Show and Tell: A Neural Image Caption Generator

Oriol Vinyals
Google

Alexander Toshev
Google

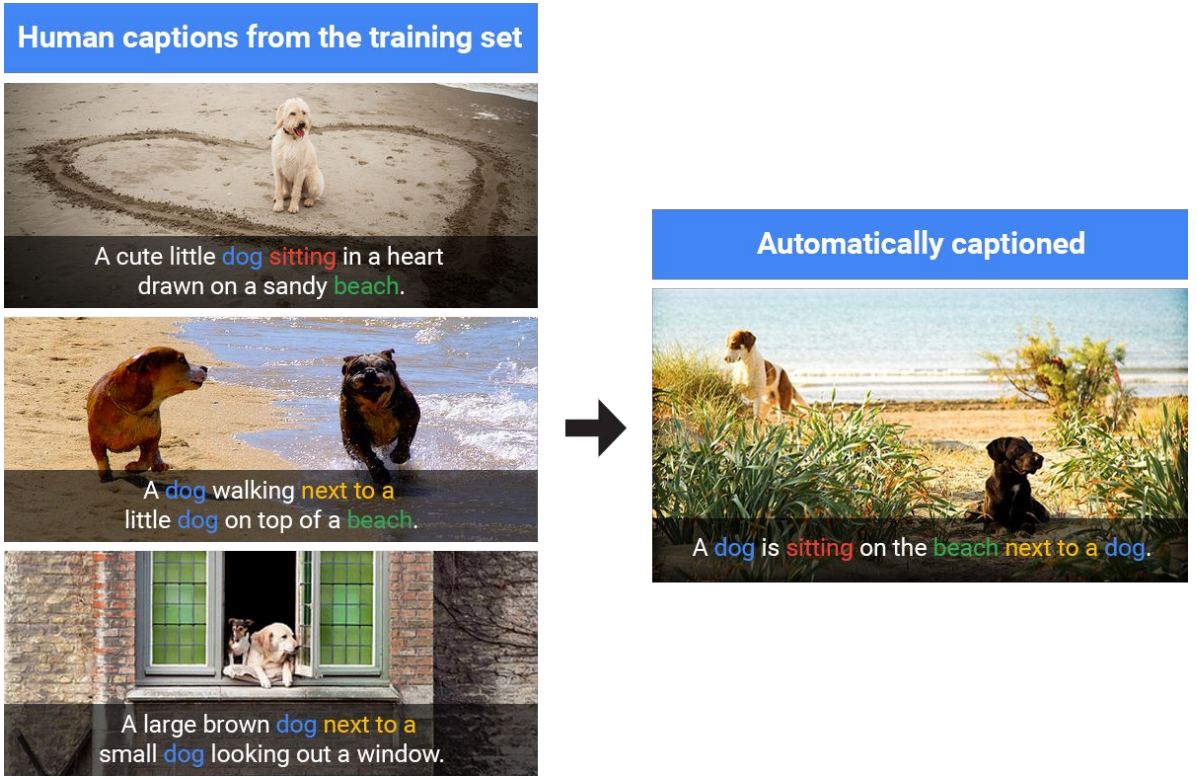
Samy Bengio
Google

Dumitru Erhan
Google

Reviewed by
Mithun

Overview

- Problem Statement
- Model Overview
- Prior Work
- Model Architecture
- Experiments
- Conclusion



Problem Statement

We want to build a system that views an image and automatically describes the content of an image in Natural Language (plain English). The system should be able to generate novel descriptions that are both diverse and high-quality.

Why is it challenging?

Interdisciplinary problem that connects both Computer Vision & NLP

Must express semantic relations between objects contained & activities they are involved in.

Very hard to evaluate how well the built system performs.

Why is it useful?

