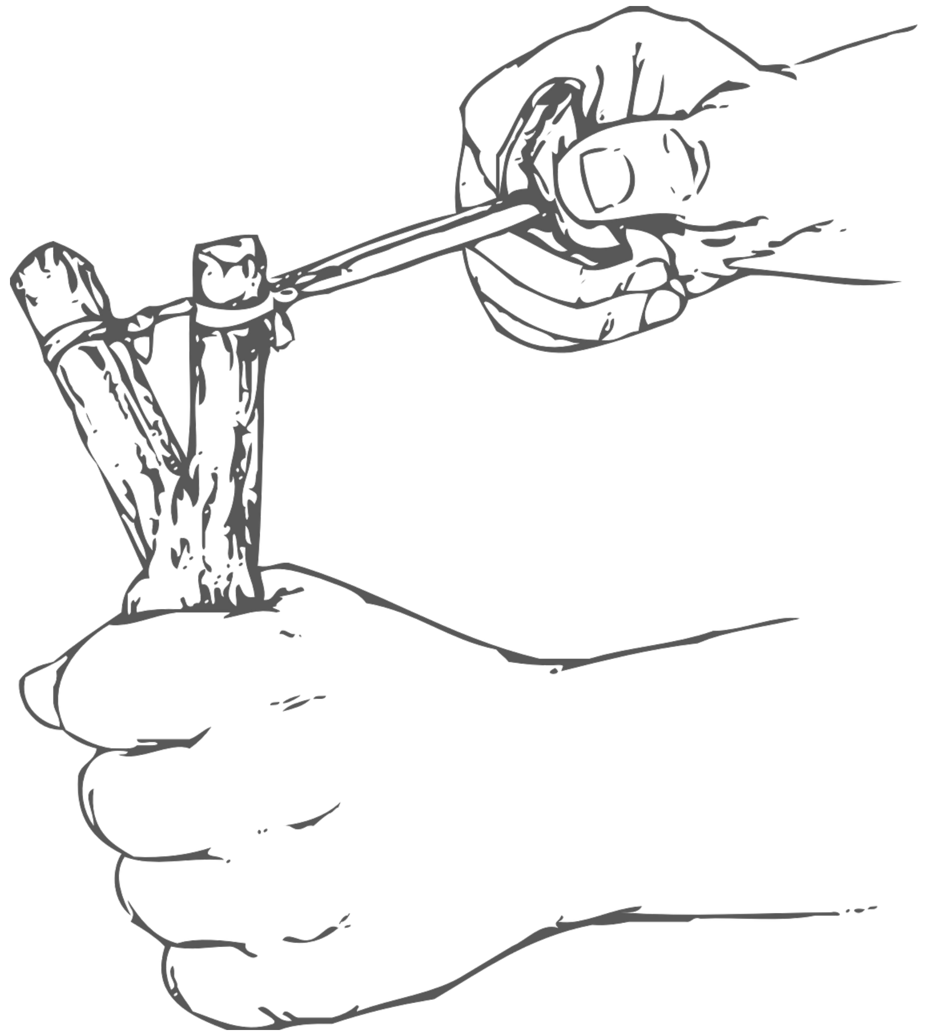
Neural·Pragmatic Natural Language Generation

N·P NLG

Learning goals



1. understand basic architectures for **grounded LMs**
   a. focus on neural image captioning
2. critically assess research papers on (grounded) LMs
3. interpret and apply common **evaluation metrics**

