Show and Tell

Presented by:
Anurag Paul

Show and Tell: A Neural Image Caption Generator

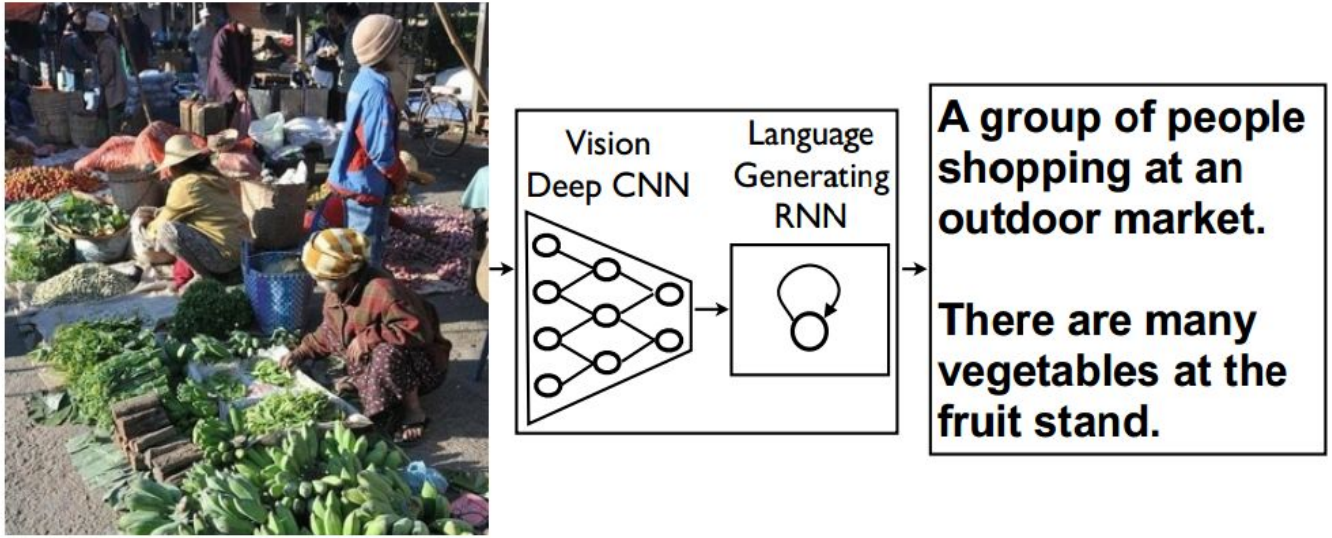
Authors: All of Google

Oriol Vinyals,
vinyals@google.com

Alexander Toshev,
toshev@google.com

Samy Bengio,
bengio@google.com

Dumitru Erhan,
dumitru@google.com



Agenda

- Related Work
- Architecture
- Metrics
- Datasets
- Analysis

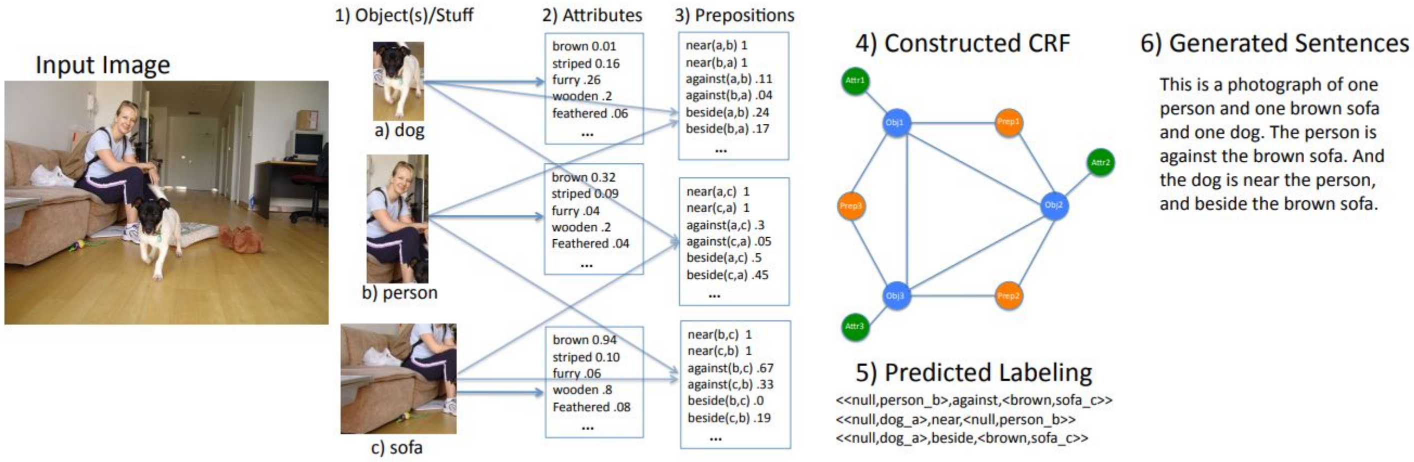
How can we do Image Captioning?