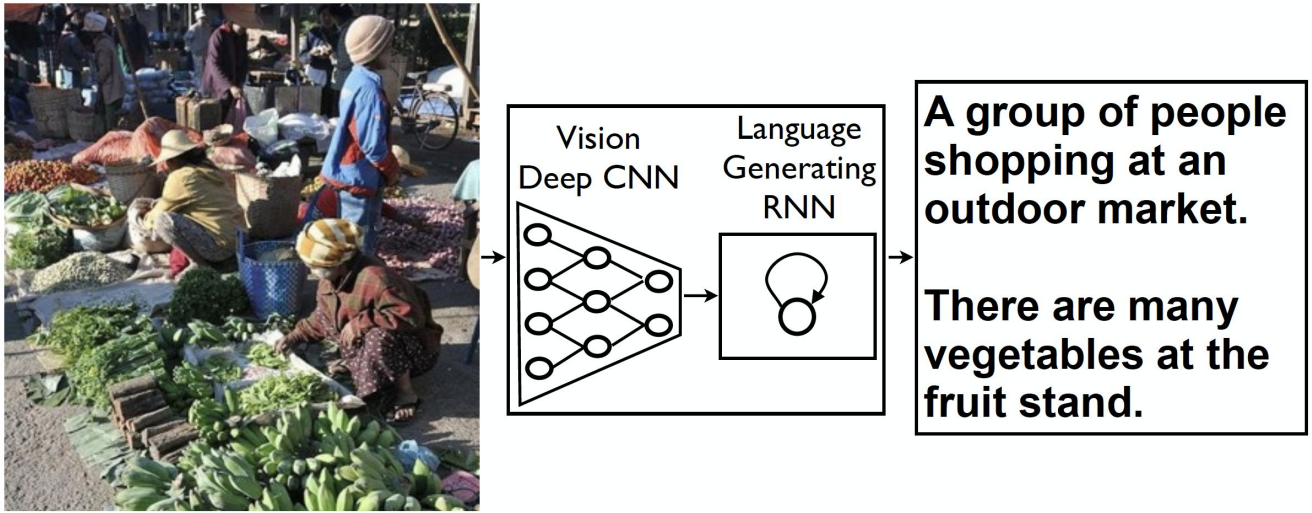
Multiple real-world appl. Could help visually impaired better understand the content of of images on the web.

Could be a possible precursor to video captioning & scene understanding.

Model Overview

Two subproblems to solve - Object detection + Generating description.

Proposes a single joint deep recurrent architecture, based on a generative model (abbreviating this model as **NIC**) that takes an image I as input and is trained to maximise the likelihood $p(S \mid I)$ of producing a target sequence of words $S = \{S_1, S_2, \ldots\}$.



Model Overview

Let's see the model's performance against contemporary state-of-the-arts.

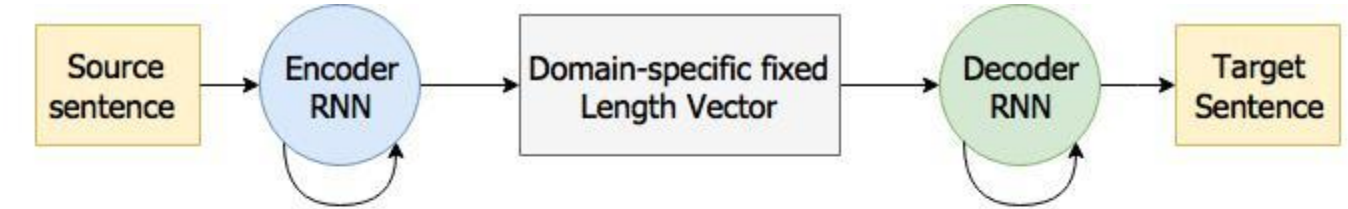
On Pascal VOC, NIC's BLEU score is **59**. Contemporary state-of-the art scored 25. *Humans scored 69.*