Examples of automatically generated image captions

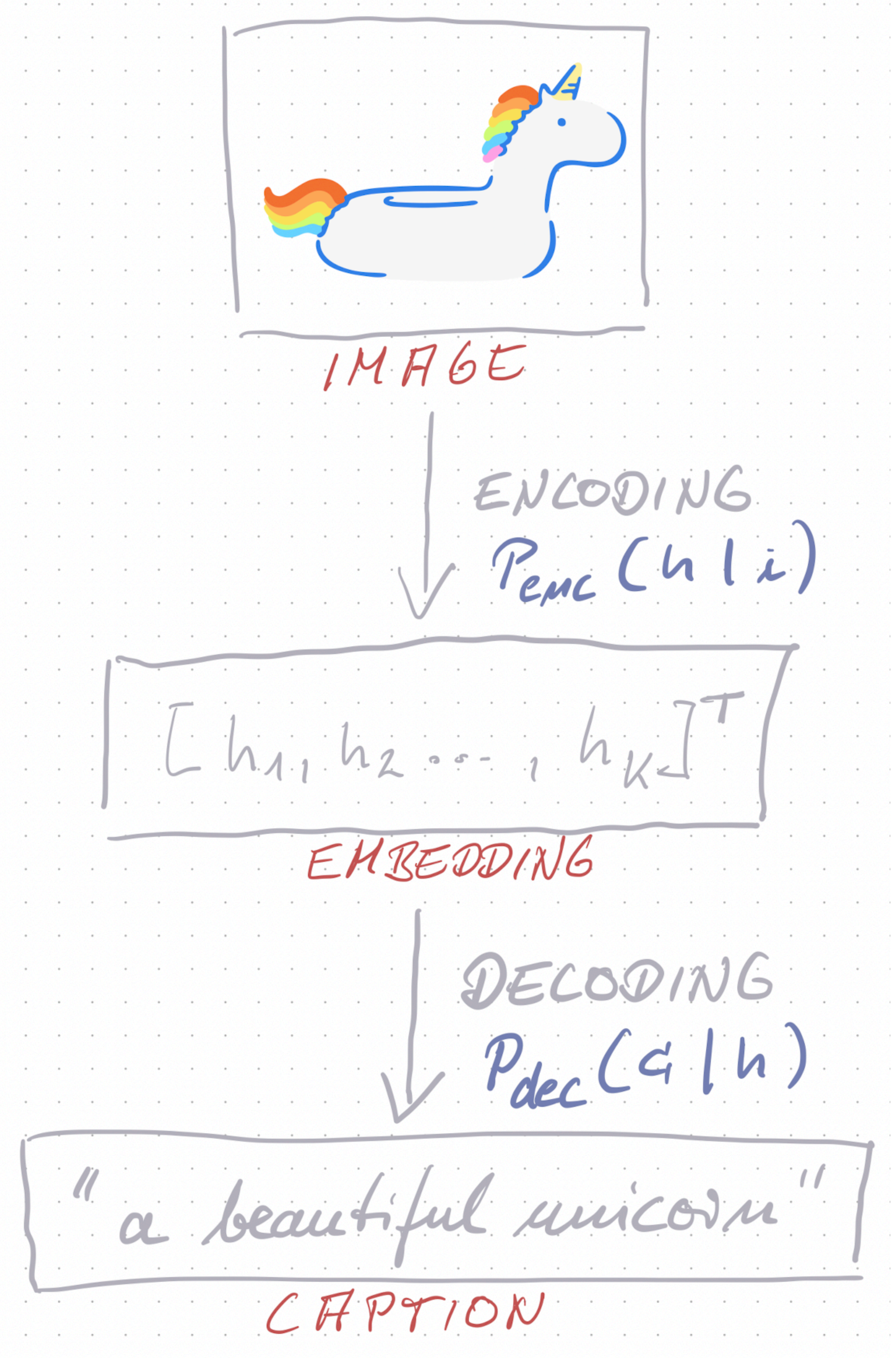
arranged by human evaluation scores

Vinyals et al. (2015) "Show and Tell: A Neural Image Caption Generator"