- If the dataset size is small (~1000 images)
- If there is only one class of data and need to capture fine-grained information (all images are of let’s say tennis)
- If there is a large amount of data (~ 1M) but it is noisy i.e. not labelled professionally

Solutions in Related Work

- Detecting Scene Elements and converting to sentence using templates
    - they are heavily hand designed and rigid when it comes to text generation.

Baby Talk



Solutions in Related Work

- Ranking descriptions for a given image Such approaches are
  - based on the idea of co-embedding of images and text in the same vector space
  - cannot describe previously unseen compositions of objects
  - avoid addressing the problem of evaluating how good a generated description is

DeFrag

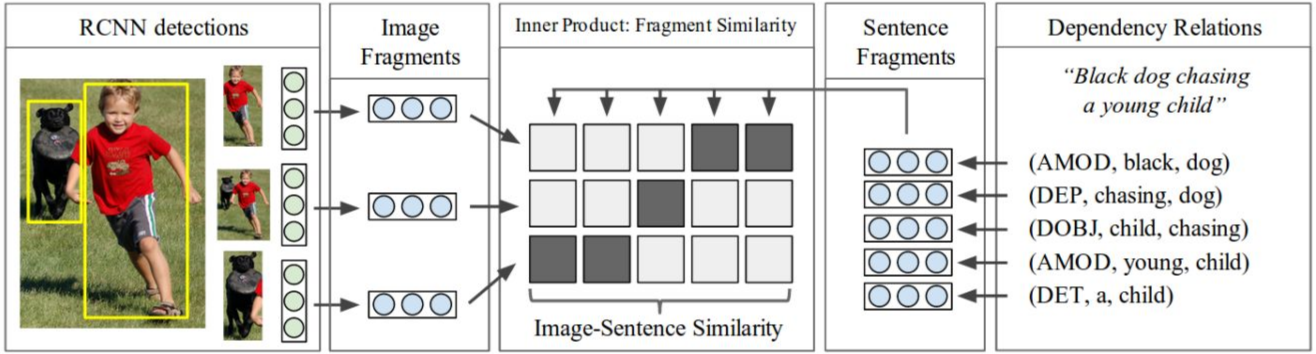
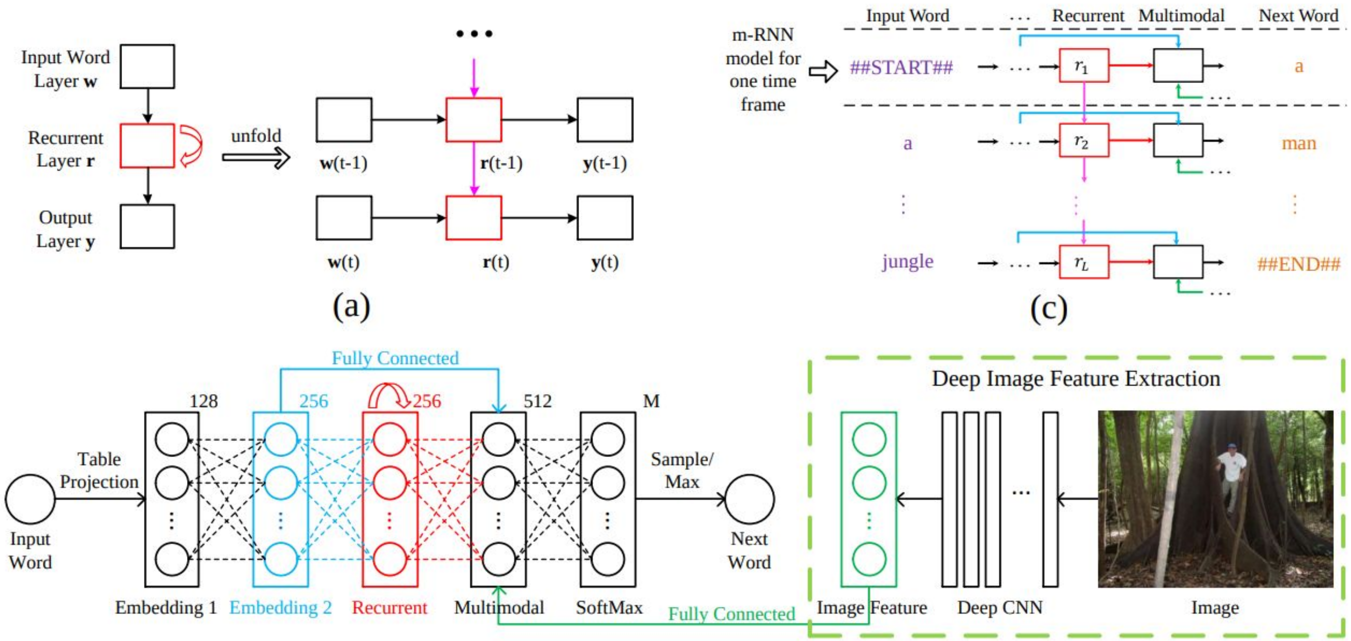


