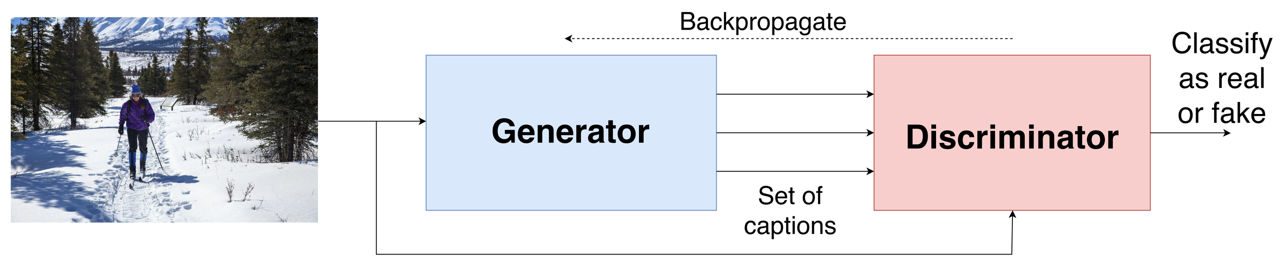
Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training

Rakshith Shetty[1], Marcus Rohrbach[2,3], Lisa Anne Hendricks[2], Mario Fritz[1], Bernt Schiele[1]
[1]Max Planck Institute for Informatics [2]UC Berkeley EECS [3]Facebook AI Research

Contact: rshetty@mpi-inf.mpg.de
https://goo.gl/3yRVnq

Summary

Motivation: Image captioning models generate correct but "safe" captions severely lacking in diversity compared to human written captions.



Generative Adversarial Network framework used to train our caption generator

Core Idea:

- Use GAN [1] framework to better match the data distribution.
- Generator produces multiple captions for an image by sampling.
- Discriminator scores this caption set on correctness and diversity.

Significantly higher diversity, larger vocabulary, more novel sentences, better match of language statistics, while maintaining same level of correctness.

Diverse captions on similar images

Ours a group of people standing around a shop | a group of young people standing around talking on cell phones | a group of soldiers stand in front of microphones | a couple of women standing next to a man in front of a store | a group of people posing for a photo in formal wear

Baseline *a group of people standing around a table*

Ours a person on skis jumping over a ramp | a skier is making a turn on a course | a person cross country skiing on a trail | a skier is headed down a steep slope | a cross country skier makes his way through the snow

Baseline *a man riding skis down a snow covered slope*

Ours a surfer rides a large wave in the ocean | a surfer is falling off his board as he rides a wave | a person on a surfboard riding a wave | a man surfing on a surfboard in rough waters | a surfer rides a small wave in the ocean

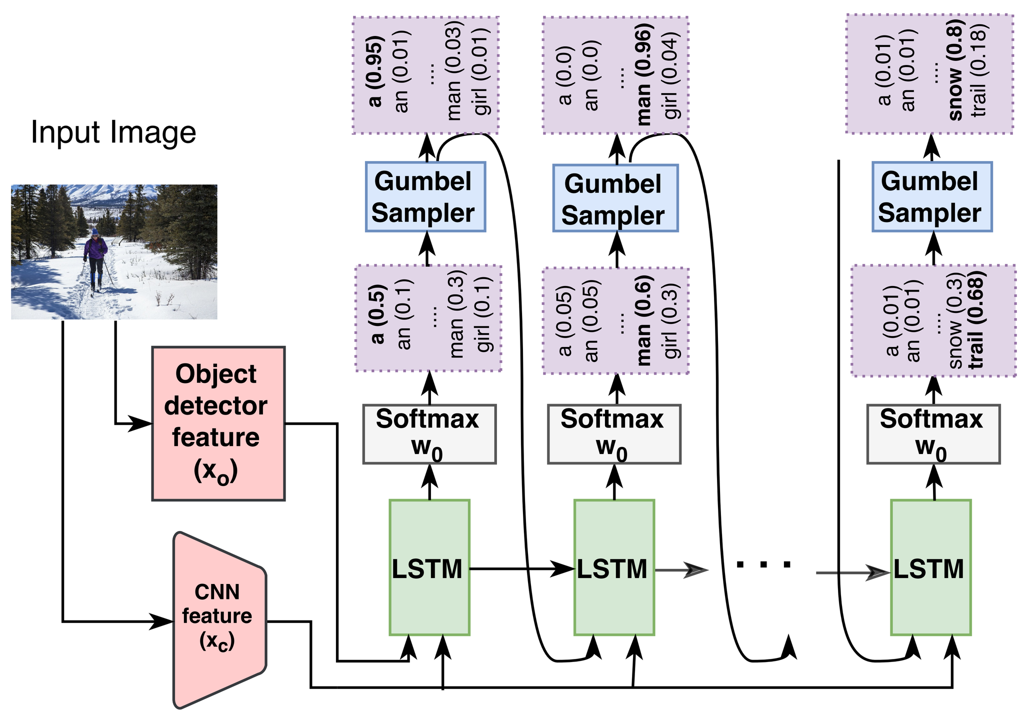
Baseline *a man riding a wave on top of a surfboard*

