Show and Tell: A Neural Image Caption Generator (2015)

(https://arxiv.org/pdf/1411.4555.pdf)

1. AIM

1.1 Automatically describing the content of an image using generative model based on a deep re- current architecture.

1.2 The model is trained to maximise the likelihood of the target description sentence given the training image.

1.3 Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions.

2. PREVIOUS RELATED WORK / INSPIRATIONS

2.1 Complex systems composed of visual primitive recognisers combined with a structured formal language, e.g. And-Or Graphs or logic systems, which are further converted to natural language via rule-based systems.

2.2 Farhadi et al. [6] use detections to infer a triplet of scene elements which is converted to text using templates.

2.3 Similarly, Li et al. [19] start off with detections and piece together a final description using phrases containing detected objects and relationships.

2.4 Kulkani et al. [16], template-based text generation.

2.5 Previous attempts have proposed to stitch together existing solutions of the above sub-problems, in order to go from an image to its description

2.6 machine translation, where the task is to transform a sentence S written in a source language, into its translation T in the target language, by maximising p(T|S). (encoder and decoder RNN)

2.7 Kiros et al. [15] who use a neural net, but a feedforward one, to predict the next word given the image and previous words

2.8 Mao et al. [21] uses a recurrent NN for the same prediction task

2.9 Lastly, Kiros et al. [14] pro- pose to construct a joint multimodal embedding space by using a powerful computer vision model and an LSTM that encodes text. (better for rankings)

PROBLEMS WITH PREVIOUS WORKS

Heavily hand- designed and rigid when it comes to text generation

Cant describe previously unseen compositions of objects, even though the individual objects might have been observed in the training data.

3. IT DOES

Describe the content of an image using properly formed English sentences

Description must capture not only the objects contained in an image, but it also must express how these objects relate to each other

4. USING

NIC, Neural Image Caption, model based end to end on a neural network consisting of a vision CNN followed by a language generating RNN (for sequence modelling).

CNN as an image “encoder”, by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences

5. PROBLEM RESOLVED/ CONTRIBUTIONS

First, we present an end-to-end system for the problem. It is a neural net which is fully trainable using stochastic gradient descent.

Second, our model combines state-of-art sub-networks for vision and language models. These can be pre-trained on larger corpora and thus can take advantage of additional data.

Finally, it yields significantly better performance compared to state-of-the-art approaches

6. MODEL / METHODOLOGY

A **neural and probabilistic framework** to generate descriptions from images given an image (instead of an input sentence in the source language), one applies the same principle of “translating” it into its description.

1. Maximise the probability of the correct description given the image by using the follow- ing formulation:

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

2. Chain rule to model the joint probability over S_0 , . . . , S_N , where N is the length of this particular example as

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

3. Variable number of words we condition upon up to t 1 is expressed by a fixed length hidden state or memory h_t.

$$h_{t+1} = f(h_t, x_t) \ .$$

4. What is the exact form of f and how are the images and words fed as inputs x_t. For f we use a Long-Short Term Memory (LSTM) net.

5. Training The LSTM model is trained to predict each word of the sentence after it has seen the image as well as all preceding words as defined by p(S_t|I, S_0, . . . , $S_{t 1}$).

6. Our loss is the sum of the negative log likelihood of the correct word at each step as follows: The above loss is minimised w.r.t. all the parameters of the LSTM, the top layer of the image embedder CNN and word embeddings W_e.

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

Used the **BeamSearch** approach in the following experiments, with a beam of size 20. Using a beam size of 1 (i.e., greedy search) did degrade our results by 2 BLEU points on average. (other technique **Sampling**)

7. EVALUATION TECHNIQUES

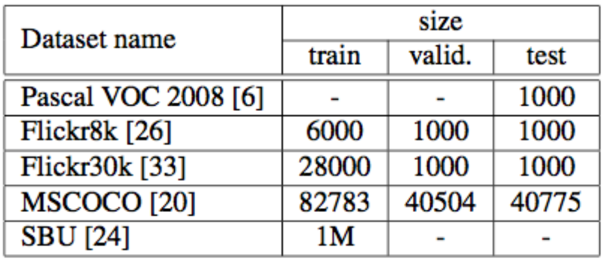
Amazon Mechanical Turk Experiment. Each image was rated by 2 workers. The typical level of agreement between workers is 65%. In case of disagreement we simply average the scores and record the average as the score

BLEU score [25], which is a form of precision of word n-grams between generated and reference sentences.

The **Perplexity** is the geometric mean of the inverse probability for each predicted word.

Proxy task of ranking a set of available descriptions with respect to a given image. **ranking metrics like recall@k.**

8. DATASETS

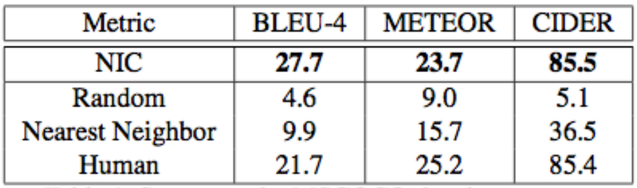


Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [6]	-	-	1000
Flickr8k [26]	6000	1000	1000
Flickr30k [33]	28000	1000	1000
MSCOCO [20]	82783	40504	40775
SBU [24]	1M	-	-

The Pascal dataset is customary used for testing only after a system has been trained on different data such as any of the other four dataset.

