GREEN: A NEW AUTOMATED EVALUATION METRIC FOR IMAGE TO TEXT

Vedantam R., Zitnick L., Parikh D.
{vrama91, parikh}@vt.edu, larryz@microsoft.com

ABSTRACT

We introduce GREEN, a new automatic metric for evaluating approaches that generate descriptions of images. We show that GREEN matches human evaluation better than existing metrics (BLEU and ROUGE). We also introduce two new datasets where each image is described with 50 reference sentences. We show that all evaluation metrics correlate much better with humans with additional references. We evaluate five state-of-the-art image description approaches as well as human generated descriptions.

Im2text

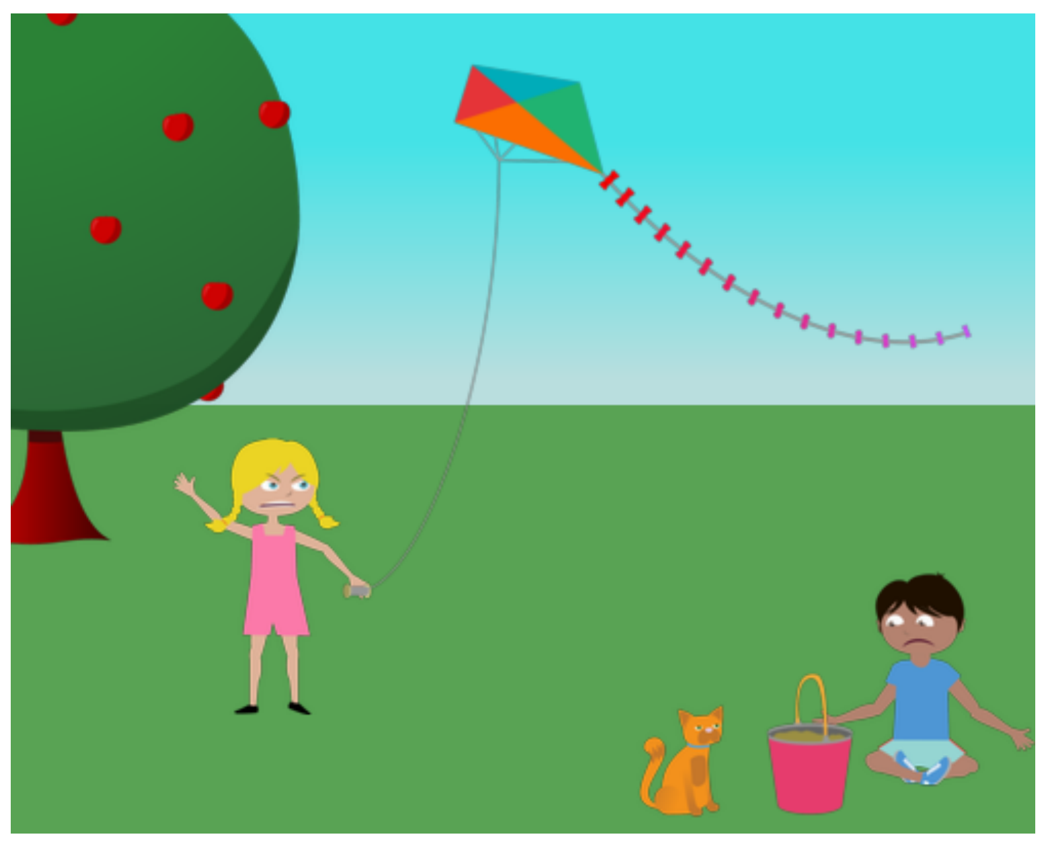
Can be viewed as:

- Text Summarization: *Summarizing Image Content*
- Machine Translation: *Translation from Image content to language content*

Im2text approaches are:

