Image Caption Generation with CNN-RNN, Soft Attention and Top-Down Bottom-Up attention

Dan DiTommaso
Robin Bhattacharya
Saurabh Kulshreshtha

Introduction

- Image Captioning is one of the most fundamental tasks in Machine Learning.

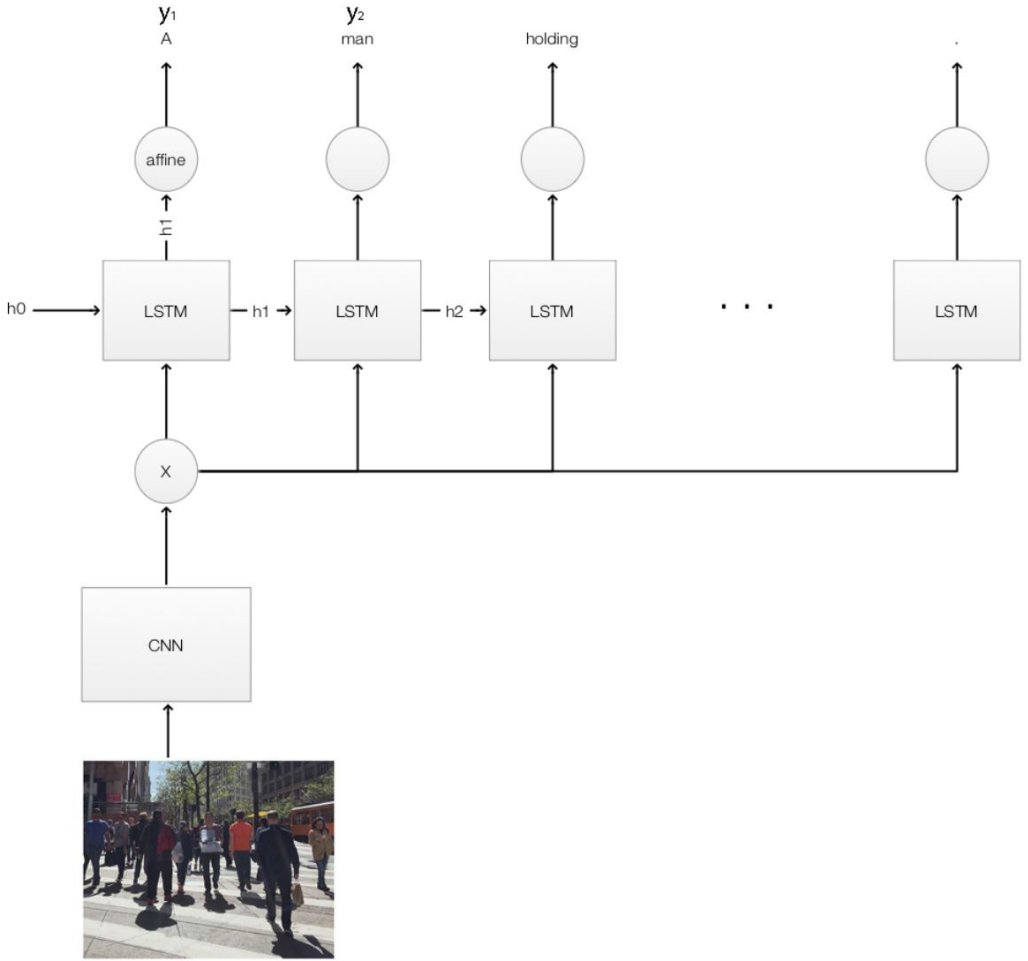
- There has been many approaches in the recent past.

- We implement 3 approaches which were all state-of-the-art at that time, evaluate them and compare it with the results given in the standard paper.

Show and Tell

ResNet CNN feeds into LSTM

Which predicts the captions



To generate an image caption with deep learning, we start the caption with a “start” token and generate one word at a time. We predict the next caption word based on the last predicted word and the image:

$$\text{next word} = f(\textit{image}, \text{last word})$$

Applying the RNN techniques, we rewrite the model as:

$$h_t = f(x, h_{t-1})$$

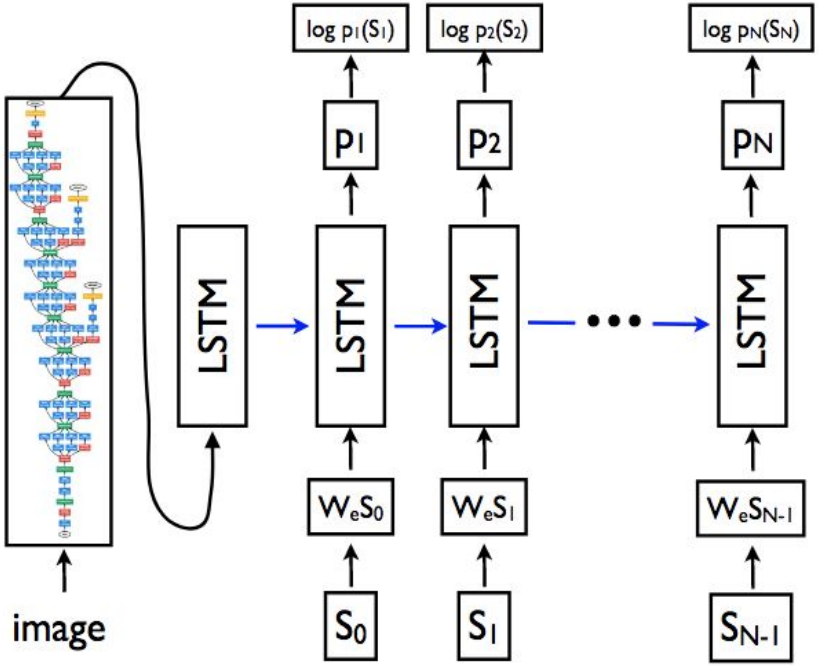
$$\text{next word} = g(h_t)$$

which x is the image, and h_t is the RNN hidden state to predict the “next word” at time step t.

Show and Tell

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$



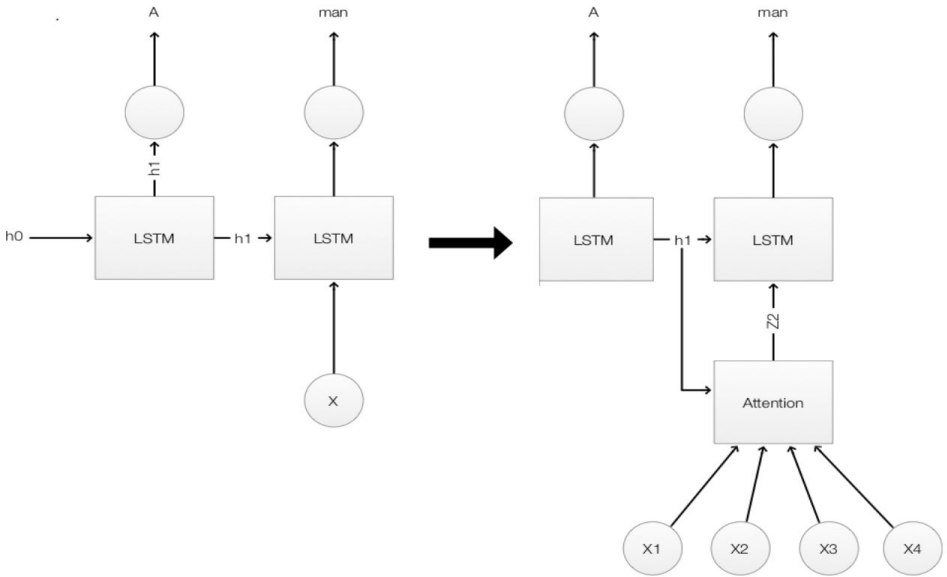
Attention

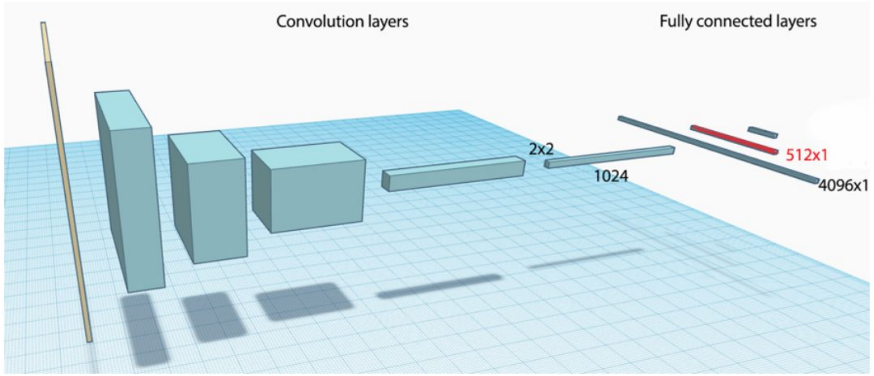
Mathematically, we are trying to replace the image x in LSTM model,

$$h_t = f(x, h_{t-1})$$

with an attention module $attention$:

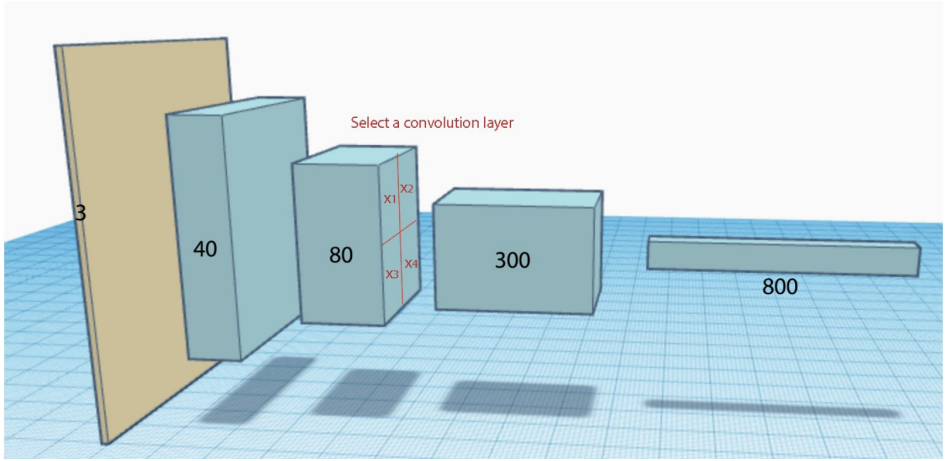
$$h_t = f(attention(x, h_{t-1}), h_{t-1})$$

