YOLO-RNN image caption

Tetsumichi(Telly) Umada

Project Overview

Combine the output of a YOLO classifier with a word embedding to understand scene context

Built on: Show and Tell: A Neural Image Caption Generator
Vinyals et al.

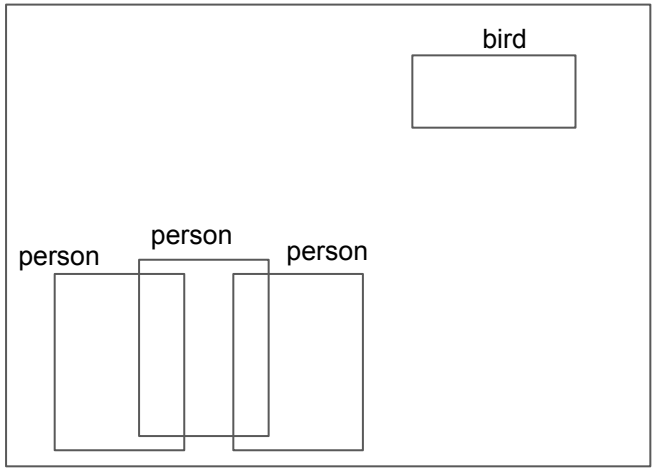
Project Overview

yolo

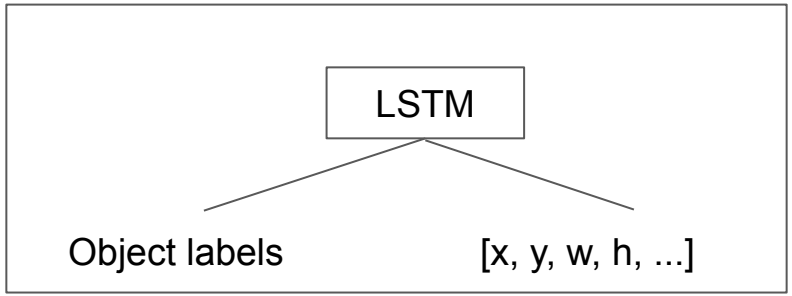
A dog is sitting next to a read bike.

Model overview

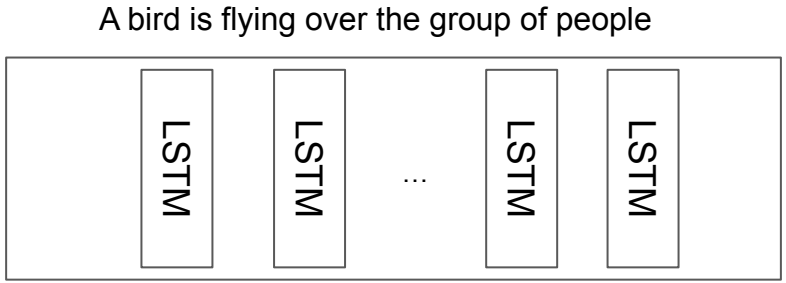
Object locations



LSTM encoders



LSTM decoders



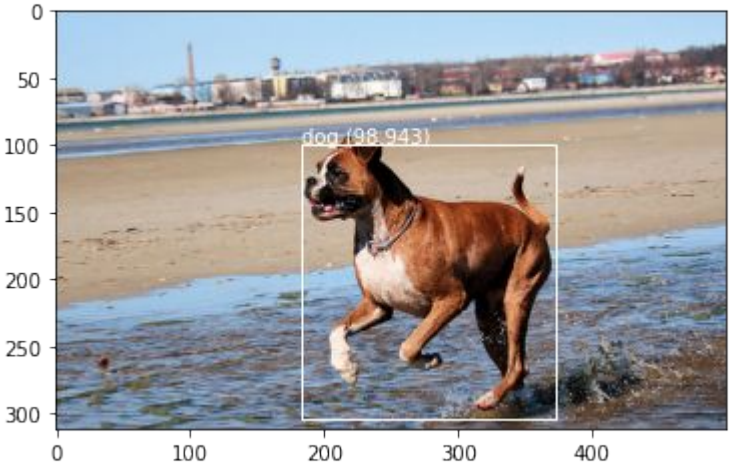
Data set and experiments

Flickr 8k Data

- 6000 training image, 1000 dev images, 1000 test images
- Each image has 5 captions

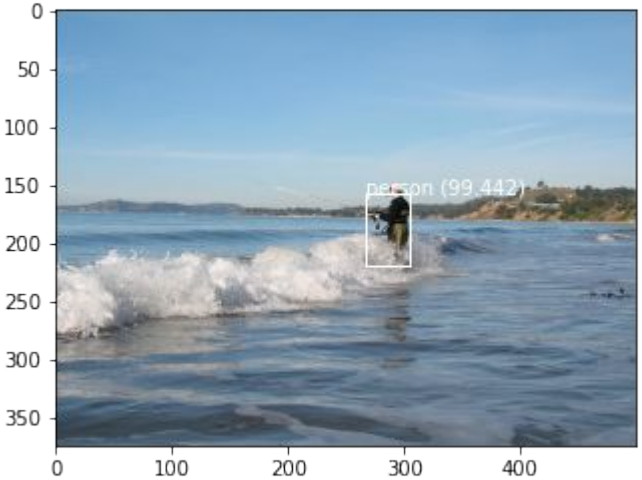
- For this project, we use 500 training image.