Figure 2: Computing the Fragment and image-sentence similarities. **Left:** CNN representations (green) of detected objects are mapped to the fragment embedding space (blue, Section 3.2). **Right:** Dependency tree relations in the sentence are embedded (Section 3.1). Our model interprets inner products (shown as boxes) between fragments as a similarity score. The alignment (shaded boxes) is latent and inferred by our model (Section 3.3.1). The image-sentence similarity is computed as a fixed function of the pairwise fragment scores.

Solutions in Related Work

m-RNN: Mao et al. uses a recurrent NN for the same prediction task.



Solutions in Related Work

MNLM: Kiros et al. propose to construct a joint multimodal embedding space by using a powerful computer vision model and an LSTM that encodes text.

- use two separate pathways (one for images, one for text) to define a joint Embedding,
- approach is highly tuned for ranking.

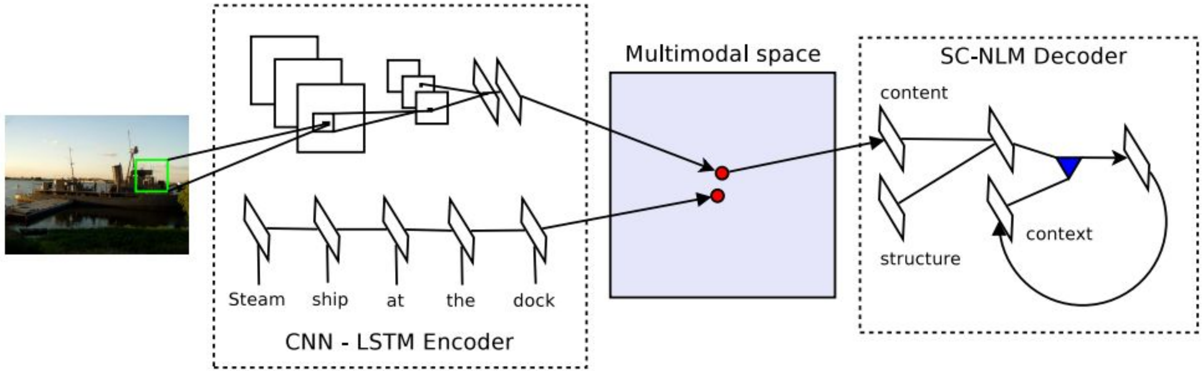
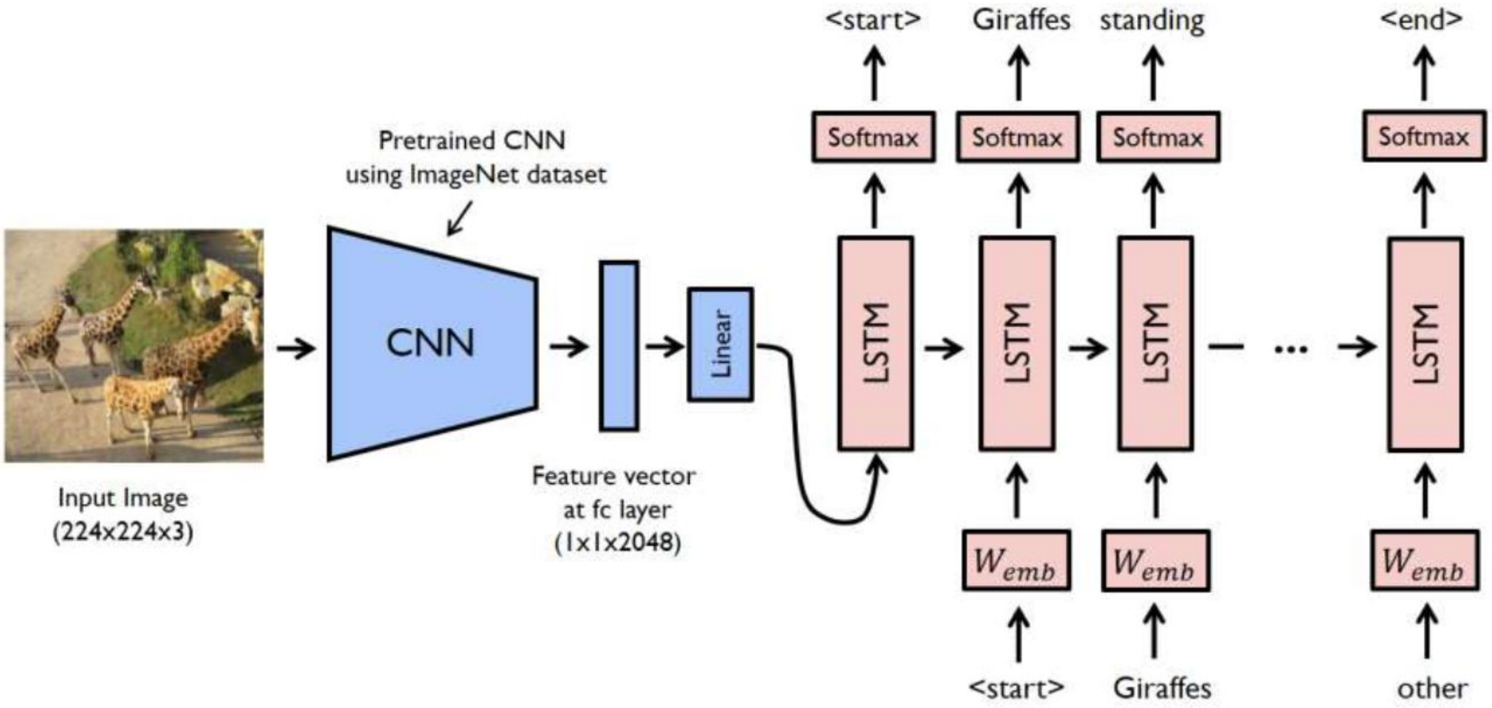
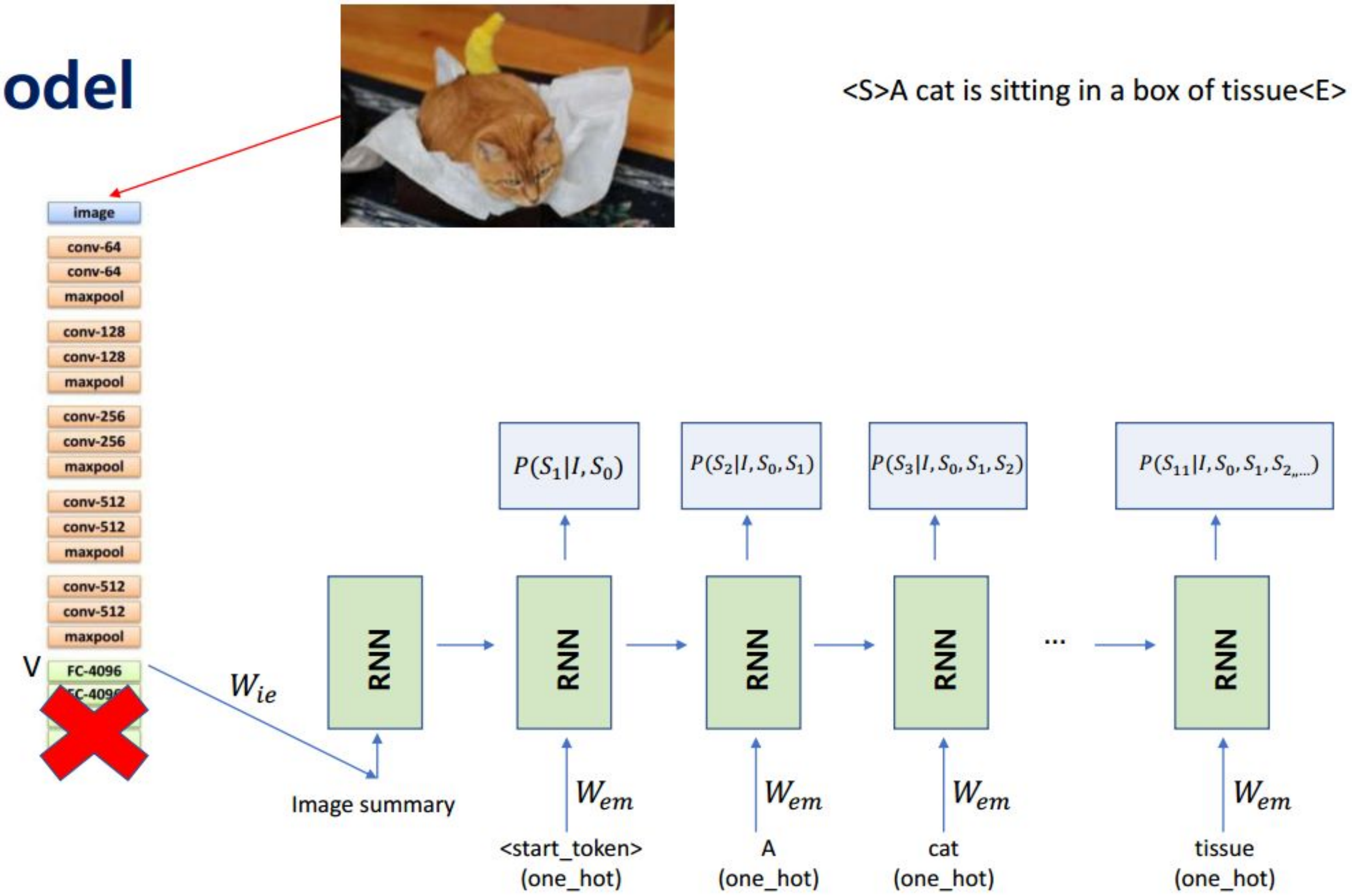


