Show and Tell: A Neural Image Caption Generator

Oriol Vinyals
Google
vinyals@google.com

Alexander Toshev
Google
toshev@google.com

Samy Bengio
Google
bengio@google.com

Dumitru Erhan
Google
dumitru@google.com

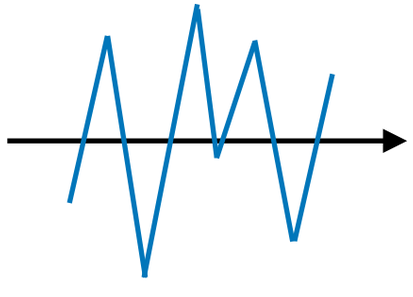
Presenter: WENWEI KANG

- Introduction

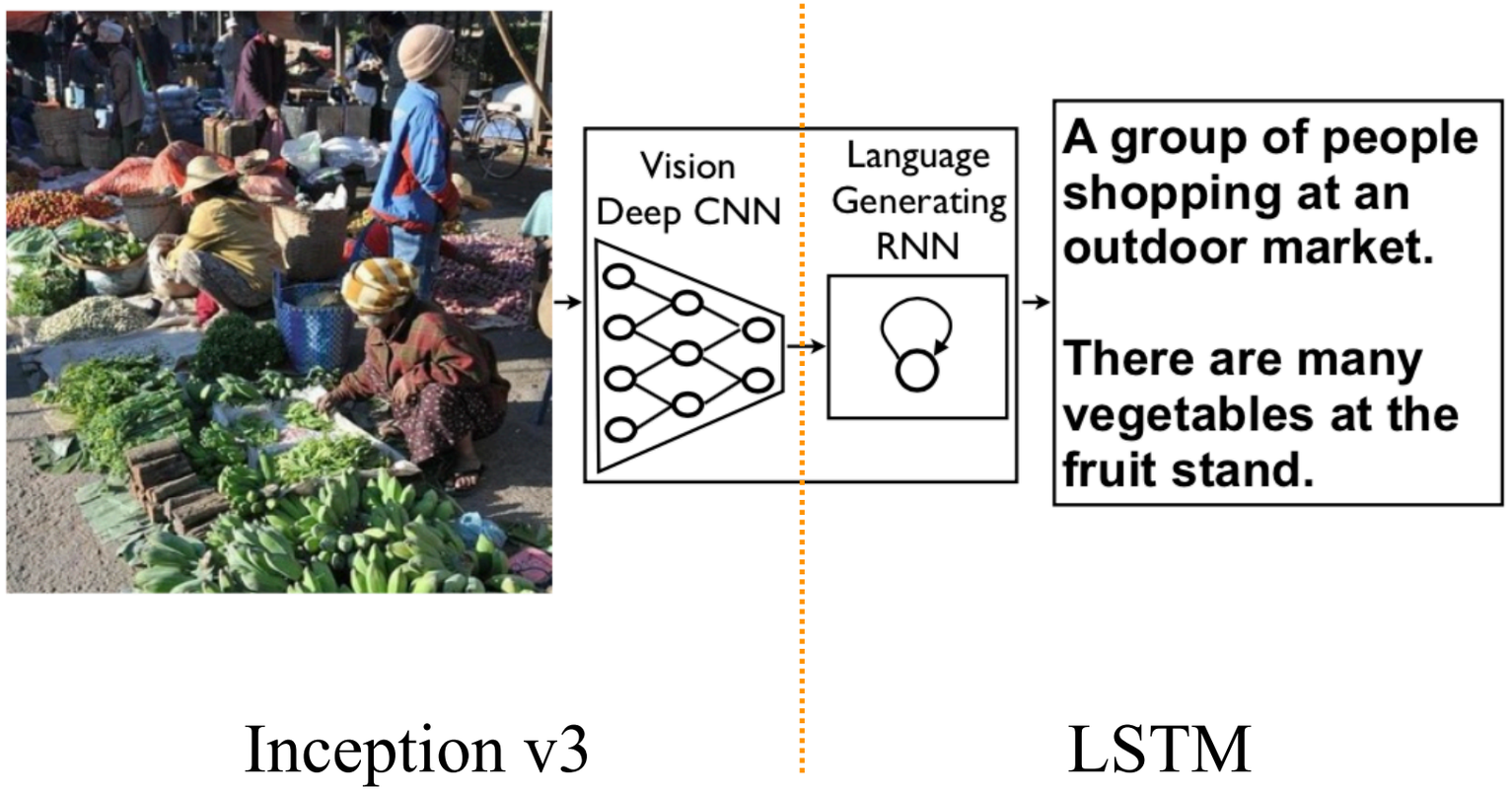
- Neural Image Caption

- Evaluation

Introduction

	Encoder	End to end	Decoder
Neural machine translation	猫は犬の横にいる	→	a cat is beside a dog
Image caption		→	a cat is beside a dog
Speech Recognition		→	a cat is beside a dog