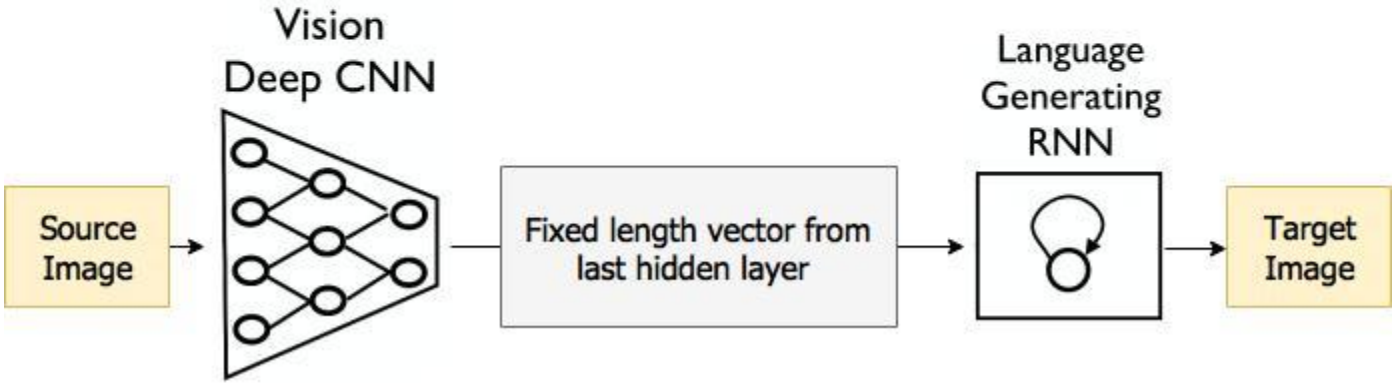
Improved scores on Flickr30k dataset from 56 to **66**. On SBU dataset, from 19 to **28**

Why is the BLEU score for human-generated descriptions so low?

Prior Work

Main inspiration - **Machine Translation** - Task is to transform sentence S written in source language into its translation T by maximizing $p(T|S)$





Prior Work

Earliest approaches used

- heavily hand-engineered features for visual elements
- rule based systems for language models.

Some tried to match with human-engineered templates and piecing together phrases containing detected objects.

Problem? **Difficult to expand models to larger training sets.**

More recent works, approached the problem by co-embedding images and text in the same vector space.

Problem? **They don't generate novel descriptions. They don't address the problem of evaluating how good a generated description is.**

Model Architecture

Directly **maximize the log-probability** of correct description given the image.

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

For a sentence of unbounded length S = {$S_1, S_2, \ldots, S_N$}

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

How to model $p(S_t \mid I, S_0, S_1, \ldots, S_{t-1})$?

Recurrent Neural Networks

Model Architecture

$$p(S_t \mid I, S_0, S_1, \ldots, S_{t-1})$$

The variables conditioned on the right hand side are expressed by Fixed length hidden states or memory $\mathbf{h_t}$. On seeing new input $\mathbf{x_t}$, a nonlinear function updates the memory.

$$\mathbf{h_t} = \mathbf{f}\,(\,\mathbf{h_t}\,,\,\mathbf{x_t}\,)$$

What is **f** ?

For **f** we use **Long-Short Term Memory** net.

LSTMs are known for sequence modelling.

How to represent $\mathbf{x_t}$?

Images are represented using **CNN**s.

Pre-trained on ImageNet to initialize weights.

Model Architecture

Long-Short Term Memory nets

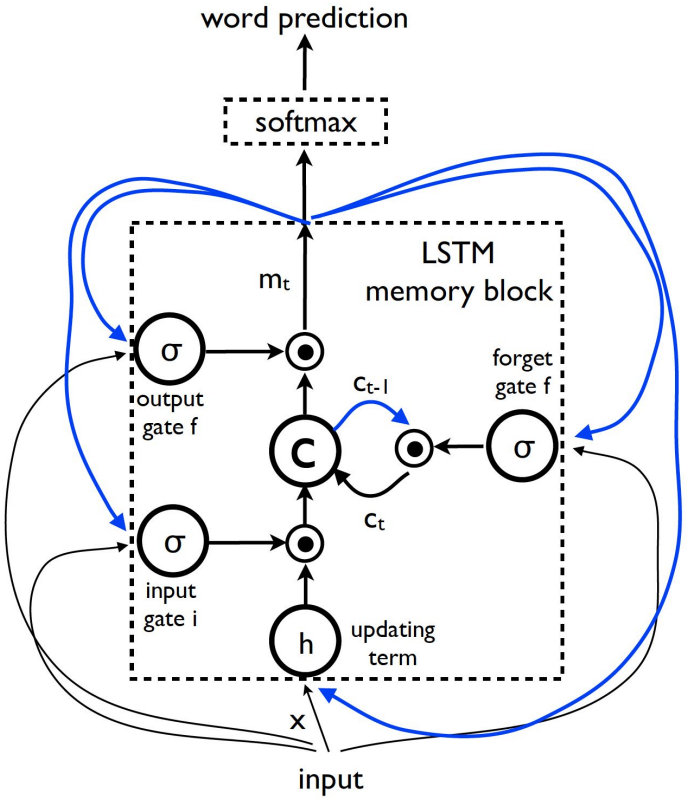
The behavior of the cell node **c** is controlled by **gates.**