Figure 2: **Encoder:** A deep convolutional network (CNN) and long short-term memory recurrent network (LSTM) for learning a joint image-sentence embedding. **Decoder:** A new neural language model that combines structure and content vectors for generating words one at a time in sequence.

Model Architecture



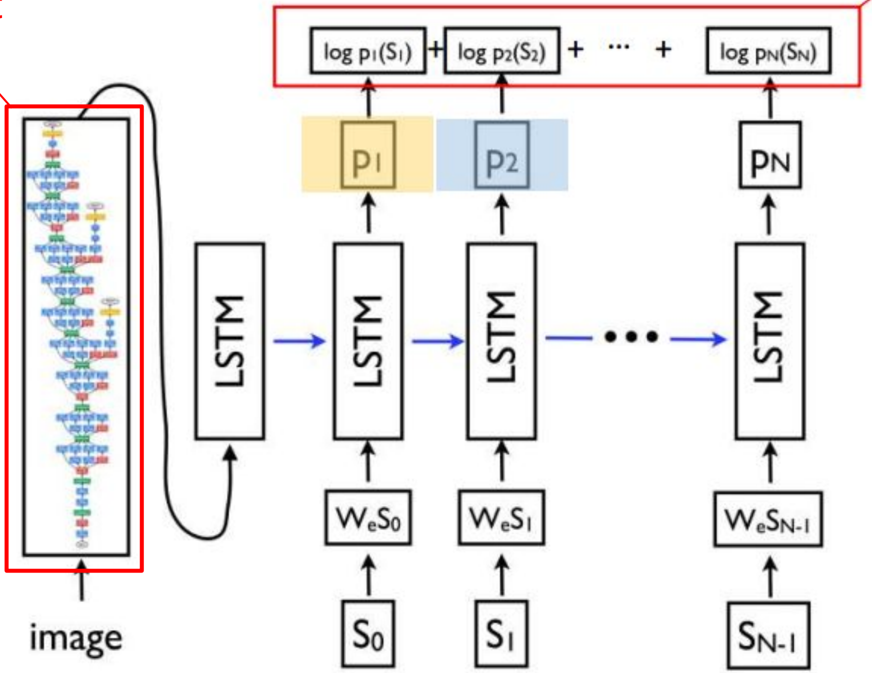
Model



<S>A cat is sitting in a box of tissue<E>
image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
V
FC-4096
W_{ie}
Image summary
RNN
RNN
RNN
RNN
...
RNN
$P(S_1|I, S_0)$
$P(S_2|I, S_0, S_1)$
$P(S_3|I, S_0, S_1, S_2)$
$P(S_{11}|I, S_0, S_1, S_{2,\ldots})$
W_{em}
W_{em}
W_{em}
W_{em}
<start_token>
(one_hot)
A
(one_hot)
cat
(one_hot)
tissue
(one_hot)