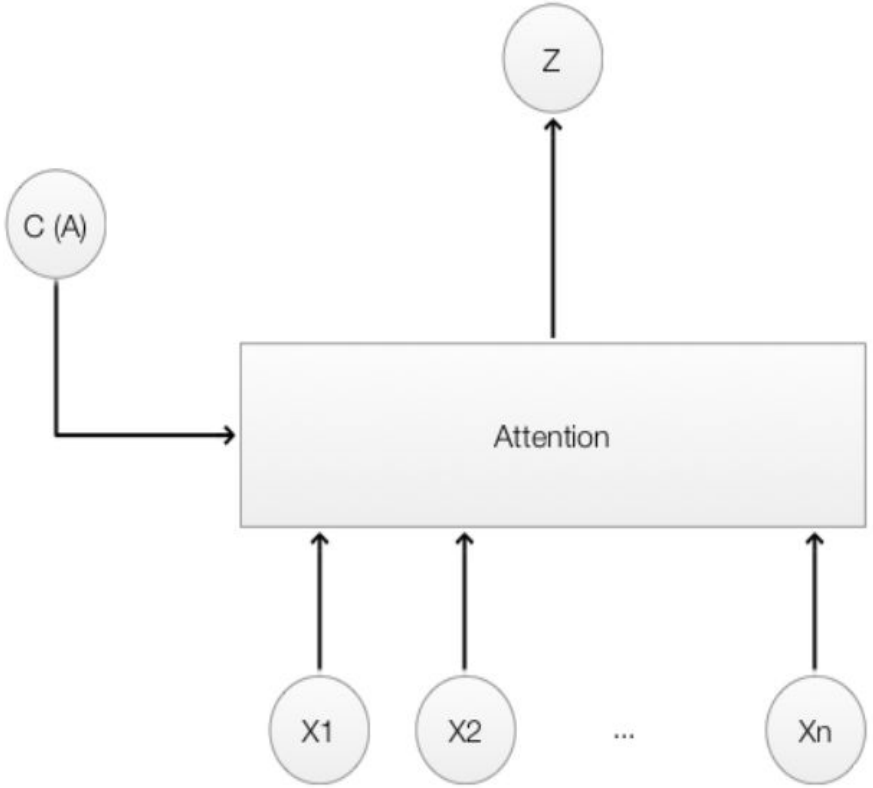


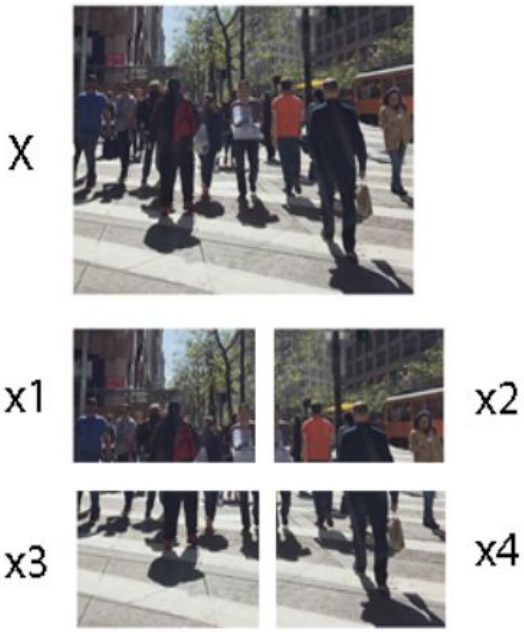


Spatial information is lost in CNN last layer

Select a Convolution Layer and divide the feature maps




Z
C (A)
Attention
X1
X2
...
Xn


X
x1
x2
x3
x4

Soft Attention

Let's show how to compute the weighted features for the LSTM. x_1, x_2, x_3 and x_4 each covers a sub-section of an image. To compute a score s_i to measure how much attention for x_i, we use (with the context $C = h_{t-1}$):

$$s_i = \tanh(W_c C + W_x X_i) = \tanh(W_c h_{t-1} + W_x x_i)$$

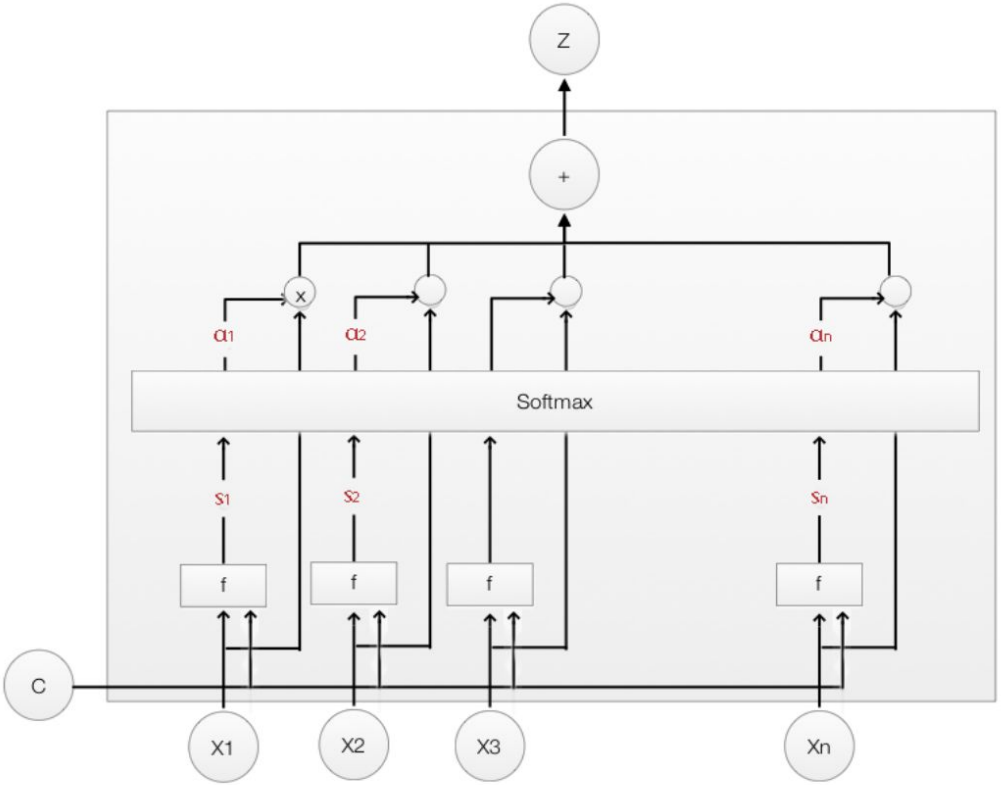
We pass s_i to a softmax for normalization to compute the weight α_i.

