Top-Down Visual Saliency Guided by Captions

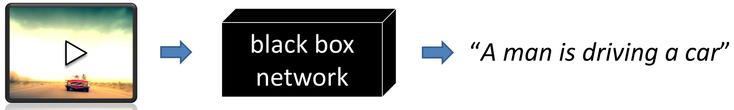
Vasili Ramanishka[1] Abir Das[1] Jianming Zhang[2] Kate Saenko[1]

[1]Boston University [2]Adobe Research

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 HONOLULU 2017

Overview



Problem:

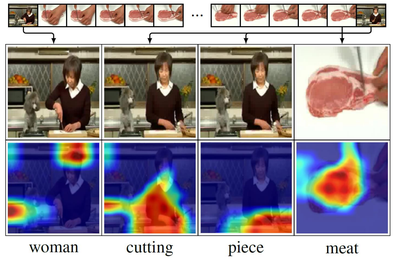
- Encoder-decoder networks for scene captioning work as black boxes
- How can we *explain* the model's captions?
- Can we extract salient regions for each generated word?

Our Contributions:

- A method to extract salient spatio(-temporal) regions for each predicted word or phrase in encoder-decoder networks
- Our method works without requiring region-level annotations or the overhead of explicit attention layers
- Query sentences not produced by the network can also be mapped to regions, allowing weakly-labeled segmentation

Motivating Example:

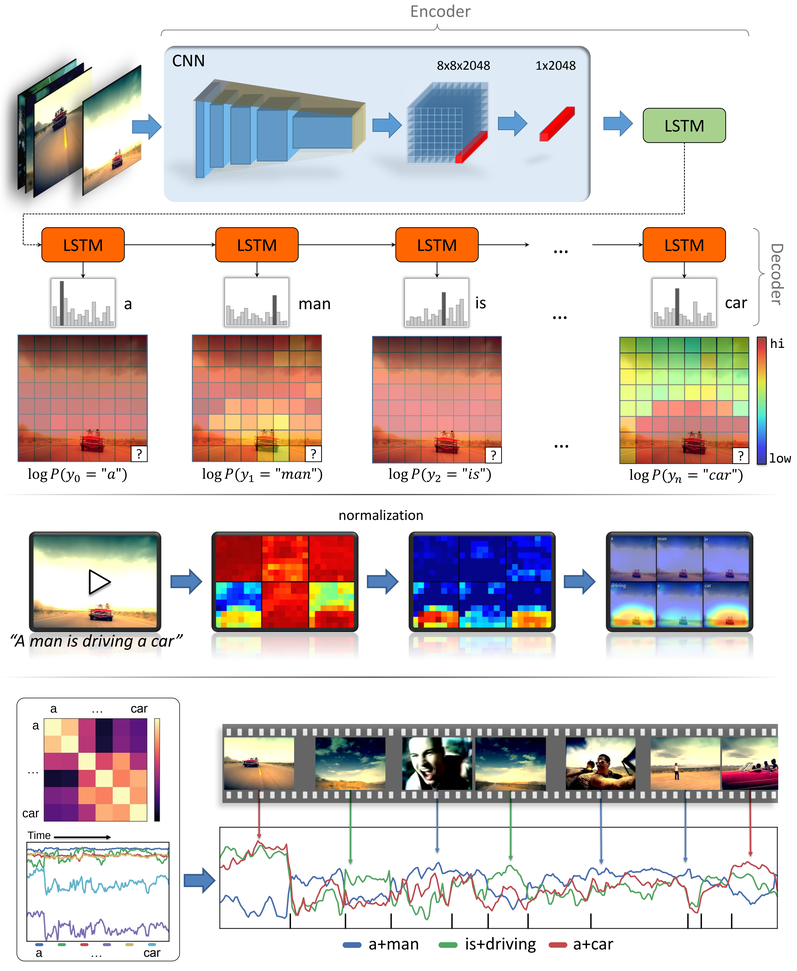
- Encoder-decoder predicts caption: *A woman is cutting a piece of meat.*
- Is "woman" generated because the model recognized a woman or merely because "A woman" is a likely way to start a sentence?



Our Solution:

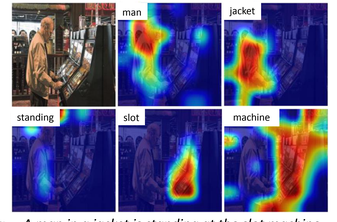
- Estimate spatiotemporal saliency for videos (or spatial saliency for images) for each word in the predicted sentence description
- Do this by measuring the drop in word probability when only one small part of the input video is fed into the network

Approach

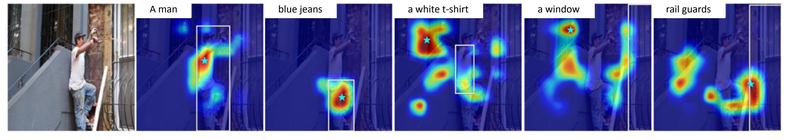


Evaluation

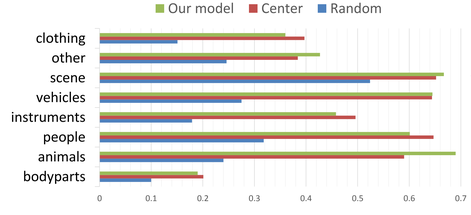
- For Flickr30kEntities [1], salient image regions of the words are obtained by sequential encoding of spatial descriptors in a similar encoder-decoder framework.
- GT bounding boxes were used **only** during evaluation



Input: *A man in a jacket is standing at the slot machine*



Pointing Game Performance



Task:

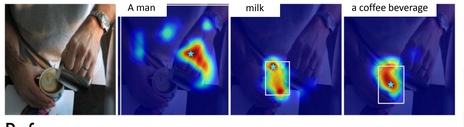
- Given an image and a noun phrases, point to the targets

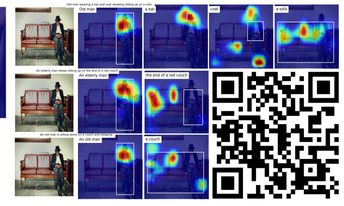
Metric:

- Mean pointing accuracy across all noun phrases
- Pointing anywhere on the targets is fine

Attention Correctness [2] is defined as a sum of pixel-level attention values which lie inside the bounding box

	Avg. per NP
Uniform Baseline	0.321
Soft Attention [3]	0.387
Soft Attention Supervised [2]	0.433
Our method	0.473





References:

[1] B. A. Plummer et al., Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, ICCV 2015

[2] C. Liu et al., Attention Correctness in Neural Image Captioning, AAAI 2017

[3] K. Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

http://ai.bu.edu/caption-guided-saliency