Model

GoogleNet
2014
ILSVRC
Winner



–*loss function*

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

$$p(S|I) = p(S_1|I,S_0) \cdot p(S_2|I,S_1,S_0) \cdots$$

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

Inference – Sentence generation

How to generate new sentence given an image (For testing, not training)?

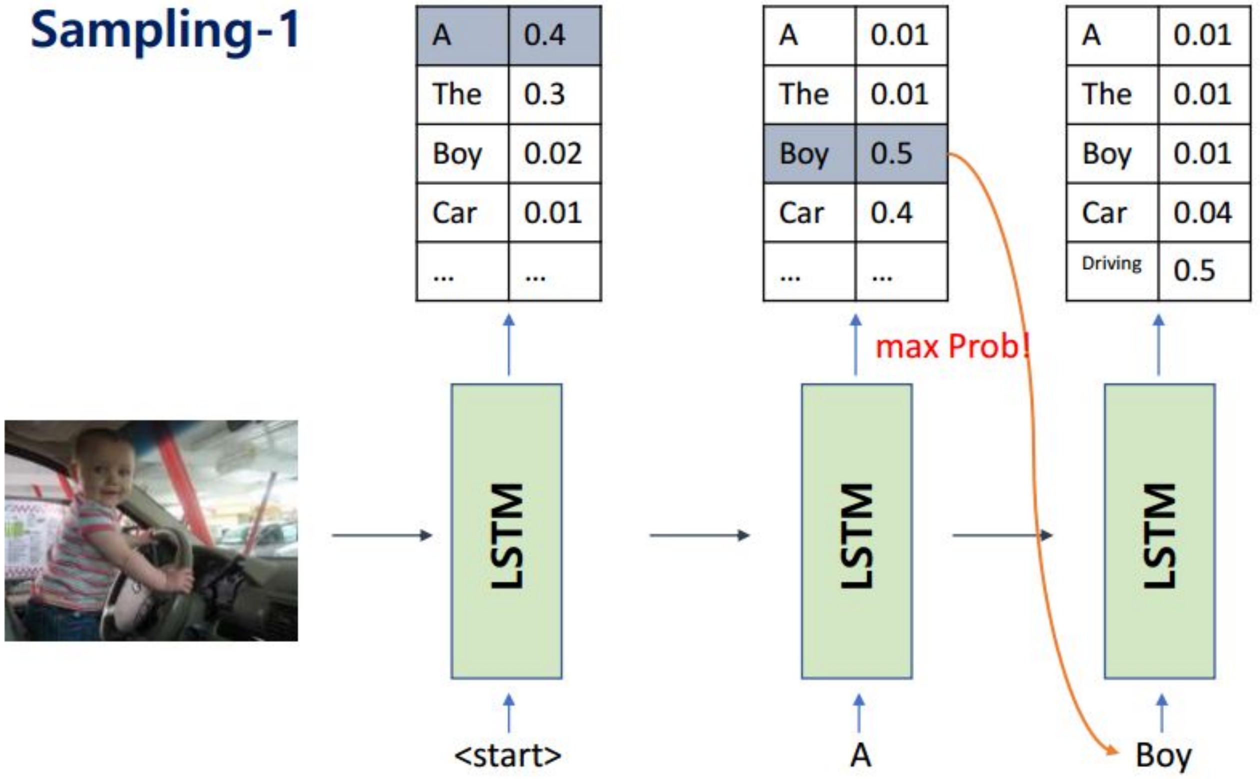
- **Sampling** - sample the first word according to p1, then provide the corresponding embedding as input and sample p2, continuing like this until we sample the special end-of-sentence token or some maximum length

- **BeamSearch**: iteratively 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. This better approximates S = arg maxS0 p(S0jI).