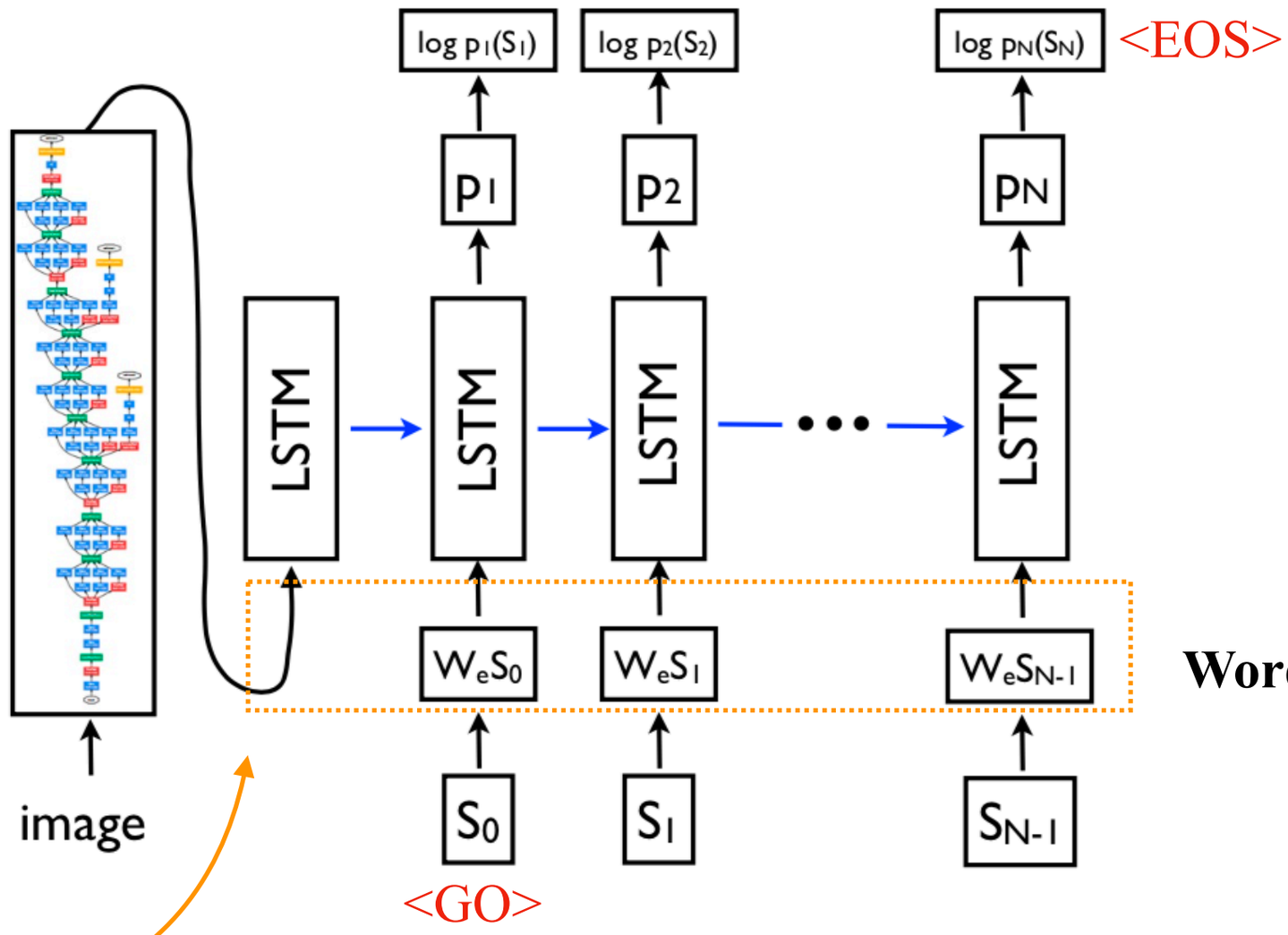
Neural Image Caption(1/3)



