Rich Image Captioning using Saliency Maps

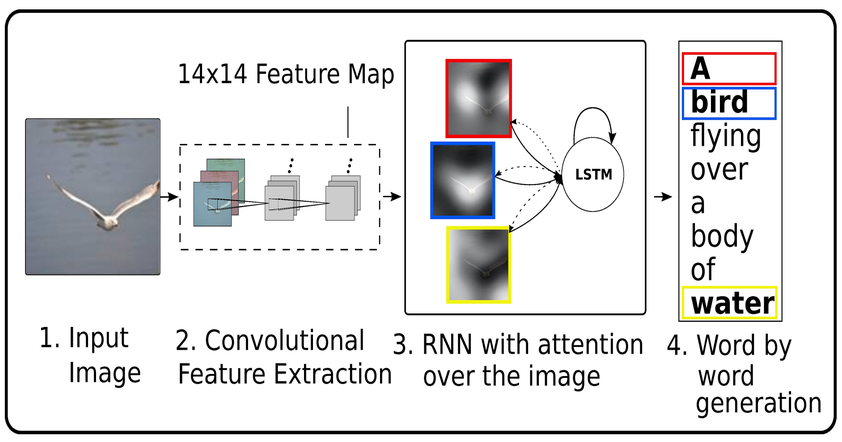
Sainandan Ramakrishnan & Rahul Solanki
(sainandancv & solankirahul)@gatech.edu

MOTIVATION

Connecting vision with language, the two most paramount tokens of human intelligence, to automatically describe content in images, with long standing applications in social media, video surveillance, information retrieval and assisting the visually impaired.

ABSTRACT

While previous models use a deep language model, they lose sense of the image after a few time steps. We restore the importance of the image by exploiting the concept of saliency.



OBJECTIVES

- To explore and demonstrate the importance of saliency maps generated from attention models, in understanding an image enough to describe it.
- To demonstrate how saliency can result in richer and more image-specific captions which have implications for scene understanding applications.
- To efficiently estimate the best saliency parameters from indirect observations.

APPROACH

- Express caption word at every time step as function of the past and the saliency map.
- Express the output of the attention model (the saliency map) as a linear combination of spatial image features.
- Extract spatial image features from the input image using a pre-trained VGG-19 net.

RESULTS

- Models solely using language models to decode suffer from generic captions that forget the image.
- Using attention, we generate captions that are truer to the scene and extract key concepts better.

Flickr30k test set	SHOW AND TELL	OURS
BLEU	0.318	**0.376**
METEOR	0.131	**0.183**
CIDeR	0.212	**0.3134**
ROUGE	0.208	**0.299**


A man in a black shirt is playing a guitar


A man with tattoo on his arm plays with a saw


A little boy in a blue shirt is playing with a toy


A little boy is playing with a toy truck

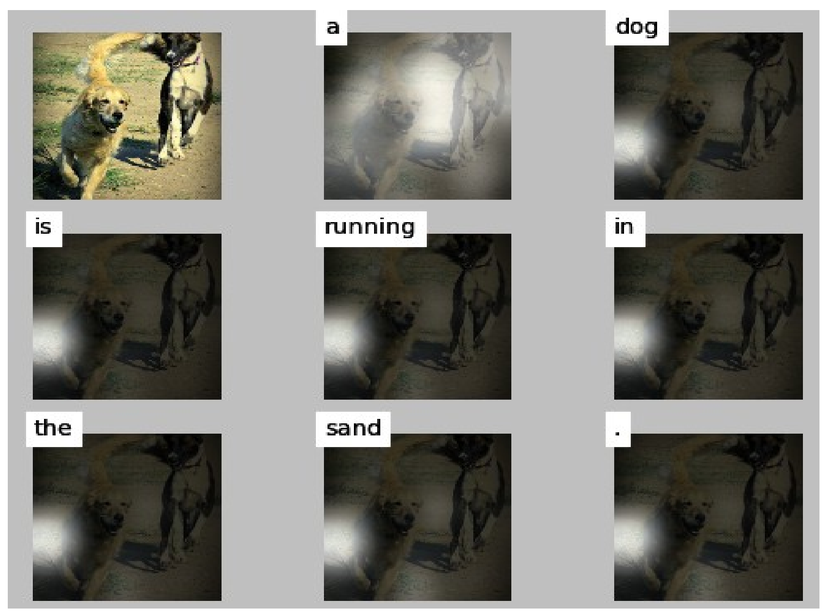

A man in a blue shirt is walking on a snowy hill

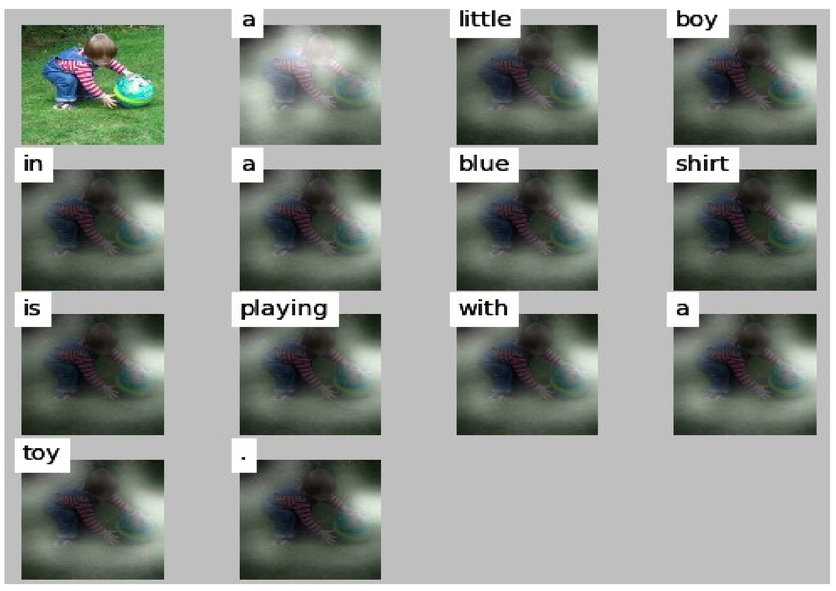

A man in a black jacket and tan pants is walking on a beach


A dog is running in the sand


A dog is shaking off water with something in its mouth

Ours (left) and *Show & Tell* (right)





CONCLUSION

- We validate both qualitatively and quantitatively, the immense improvement in captioning quality achieved by using saliency maps.
- The approach of learning attention from data can be extrapolated to any problem where selective restoring of information is required.
- We analyze the interpretability and meaning of saliency maps generated, and note its implications.