- Authors used the Beam Search approach in the paper with a beam of size 20.

- Using a beam size of 1 (i.e., greedy search) did degrade their results by 2 BLEU points on average.

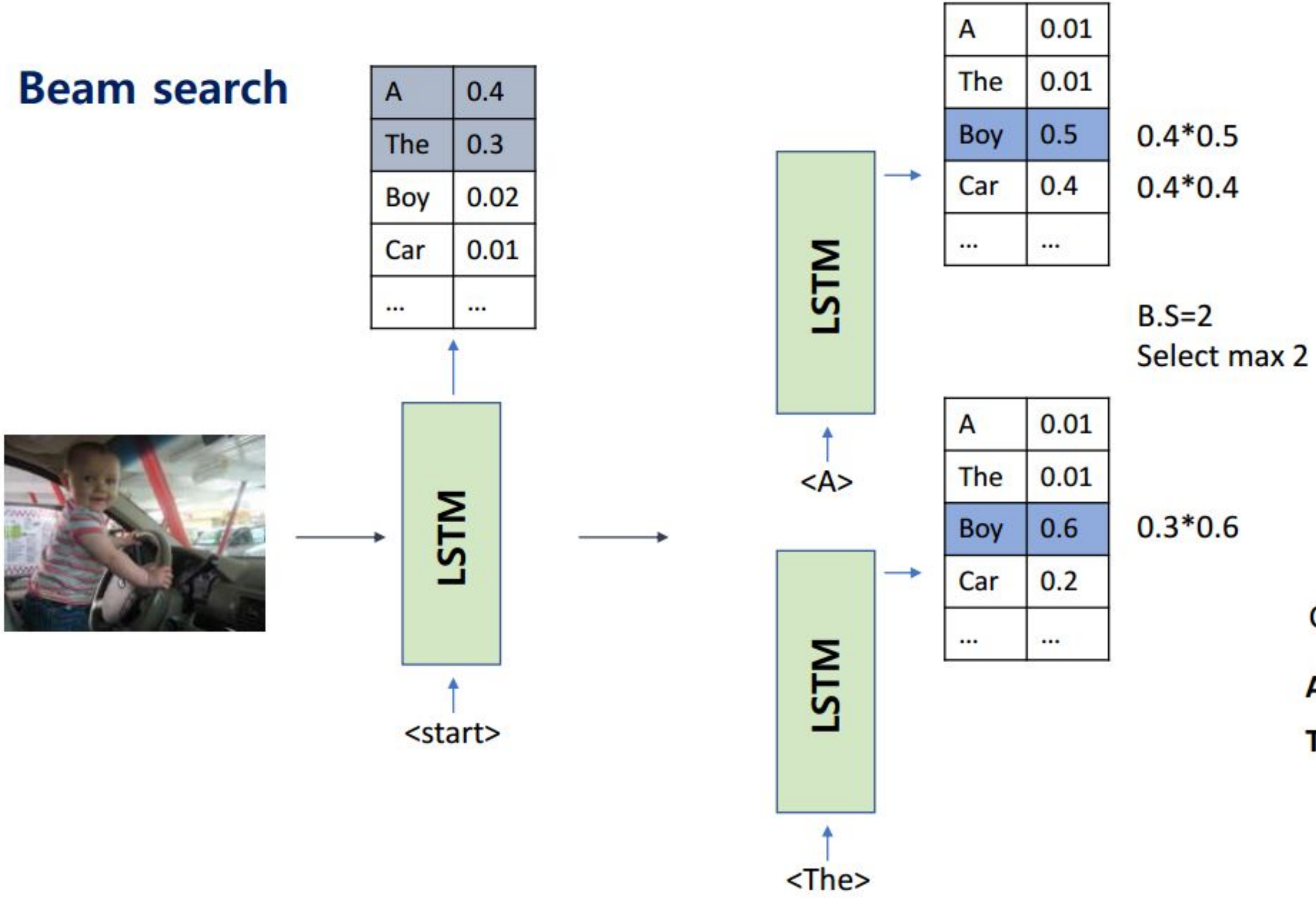
Sampling-1



Generate same caption
for the same image
-> deterministic

A boy driving a car

Beam search



Can get top-N captions

A boy driving a car

The boy driving a car

Training Details

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	-	-

Most of the datasets are quite small compared to the ones we have for Image Classification

Challenge: Overfitting

Solutions:

1) Pretrained CNN model
2) Dropout
3) Ensembling

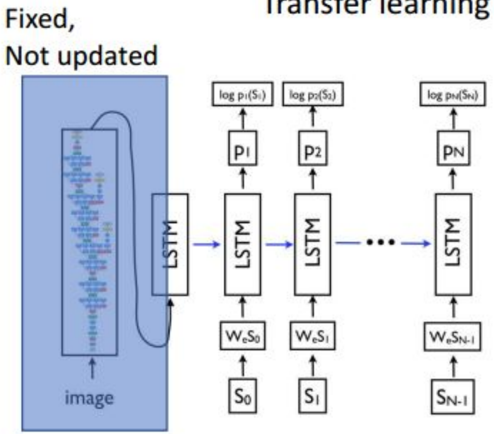
All weights randomly initialised except for the CNN

Training Details

- Loss function $L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$
- CNN pre-trained on ImageNet
- Minimize w.r.t. LSTM parameters, W_e and CNN top layer
- SGD on mini-batches
- Dropout and ensembling
- 512 dimensional embedding

+ fixed learning rate, no momentum

Transfer learning



Results


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

Evaluation Metrics

BLEU Score

$$P(i) = \frac{Matched(i)}{H(i)}$$

$$Matched(i) = \sum_{t_i} \min\{C_h(t_i), \max_j C_{hj}(t_i)\}$$

$$BLEU_a = \{\prod_{i=1}^{N} P(i)\}^{1/N}$$

$$\rho = \exp\{\min(0, \frac{n-L}{n})\}$$

$$BLEU_b = \rho \; BLEU_a$$

Evaluation Metrics

CIDEr

A measure of consensus would encode how often n-grams in the candidate sentence are present in the reference sentences.

- n-grams not present in the reference sentences should not be in the candidate sentence.
- n-grams that commonly occur across all images in the dataset should be given lower weight

TF-IDF

- TF places higher weight on n-grams that frequently occur in the reference sentence describing an image, while
- IDF reduces the weight of ngrams that commonly occur across all images in the dataset

Evaluation Metrics

CIDEr

- CIDEr_n score for n-grams of length n is computed using the average cosine similarity between the candidate sentence and the reference sentences

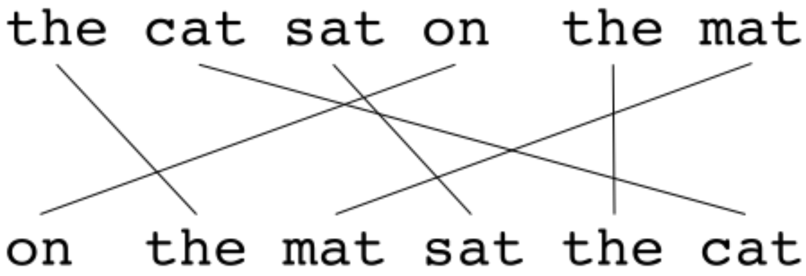
$$CIDEr_n(a,b) = \frac{1}{|b|} \sum_{j=1}^{|b|} \frac{\mathbf{g^n}(a) \cdot \mathbf{g^n}(b_j)}{\|\mathbf{g^n}(a)\| \|\mathbf{g^n}(b_j)\|}$$

$\mathbf{g^n}(x)$: vector formed by TF-IDF scores of all n-grams in x.

$$CIDEr(a,b) = \sum_{n=1}^{N} w_n CIDEr_n(a,b)$$

Evaluation Metrics

METEOR



$$P = \frac{m}{w_t} \qquad R = \frac{m}{w_r} \qquad F_{mean} = \frac{10PR}{R + 9P}$$

m is the number of unigrams in the candidate translation that are also found in the reference translation,
w_T is the number of unigrams in the candidate translation
w_R is the number of unigrams in the reference translation

$$p = 0.5\left(\frac{c}{u_m}\right)^3$$

$$M = F_{mean}\,(1 - p)$$

c is the number of chunks, and
u_m is the number of unigrams that have been mapped

Comparisons

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.

Comparisons

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.

High Diversity: Novel Sentences

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.

Analysis of Embeddings

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

Conclusion

- NIC, an end-to-end neural network system that can automatically view an image and generate a reasonable description in plain English
- NIC is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence
- The model is trained to maximize the likelihood of the sentence given the image
- Authors believe that as the size of the available datasets for image description increases, so will the performance of approaches like NIC