- forget gate **f**
- input gate **i**
- output gate **o**

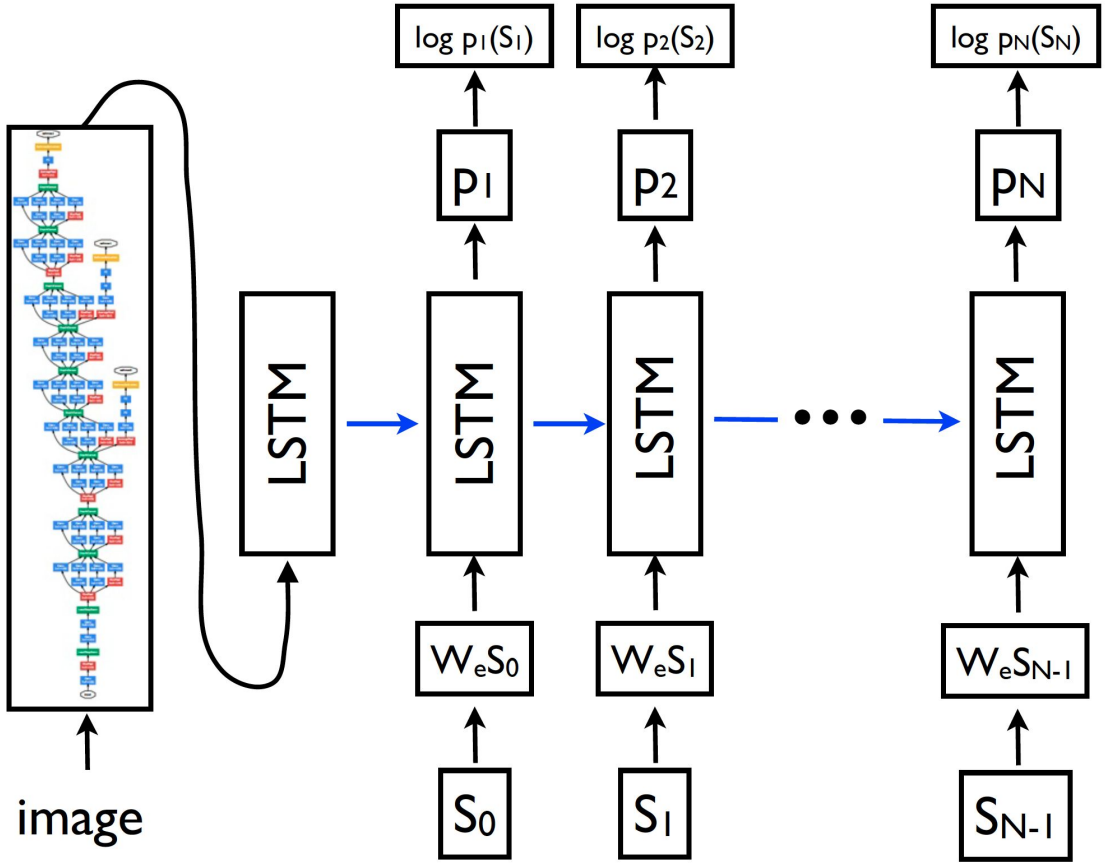
LSTMs are known to deal well with vanishing or exploding gradients problem.

Gates use a sigmoid nonlinearity while the cell works with a hyperbolic tangent.

m_t is fed to softmax to obtain probability distribution over all words at time t.



Model Architecture



Think of the LSTMs in an unrolled form. All the LSTMs share the same parameters.

Recurrent connections are transformed to feed-forward connections.

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

Model Architecture

Training

Both image and text mapped to same vector space - Image using **CNN** and the words using **Word Embeddings**.