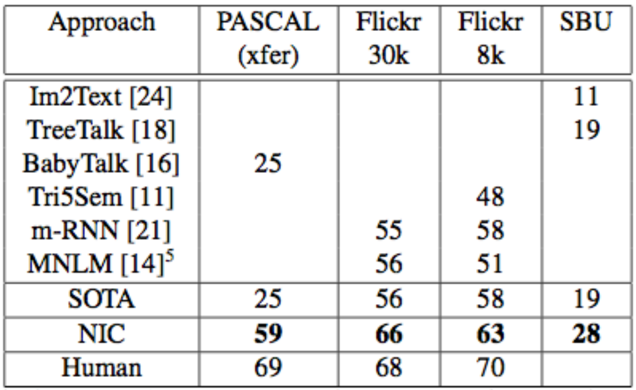
9. RESULTS

We performed experiments on five different datasets:



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

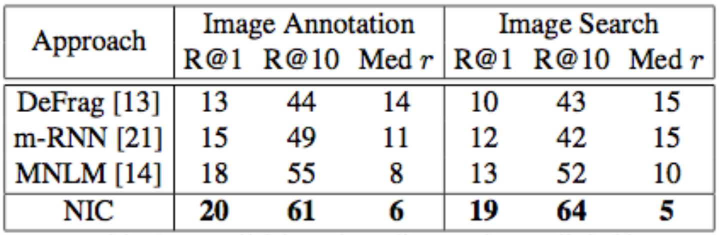
Table 1. Scores on the MSCOCO development set.



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	

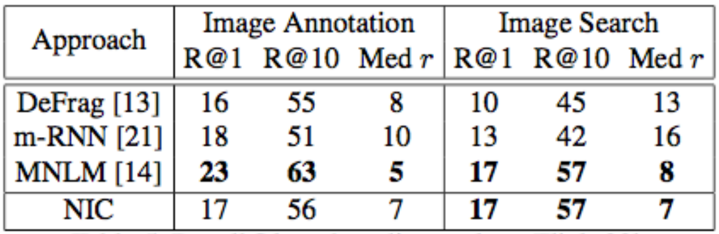
Table 2. BLEU-1 scores. We only report previous work results when available. SOTA stands for the current state-of-the-art.

Since PASCAL does not have a training set, we used the system trained using MSCOCO.



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.

10. CONCLUSION

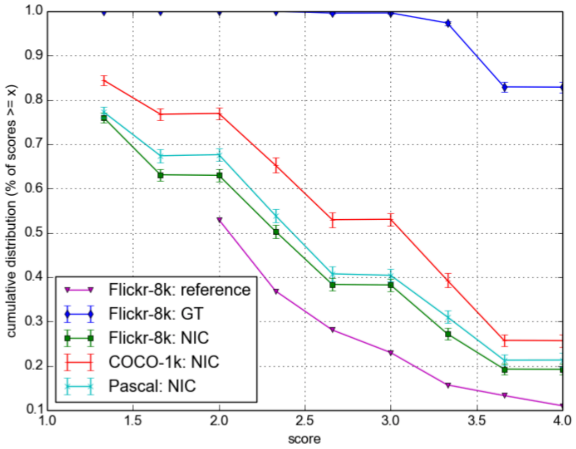


Figure 4. *Flickr-8k: NIC*: predictions produced by NIC on the Flickr8k test set (average score: 2.37); *Pascal: NIC*: (average score: 2.45); *COCO-1k: NIC*: A subset of 1000 images from the MSCOCO test set with descriptions produced by NIC (average score: 2.72); *Flickr-8k: ref*: these are results from [11] on Flickr8k rated using the same protocol, as a baseline (average score: 2.08); *Flickr-8k: GT*: we rated the groundtruth labels from Flickr8k using the same protocol. This provides us with a "calibration" of the scores (average score: 3.89)

Figure 4 shows the result of the human evaluations of the descriptions provided by NIC, as well as a reference system and groundtruth on various datasets. We can see that NIC is better than the reference system, but clearly worse than the groundtruth, as expected. This shows that BLEU is not a perfect metric, as it does not capture well the difference between NIC and human descriptions assessed by raters.

It is clear from these experiments that, as the size of the available datasets for image description increases, so will the performance of approaches like NIC.

11. CHALLENGES FACED

Many of the challenges that we faced when training our models had to do with overfitting. The most obvious way to not overfit is to initialize the weights of the CNN component of our system to a Pre trained model (e.g., on ImageNet)

tried dropout [34] and ensembling model

We trained all sets of weights using stochastic gradient descent with fixed learning rate and no momentum. All weights were randomly initialised except for the CNN weights, which we left unchanged because changing them had a negative impact. We used 512 dimensions for the embeddings and the size of the LSTM memory.

12. FUTURE SCOPE

how one can use unsupervised data, both from images alone and text alone, to improve image description approaches.