Ours a bathroom with a walk in shower and a sink | a dirty bathroom with a broken toilet and sink | a view of a very nice looking rest room | a white toilet in a public restroom stall | a small bathroom has a broken toilet and a broken sink

Baseline *a bathroom with a toilet and a sink*

Model

Generator



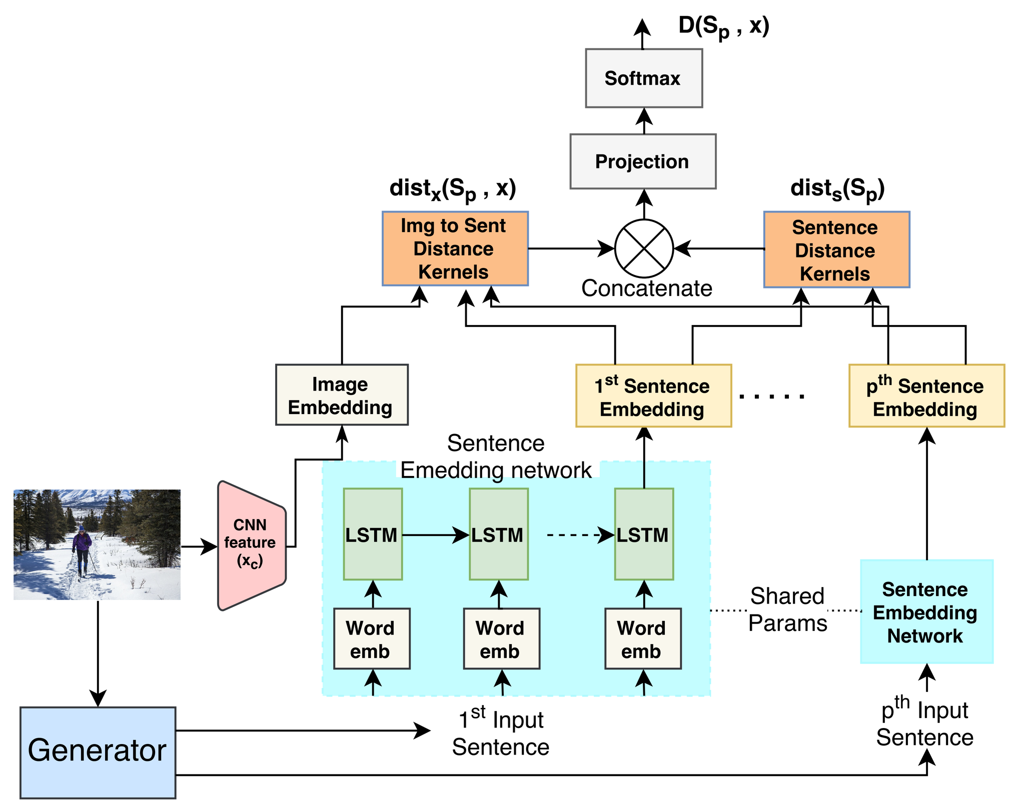
- 3-layers LSTM with residual connections.
- Use **Gumbel-softmax approximation** [2] for differentiability.