Loss of the system is defined as the sum of the **negative log-likelihood** of the correct word at each step.

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

Inference

- Sampling - sample words one after another till end-word token
- BeamSearch - k-best sentences. Typically used for ranked retrievals.

Experiments

Evaluation Metrics

Evaluating whether a generated description matches the ground-truth description is in itself a challenging task. Following metrics were used throughout.

Human raters on Amazon Mechanical Turk - time consuming but most reliable

BLEU (Bilingual Evaluation understudy) - most prevalent in literature, inexpensive

METEOR – Metric for Evaluation of Translation with Explicit ORdering

CIDEr – Consensus Based Image Description Evaluation

Recall@K – This metric is typically used to evaluate ranking results. (BeamSearch)

Perplexity - Wasn't reported. But used to fine-tune model hyperparameters.

Experiments

Datasets

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	-	-

SBU is captioned by image owners. Not guaranteed to be unbiased. Thus SBU has more noise.

Other datasets are standard, each image has been annotated by labelers with 5 sentences that are relatively visual and unbiased

Training Details

Initialize weights (**Image embeddings**) of the CNN to a pretrained model (ImageNet)

Weights (**512 Word embeddings**) of RNN are left uninitialized.

Dropouts and ensembling methods were employed.

Trained using SGD. Fixed learning rate. No momentum

Experiments

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

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	

Experiments

How novel & diverse are the generated descriptions?

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.**

Some samples returning K-best list. Notice how samples are diverse and show different aspects of same image.

High quality descriptions

Agreement in BLEU scores (**58%**) of top 15 candidates is equivalent to that of humans among them (**65%**)

Diverse descriptions

Almost 7-8 out of the top-15 descriptions are novel (not from training set), but having similar BLEU scores.

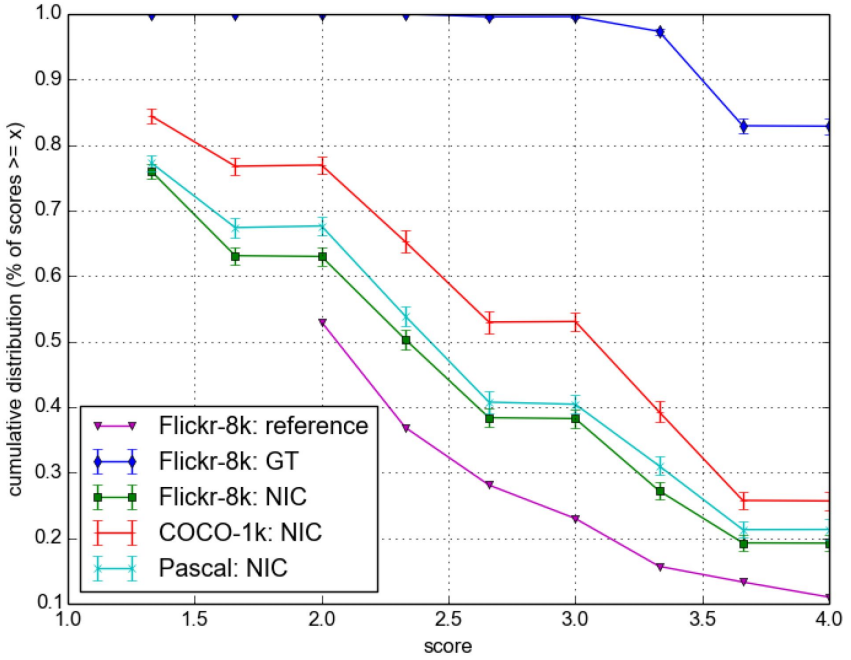
Experiments

Flickr8k

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**

Flickr30k

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**



Experiments

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

Conclusion

Image embeddings can be more specific to the task rather than just supplying the last layer of a typical CNN.

Can use an LSTM with more word embeddings to increase the richness of the description.

The system gets better with more and more data as empirically proven. With more data flowing in we can expect much more novel & diverse descriptions.

Can be extended to a video captioning model.

Can devise a better evaluation metric that can help gauge the performance of our system better.

Learning is currently supervised. It'll be interesting to see how the model reacts to unsupervised data.