iteration**, we generate a **patch** in the output **canvas *c***, given a **sentence representation** and a **latent variables z** which capture salient information about the training images.

txt2im

Experiments

Data	MS COCO
Number of Epochs	6 (18 in the paper)
Run-Time per Epoch	**15 h of BW**
Vocabulary size	25323
Processing step	convert all capital letters to small letters
Tool	Theano

“A black and white dog swims in the water.”

“A black and white dog swims in the water.”

“A brown dog jumps into the water.”

“A man in a red jacket is skiing down a snowy hill.”

“A brown dog is running through the grass.”

Conclusions

- LSTM outperformed RNN and GRU
- Nonunique representation of image captions
  - For the same image, even people might focus on different things and generate different captions
- Caption evaluation uncertainty
  - BLEU score drawbacks

References

- Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- Vinyals, Oriol, et al. "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016).
- 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.