Neural Image Caption(2/3)

Inception v3

LSTM

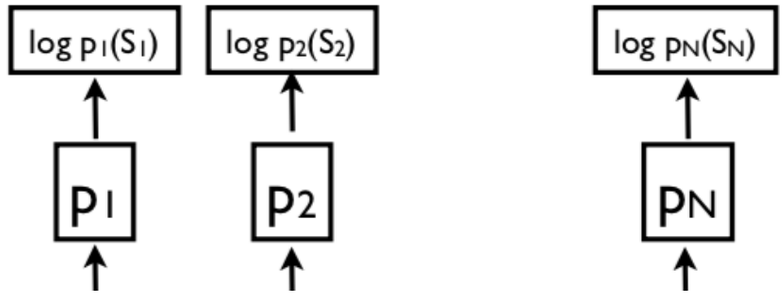


$$
\begin{aligned}
x_{-1} &= \text{CNN}(I) \\
x_t &= W_e S_t, \quad t \in \{0 \dots N-1\} \\
p_{t+1} &= \text{LSTM}(x_t), \quad t \in \{0 \dots N-1\}
\end{aligned}
$$

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

Neural Image Caption(3/3)

Inference



- **Sampling:** Sample the first word according to p_1
- **Beam Search:** Consider the set of the *k* best sentences up to time *t* as candidates to generate sentences of size *t* + *1*, and keep only the resulting best *k* of them.

Beam Search = 2, vocabulary size = 3, {a,b,c}

t = 0, p_1 = {a, b}
t = 1, p_2 = {aa, ab, ac, ba, bb, bc}
t = 3, ……

Evaluation(1/)

Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [6]	-	-	1000
Flickr8k [26]	6000	1000	1000
Flickr30k [33]	28000	1000	1000
MSCOCO [20]	82783	40504	40775
SBU [24]	1M	-	-

Pascal Visual Object Classes: image classification, segmentation, caption.

Microsoft COCO:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

SBU: consist of descriptions given by image owners when they uploaded to Flickr.

Evaluation(2/)

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.

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [24]				11
TreeTalk [18]				19
BabyTalk [16]	25			
Tri5Sem [11]			48	
m-RNN [21]		55	58	
MNLM [14][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Table 2. BLEU-1 scores. We only report previous work results when available. SOTA stands for the current state-of-the-art.

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.**

Evaluation(3/)

Approach	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
DeFrag [13]	13	44	14	10	43	15
m-RNN [21]	15	49	11	12	42	15
MNLM [14]	18	55	8	13	52	10
NIC	**20**	**61**	**6**	**19**	**64**	**5**

