Show and Tell: A Neural Image Caption Generator

Oriol Vinyals, Alexander Toshev,
Samy Bengio, Dumitru Erhan

CVPR 2015

(overview by Abhishek Jha)

Summer School on Deep Learning for Computer Vision
July 11-16, 2016: CVIT, IIIT Hyderabad

Results: Dataset and Evaluation Measures

SBU

I don't chew up the couch and pee in the kitchen mama!

MSCOCO

The man at bat readies to swing at the pitch while the umpire looks on.

PASCAL VOC 08

Children play racing games in an arcade.
A group of kids playing games.
A few kids playing arcade games.
some kids in an arcade.
Kids are playing racing games.

Flickr 8k/30k

A group of people sit at a table in front of a large building.
People are drinking and walking in front of a brick building.
People are enjoying drinks at a table outside a large brick building.
Two people are seated at a table with drinks.
Two people are sitting at an outdoor cafe in front of an old building.

Evaluation Metrics:

1) BLEU-1 / BLEU-4: BLEU-n is a geometric average of precision over 1- to n-grams.

Eg.:

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

BLEU-1 score for candidate w.r.t to two references= 2/7

2) Manual rating: Each Image rated by 2 AMT* workers (mutual agreement=65%). Average score was considered.

3) Perplexity: Geometric Mean of the inverse probability for each predicted word.

4) Other Metric used: METEOR and Cider.

*AMT: Amazon Mechanical Turk.

Generation Results

MS-COCO:

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.

- Model fares strongly versus human raters.
- On official test set, BLEU-4 = **27.2**

Other Datasets:

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text				11
TreeTalk				19
BabyTalk	25			
Tri5Sem			48	
m-RNN		55	58	
MNLM		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Table 2. BLEU-1 scores. (SOTA: State-of-the-art)

- For PASCAL training has been done on MS-COCO
- One human caption is compared against remaining 4 caption.
- Each generated caption is compared against 5 human captions in the test set.
- The average difference of having 5 reference has been added back to the human score.

Transfer Learning

Transferring a model trained on one dataset to a different dataset, to study how much the mismatch in domain would be compensated.\

Flickr 30k to Flickr 8k:

(4X training data)

- Dataset collection **process same**.
- BLEU score **increased** by 4 points.

Inference: More data, better performance

MSCOCO to Flickf 8k:

(20X training data)

- Dataset collection process different
- BLEU score **decreased** by 10 points.

Inference: Mismatch in vocab, despite larger training dataset

MSCOCO to PASCAL: BLEU-1 = 59

Flickr 30 to PASCAL: BLEU-1 = 53

MSCOCO to SBU:

- SBU has Noisy and non visual descriptions.
- BLEU score **decreased** from 28 to 16 points.

Inference: Mismatch in vocab., weak labeling

Generation Diversity

Target: Model generates novel,diverse and high quality captions.

- 80% of the best candidate sentences come from training set
- Top-15 generated sentences: 50% comes from training and BLEU score = 58(~ Human)

Inference: Both diversity and quality are good

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.

Table 3. N-best examples from the MSCOCO test set. Bold lines indicate a novel sentence not present in the training set.

Ranking Results

Recall @ K

Flickr 8k

Approach	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
DeFrag	13	44	14	10	43	15
m-RNN	15	49	11	12	42	15
MNLM	18	55	8	13	52	10
NIC	**20**	**61**	**6**	**19**	**64**	**5**

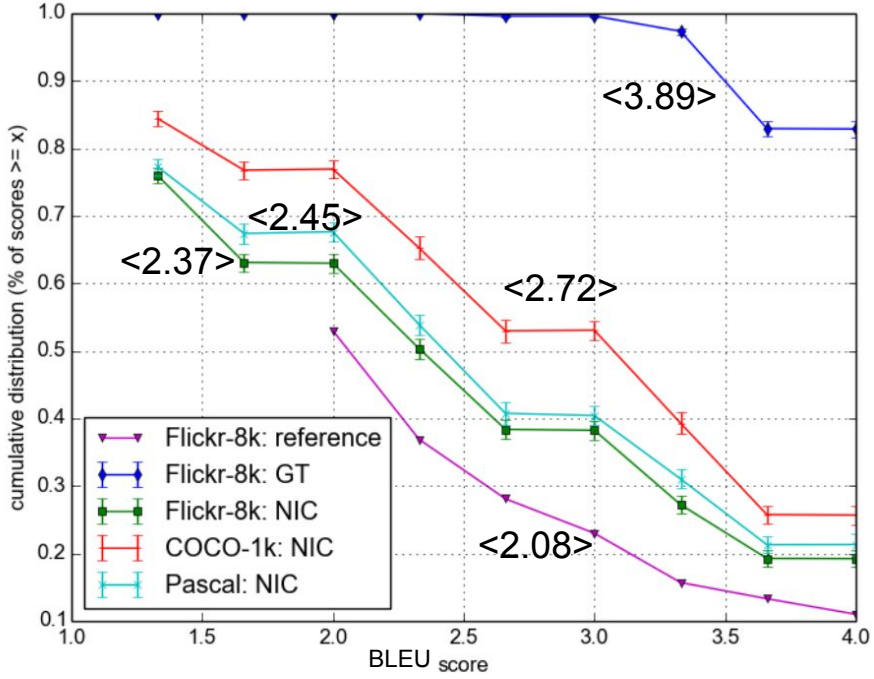
Table 4.

Flickr 30k

Approach	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
DeFrag	16	55	8	10	45	13
m-RNN	18	51	10	13	42	16
MNLM	**23**	**63**	**5**	**17**	**57**	**8**
NIC	17	56	7	**17**	**57**	**7**

Table 5.

Human Evaluation



- NIC is better than reference but worse than ground truth (GT).
- <average scores> shown in graph.

Inference: BLEU is not a perfect metric, as it does not capture well the difference between NIC and human descriptions assessed by raters.

Note: COCO-1k: A subset of 1000 images from the MSCOCO test set with descriptions produced by NIC.

Example of generated descriptions:

Describes without errors

Describes with minor errors

Somewhat related to image

Unrelated to image

Analysis of Embedding

- Using ‘word2vec’ have advantage of size of training LSTM irrespective of vocabulary.

- The resultant word vectors of this embedding also have semantic properties:

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

Table 6. Nearest neighbors of a few example words

- Relationships learned by word embedding helps the visual component.

- In cases where there’s a new class, the proximity to the nearest known class in the vocabulary would provide meaningful information to the model to generate sentences.

word2vec: T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In ICLR, 2013.