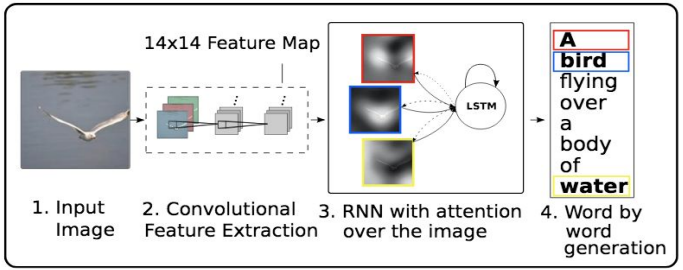
$$\alpha_i = softmax(s_1, s_2, \ldots, s_i, \ldots)$$

With softmax, α_i adds up to 1, and we use it to compute a weighted average for x_1, x_2, x_3 and x_4

$$Z = \sum_i \alpha_i x_i$$

Show, Attend and Tell





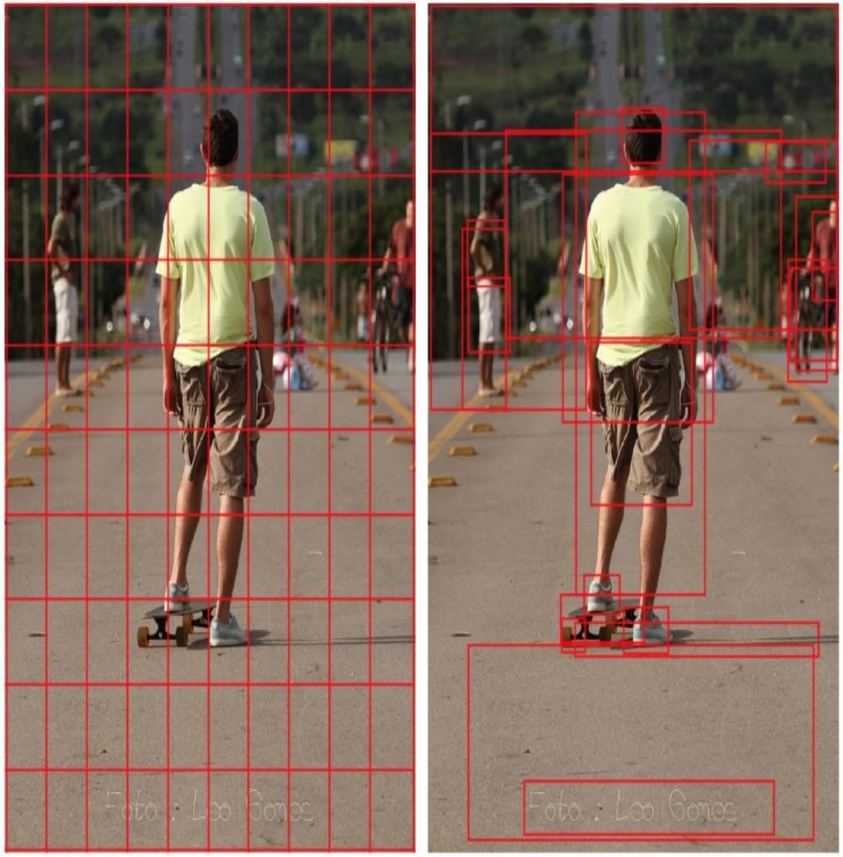
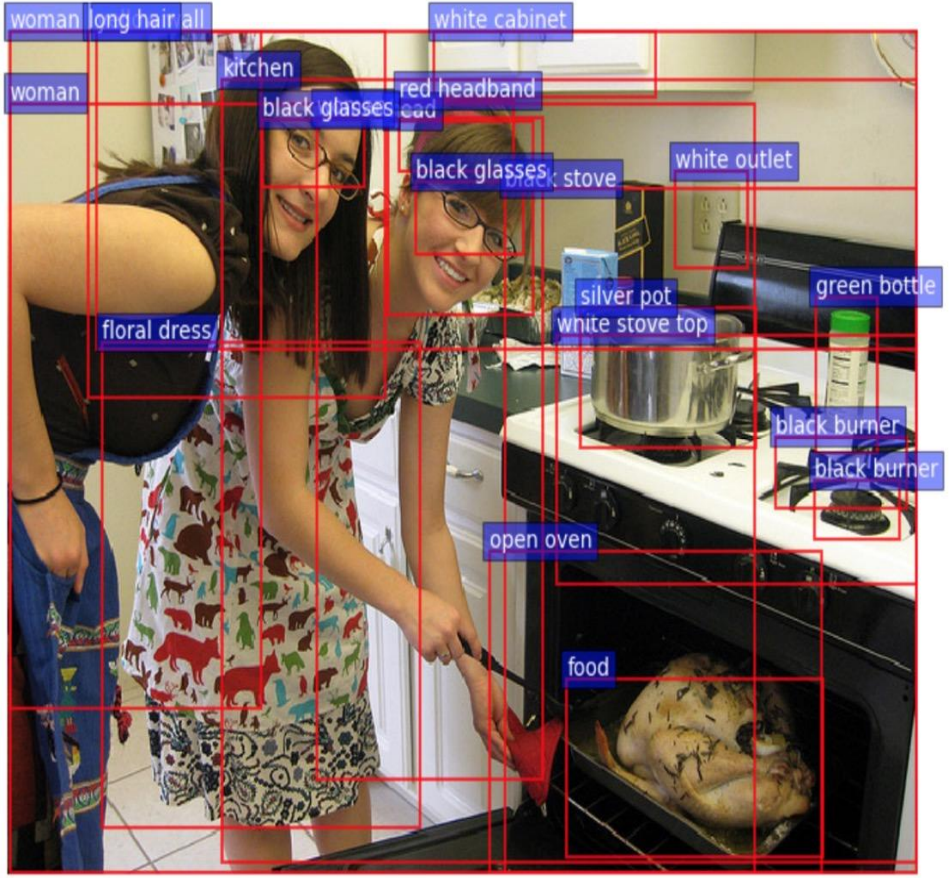
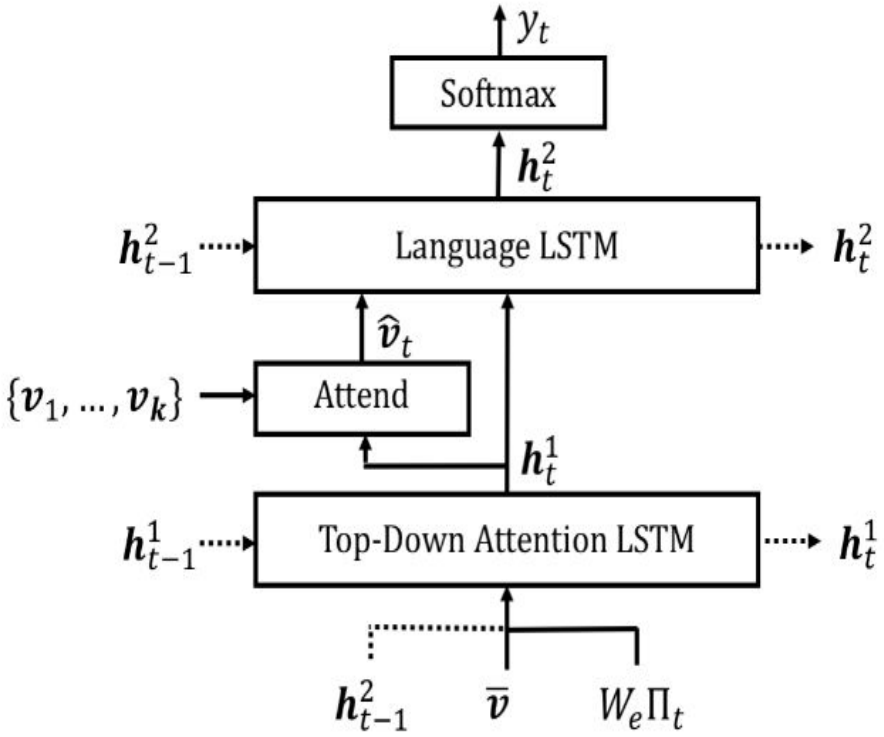