- Retrieval based
- Generation Based

Datasets



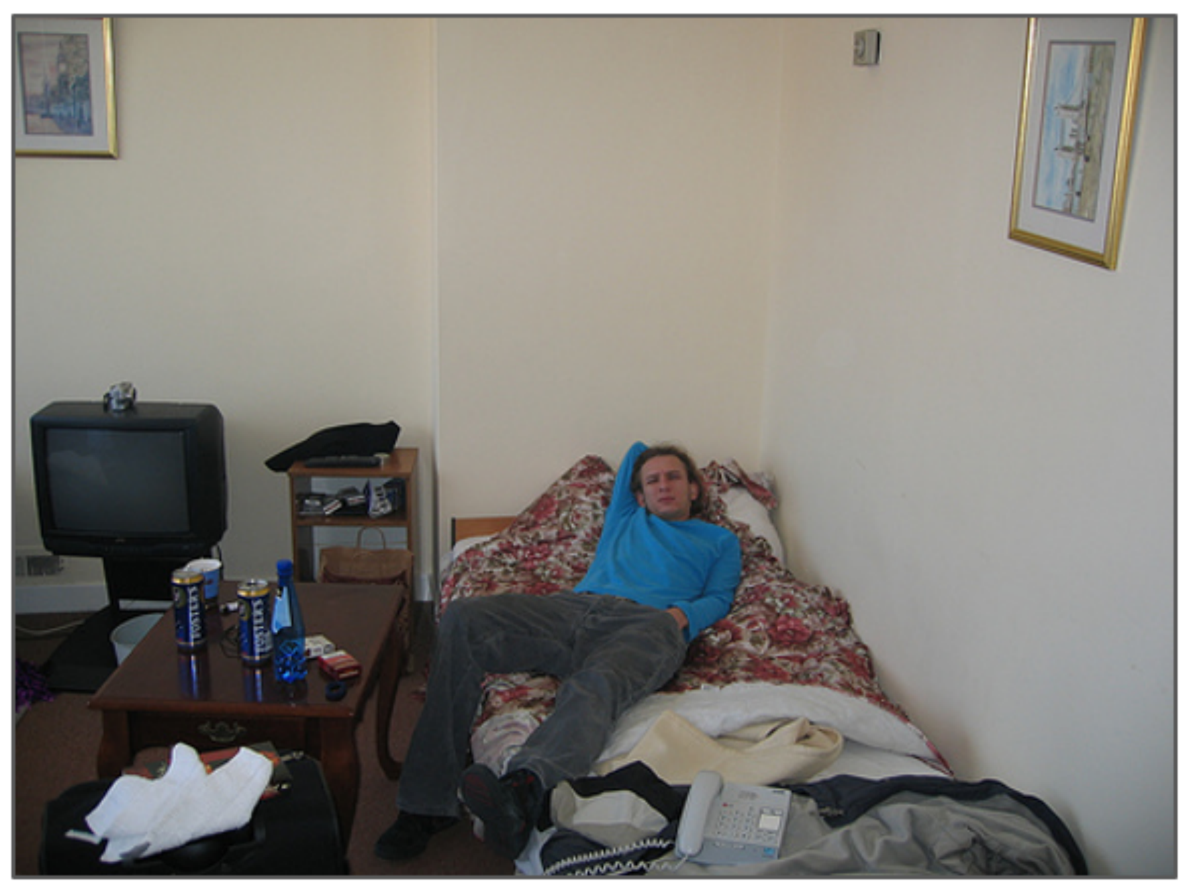
ABSTRACT-50S
500 Images- 50 Sentences

PASCAL-50S
1000 Images – 50 Sentences

Human Agreement

Rate: How well does sentence describe image?
1 to 5 scale
Human scores compared to metric scores

Variation in Sentences



A man laying on a bed with beer cans on a nearby table.

A man lies in a bed in a cluttered room.

a middle-aged man relaxes after drinking a few beers

A person in a blue shirt reclines near a coffee table and television.

Young man in blue shirt lying down on couch.

A black train engine is facing me on the tracks with its light on in the woods during the day.

A close-up of a black train engine.

A front view of a black train engine on the tracks with people standing on either side.

A group watches an old freight train engine.