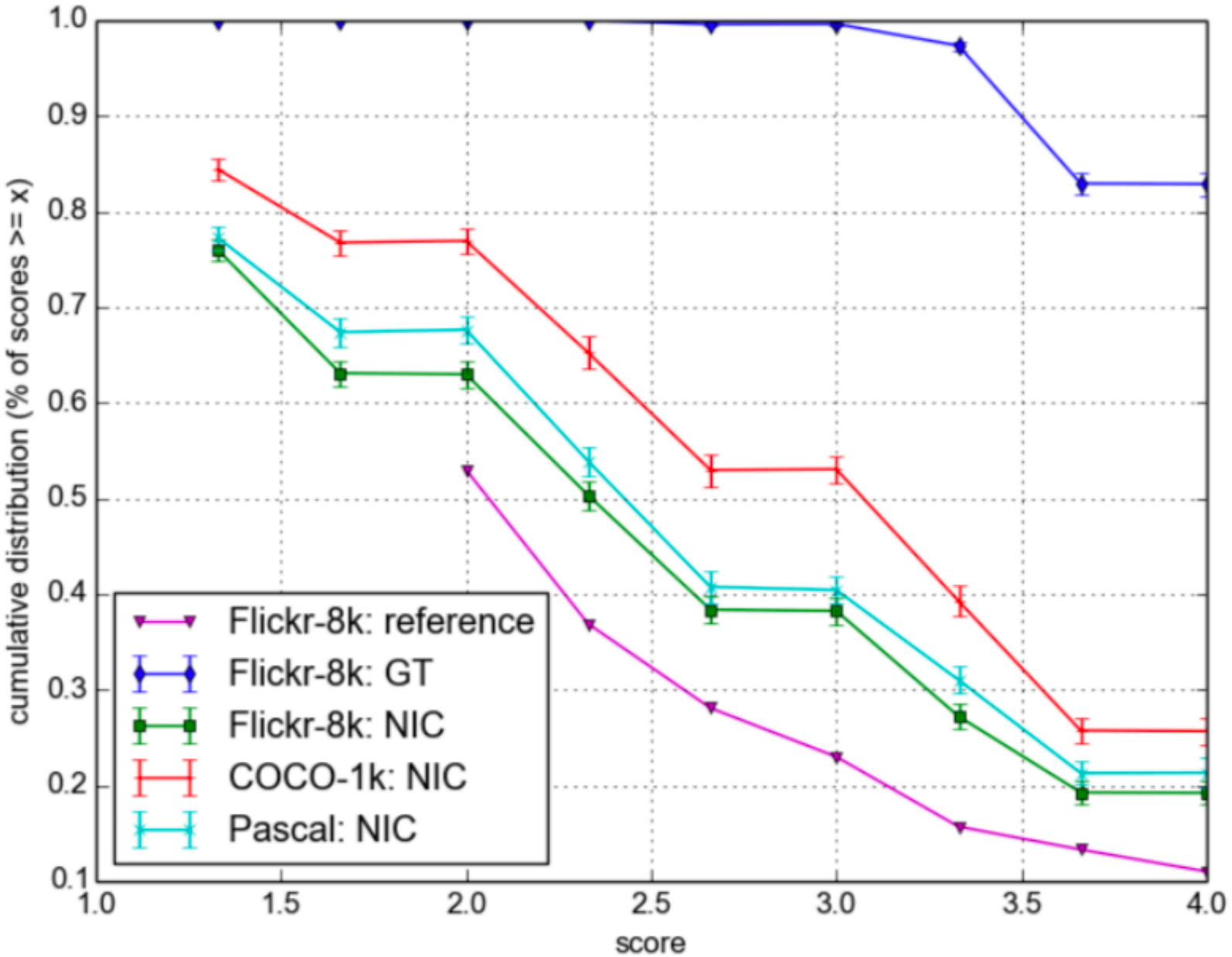
Table 4. Recall@k and median rank on Flickr8k.

Approach	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
DeFrag [13]	16	55	8	10	45	13
m-RNN [21]	18	51	10	13	42	16
MNLM [14]	**23**	**63**	**5**	**17**	**57**	**8**
NIC	17	56	7	**17**	**57**	**7**

Table 5. Recall@k and median rank on Flickr30k.

Image Annotation: return a ranked list of the 1000 images in I_{test} for each of the captions in S_{test}
Image Search: return a ranked list of the 1000 captions in S_{test} for each of the images in I_{test}

Evaluation(4/)



Evaluation(5/)

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.

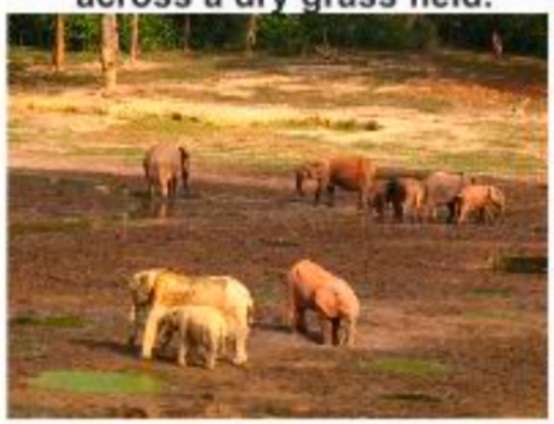
A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

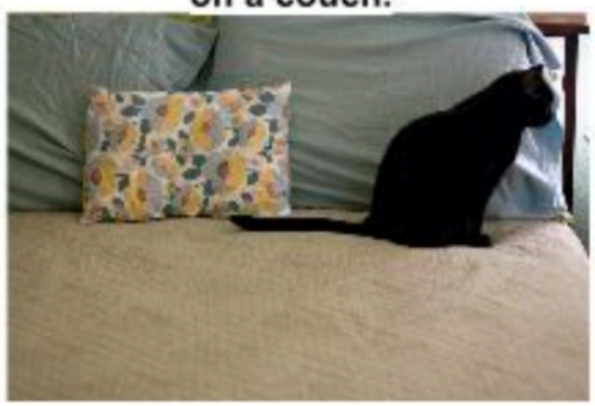
A little girl in a pink hat is blowing bubbles.

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