Encoder-decoder architectures

for grounded language modeling

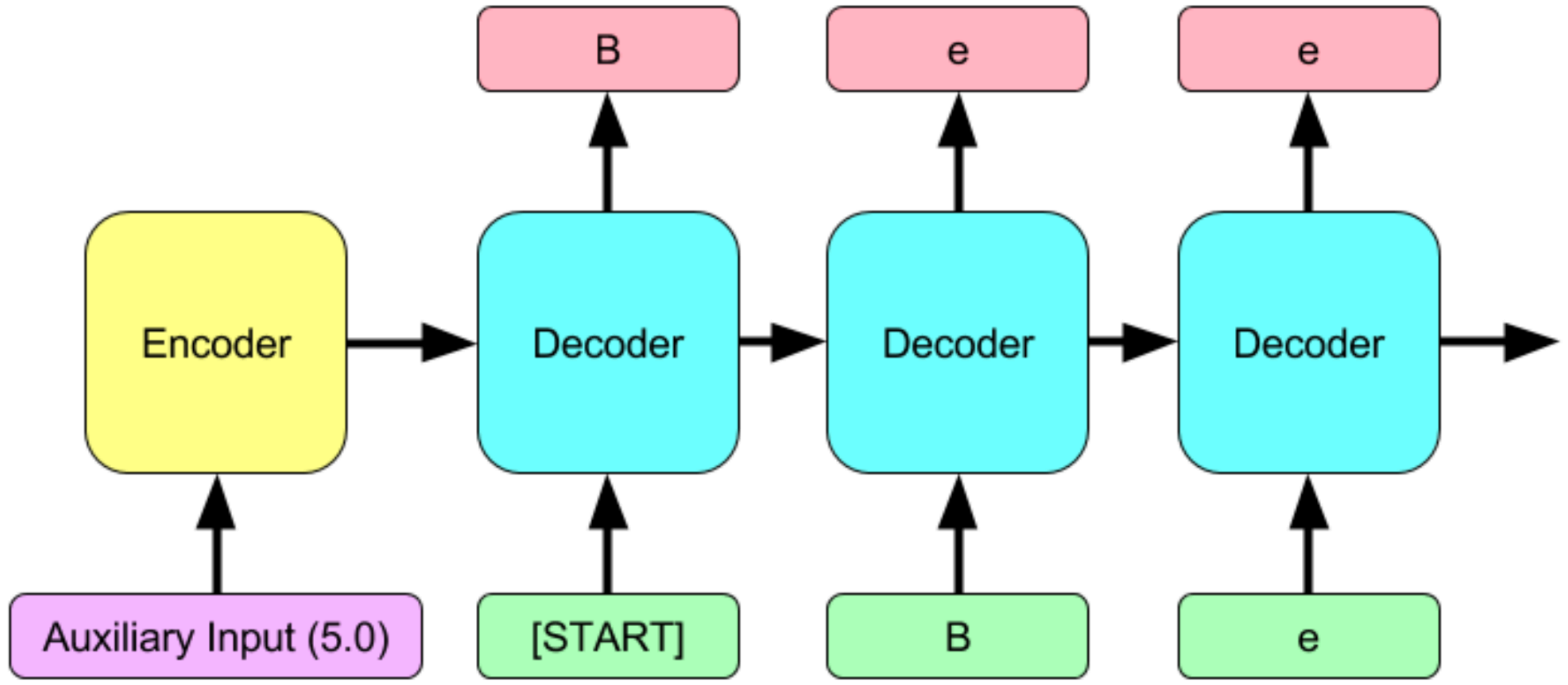
- training data: pairs $\langle i, c \rangle$ of image & caption
  - $c = w_1 \dots w_n$
- objective: approximate true $P(c \mid i)$
- "classical" approach:
  - image → objects, relations → "classical" NLP
- neural approach: encoder-decoder architecture
  - encoder: $P_{enc}(h \mid i)$
    - image embedding (RNN, CNN, ...)
  - decoder: $P_{dec}(c \mid h)$
    - (causal) language model (RNN, LSTM, ...)