The train is very old and has the number 1225 on it

GREEN

Sentence level importance

Images

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_l h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{i=1}^{|I|} \min(1, \sum_l h_k(s_{il}))}\right)$$

TF

IDF

Captures Saliency / Importance

$$GREEN_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g}(c_i) \cdot \boldsymbol{g}(s_{ij})}{\|\boldsymbol{g}(c_i)\| \|\boldsymbol{g}(s_{ij})\|}$$

Cosine Similarity

Sentence level accuracy + mean

$$\boxed{GREEN(c_i, S_i) = \sum_{n=1}^{N} w_n GREEN_n(c_i, S_i)}$$

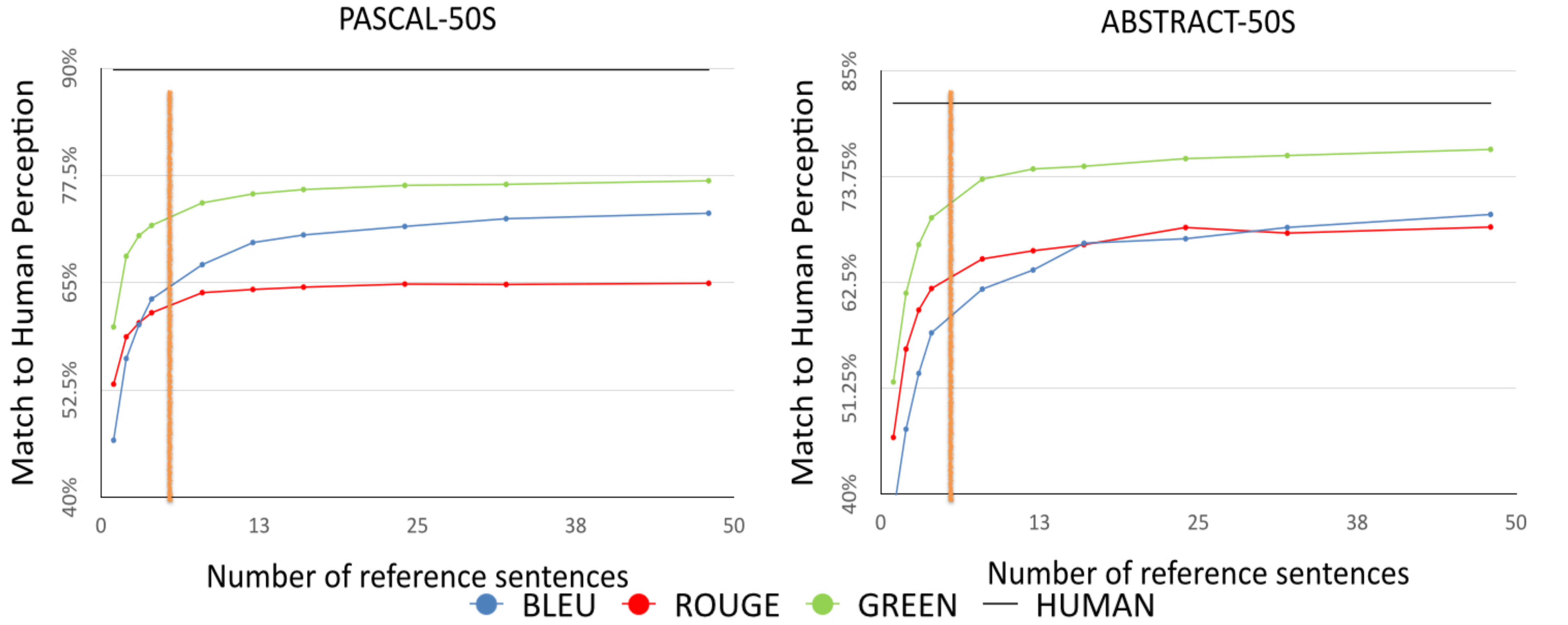
Grammaticality

Typically N=4 is chosen and higher n-grams are given higher weights

ROUGE VS BLEU VS GREEN

- ROUGE and BLEU do not measure saliency
- BLEU does a max across reference sentences (sensitive to outliers)
- BLEU computes precision, ROUGE computes recall, GREEN computes accuracy
- ROUGE is recall based, can be gamed with an excessively long sentence

Results



Midge (Mitchell et. al) is the best performing Im2text method on both GREEN and HUMAN

Conclusions

- GREEN Correlates best to humans
- Even at 5 sentences, GREEN is best
- More sentences make a big difference
- Code and dataset will be released