Sample outputs



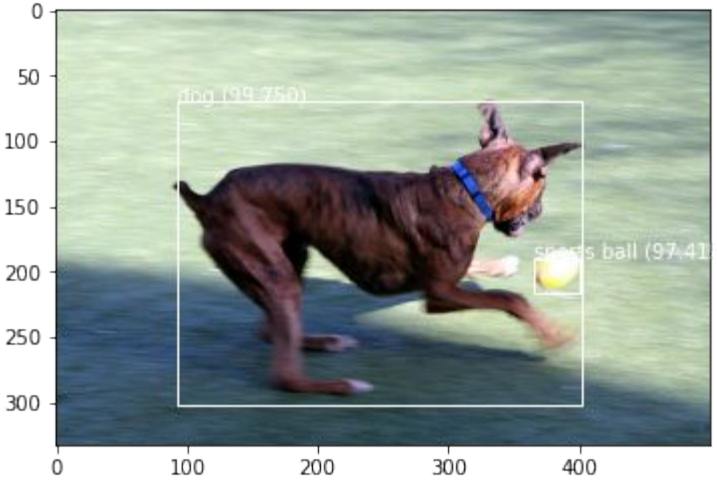
a brown dog is running through the grass.

Sample outputs



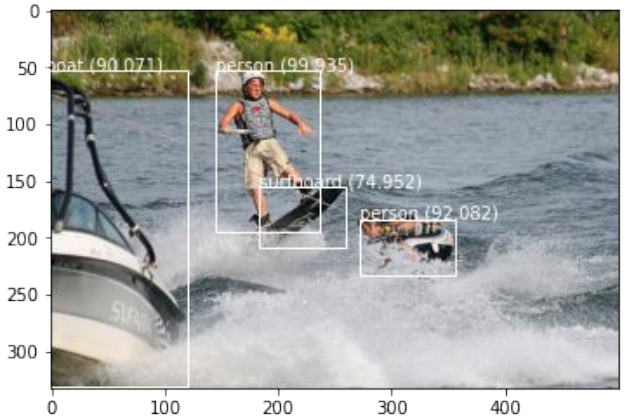
a man is running on a <UNK> .

Sample outputs



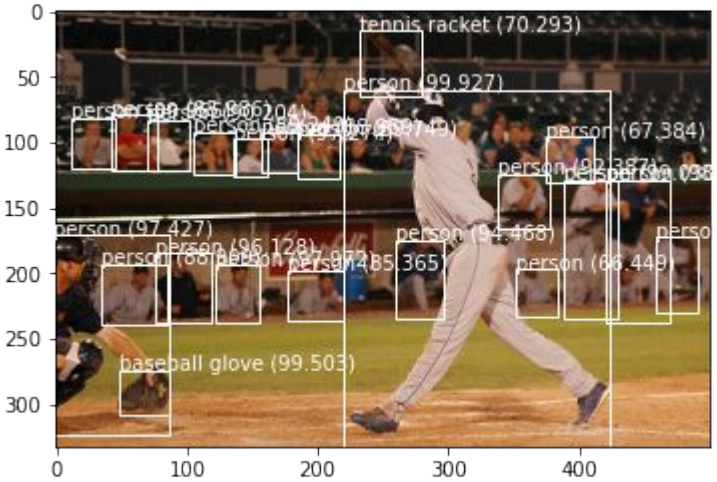
Generated: A dog running in the grass .
Annotation: A dog chases after a yellow ball .

Sample outputs



a man in a <UNK> <UNK> <UNK> .

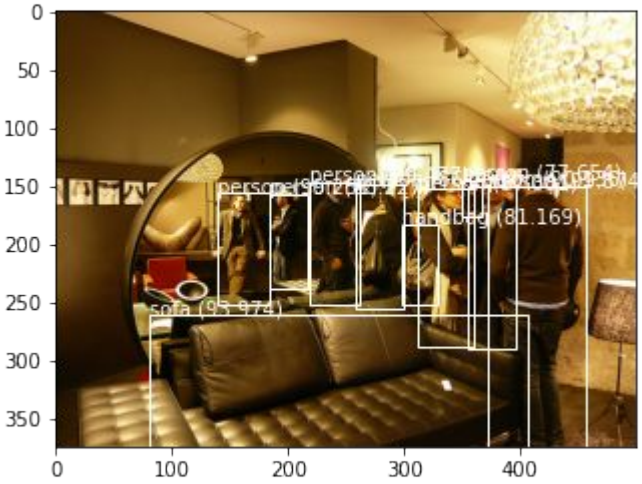
Sample outputs



Generated: a <UNK> <UNK> a <UNK> a a <UNK> . a <UNK> . a <UNK> . a <UNK> . a <UNK> . a <UNK> . a <UNK> . a <UNK> . a <UNK> . a <UNK> . a

annotation: A baseball player near home plate .

Sample outputs



a man in a <UNK> <UNK> <UNK> .

Future work

- Train with full data with more epochs
- Filter the images with more objects and spatial relation (select images if the caption contains in/over/on etc..)
- Add CNN to encode image features

Any suggestions? Questions?