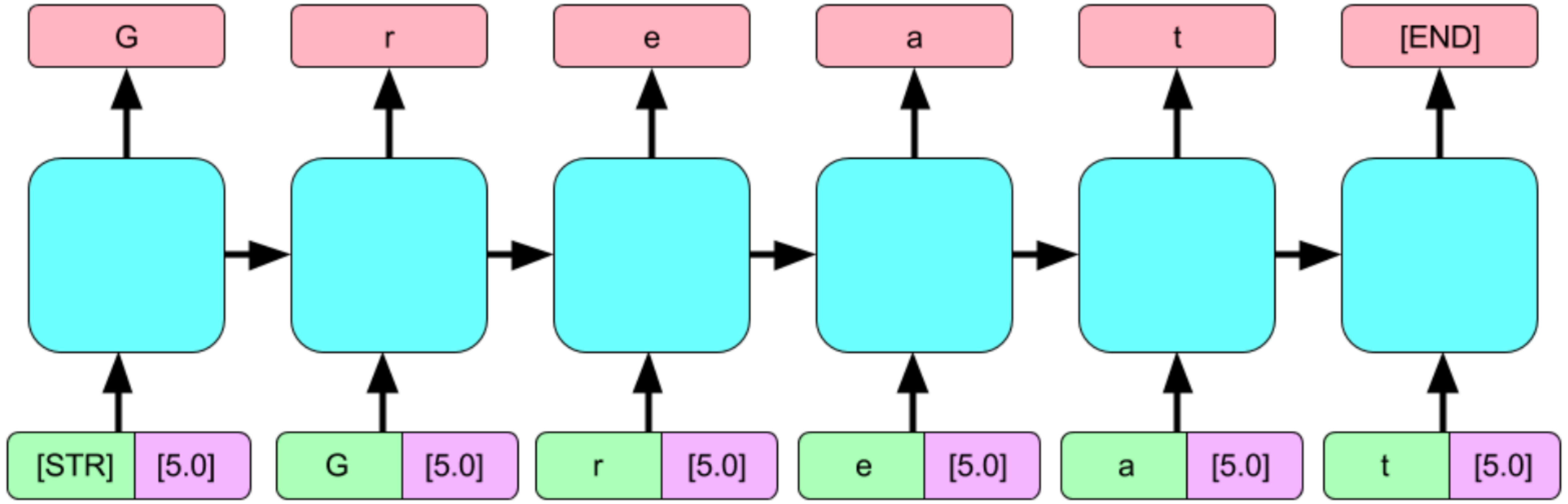
Where to supply the encoding?