Gumbel-Max Trick: $r = \text{one\_hot}\left[\arg\max_i (g_i + \log\theta_i)\right]$

Softmax approximation: $r' = \text{softmax}(g_i + \log\theta_i)$

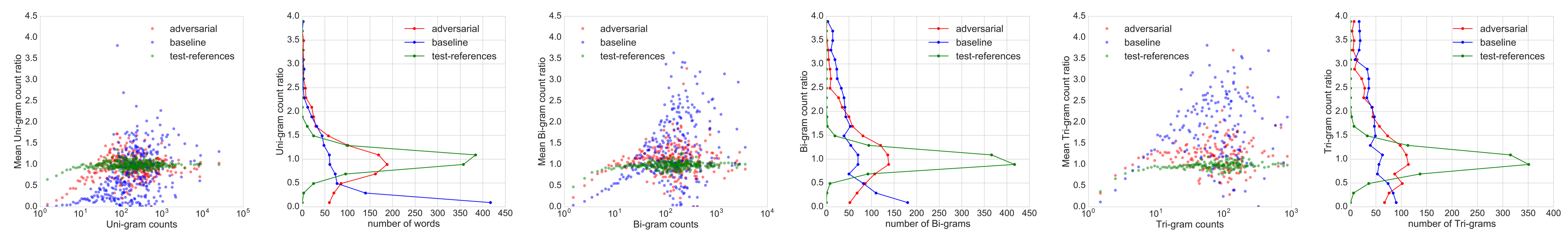
- Feature matching loss [3] helps.

Discriminator



- Evaluate a set of five captions per image.
- Compute two distances with image and sentence embeddings:
  - **Image to sentence distances** – for semantic correctness
  - **Intra-sentence distances** – for sufficient diversity

Quantitative Results



Adversarial model better matches the n-gram distribution of the dataset. Figure compares n-gram count ratios of the generated captions to true test set captions. Scatter plots show the n-gram count-ratios as a function of counts on training set. Adjoining the scatter plots on the right are the histogram plots of the count-ratios.

Method	n	Div-1	Div-2	mBleu-4	Vocab-ulary	% Novel Sentences
Base-beamsearch	1 of 5	–	–	–	756	34.18
	5 of 5	0.28	0.38	0.78	1085	44.27
Base-sampling	1 of 5	–	–	–	839	52.04
	5 of 5	0.31	0.44	0.68	1460	55.24
Adv-beamsearch	1 of 5	–	–	–	1508	68.62
	5 of 5	0.34	0.44	0.70	2176	72.53
Adv-sampling	1 of 5	–	–	–	1616	73.92
	5 of 5	**0.41**	**0.55**	**0.51**	**2671**	**79.84**
Human captions	1 of 5	–	–	–	3347	92.80
	5 of 5	0.53	0.74	0.20	7253	95.05

Adversarial model has significantly better diversity statistics. Div-1 and Div-2 measure the n-gram uniqueness in the 5 samples. mBleu-4 measures the similarity in terms of Bleu-4. Vocabulary size increases by 100% and 82% when using beamsearch and sampling respectively with the adversarial model.

Comparison	Adversarial - Better	Adversarial - Worse
Beamsearch	36.9	34.8
Sampling	35.7	33.2

Human evaluation shows that the adversarial model is on par to the baseline model in correctness. Five human evaluators were asked to pick the more correct caption of the two on 482 random images.

Image Feature	Evalset size (p)	Feature Matching	Meteor	Div-2	Vocab. Size
Baseline Model with VGG features			0.247	0.44	1367
VGG	1	No	0.179	0.40	812
VGG	5	No	0.197	0.52	1810
VGG	5	yes	0.207	**0.59**	2547
ResNet	5	yes	**0.236**	0.55	**2671**

Ablation study shows that multi-caption evaluation and feature matching are key to increasing diversity Comparison done on the validation set. Switching to ResNet features help improve semantics

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[2] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*, 2016.

[3] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, 2016.