initially or repeatedly

initial supply



repeated supply



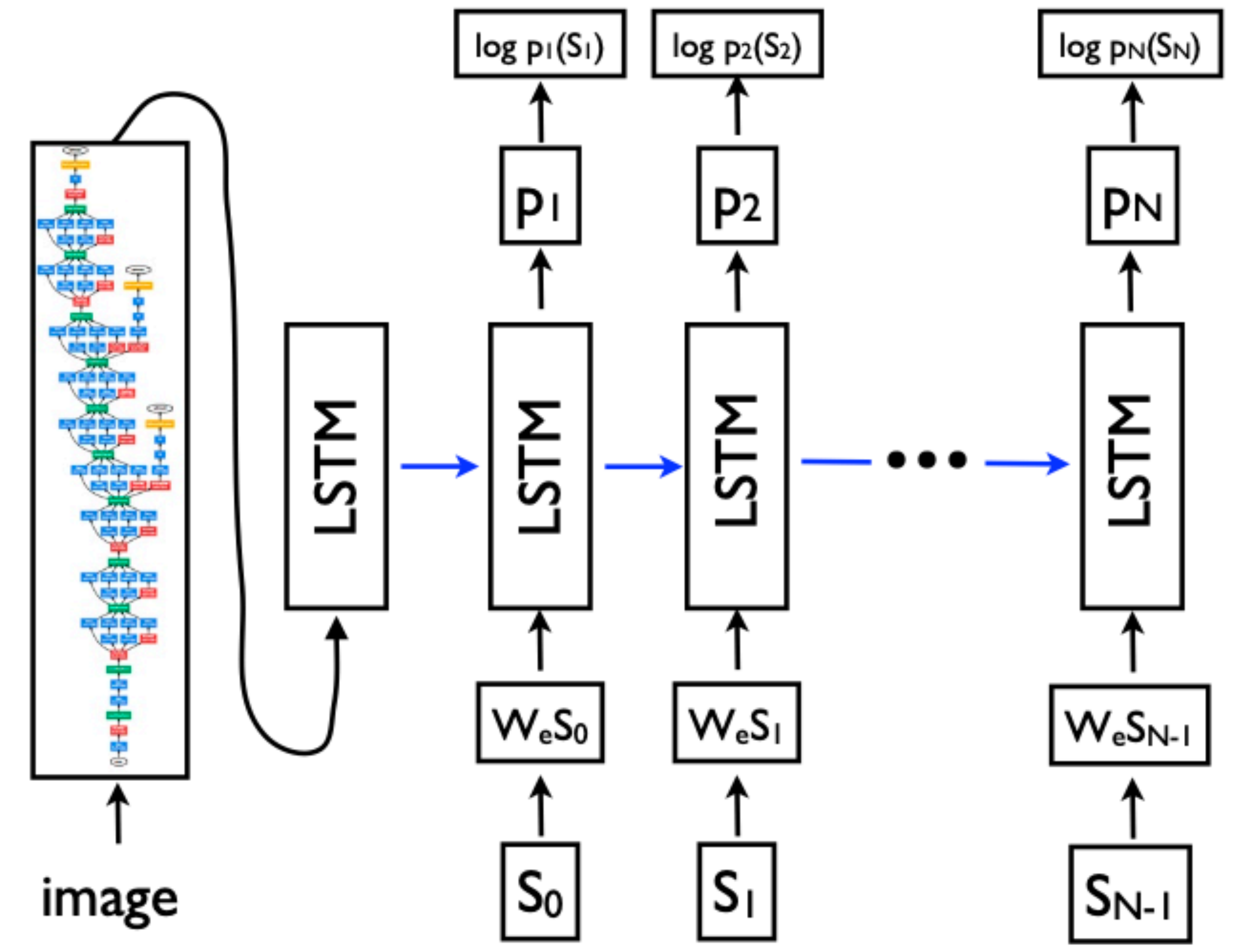
“Show & Tell: A Neural Image Caption Generator”

Vinyals et al. (2015)

Neural Caption Generator

Vinyals et al. (2015)

- encoder:
  - CNN
  - pretrained on ImageNet
- decoder:
  - LSTM, (hidden layer size: 512)
  - initialized with random embeddings
- decoding strategies:
  - pure sampling
  - beam search (beam size 20)
- training specs:
  - objective function: surprisal
    $$-\log P(c \mid i) = -\sum \log(w_{i+1} \mid w_{1:i}, c)$$
  - vanilla gradient descent



initial supply of image embedding

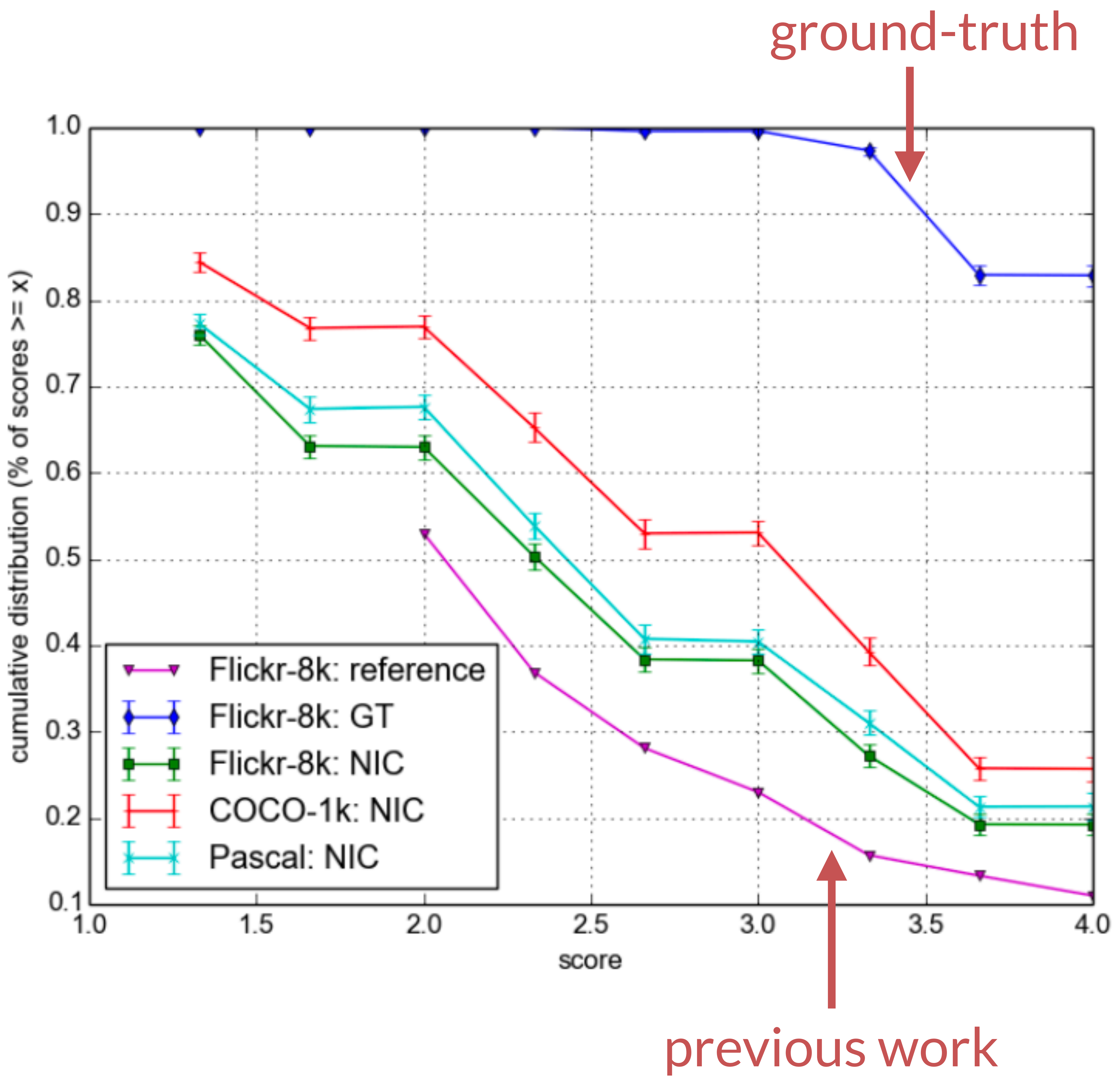
Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [6]	-	-	1000
Flickr8k [26]	6000	1000	1000
Flickr30k [33]	28000	1000	1000
MSCOCO [20]	82783	40504	40775
SBU [24]	1M	-	-

data sets & their split sizes

Human Evaluation

Vinyals et al. (2015)

- each image rated by two human rater
- scale from 1 to 4
- images paired with model-generated captions or a ground-truth caption from the data set



Evaluation metrics

Vinyals et al. (2015)

- perplexity
  - used only for model comparison and tracking training progress
- BLEU-n
  - co-occurence on n-grams between generated and reference sequences (Papineni et al., 2002)
  - correlates well with human quality judgements
  - easy to compute but may depend on tokenizer (what counts as a word)
- METEOR
  - based on harmonic mean of unigram precision and recall (Banerjee & Lavie 2005)
  - intended as improvement over BLEU
  - matching target and output via exact matching, synonymy, stem-identity ...

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.

- CIDER
  - specific to image captioning (Vedantam 2014)
  - score each caption to set of ground-truth reference captions
  - use only stem/root forms
  - score based on:
    - how often n-gram is present in reference set
    - how often it occurs in any other reference set