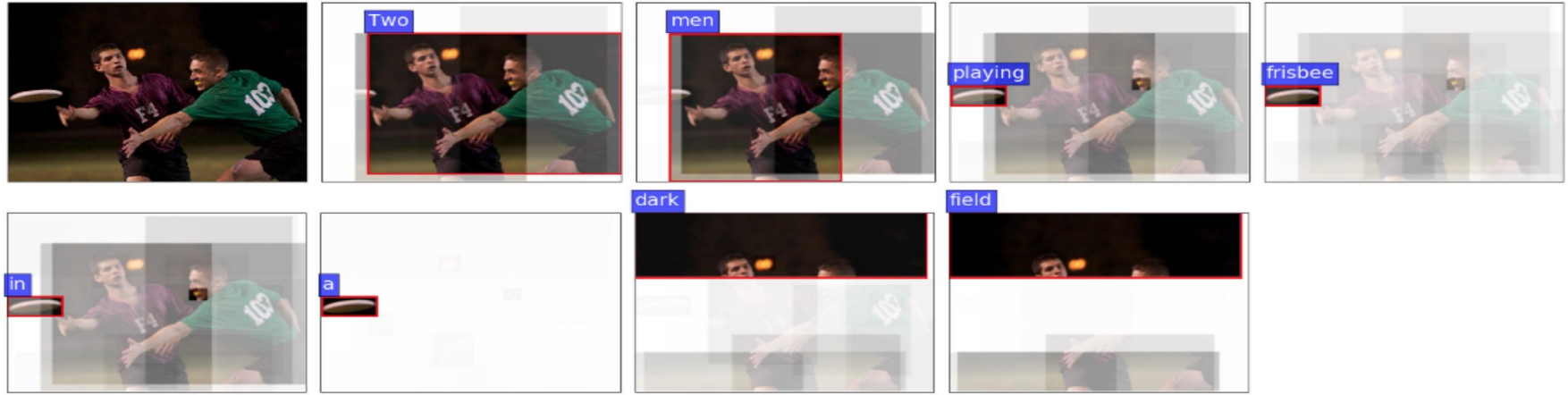
Foto : Leo Gomes

Foto : Leo Gomes


woman long hair all
white cabinet
kitchen
woman
red headband
black glasses
white outlet
black glasses
stove
green bottle
silver pot
floral dress
white stove top
black burner
black burner
open oven
food

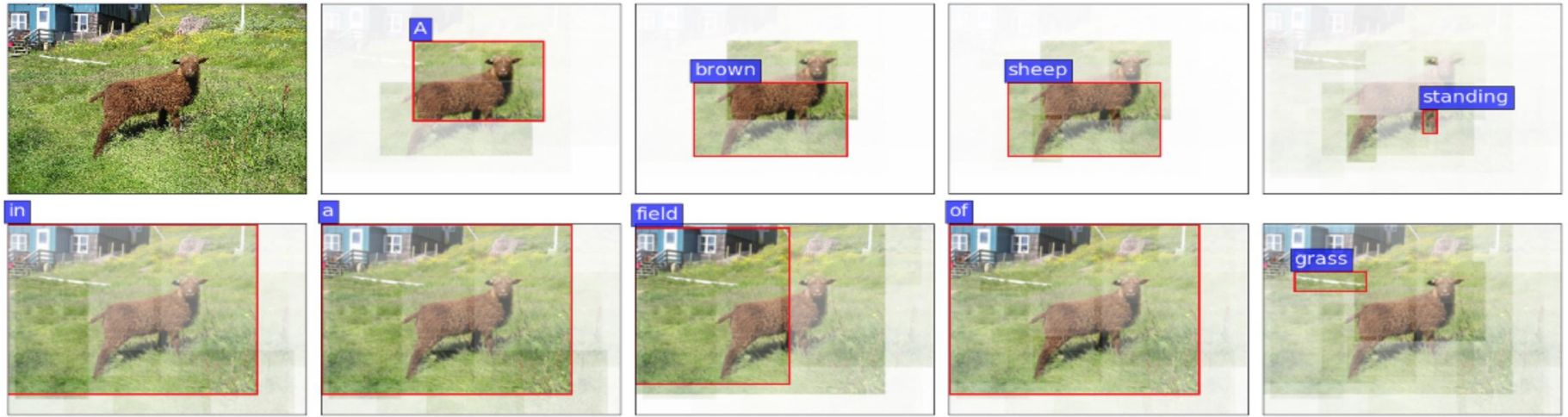
Top Down and Bottom Up Attention Model





Two men playing frisbee in a dark field.

A brown sheep standing in a field of grass.



Details

Dataset: Flickr8k (faster and smaller than the other benchmark ones)

Framework: PyTorch

Trained with Stochastic Gradient Descent with adaptive learning rate (Adams)

Batch Size : 10

Learning Rate: 5 x 10 ^-4

Epochs: 25

Results

(STANDARD/OURS)	BLEU 1	BLEU 2	BLEU 3	BLEU 4	CIDEr	ROUGE-L	SPICE	METEOR
Show and Tell	(0.63/0.61)	(0.41/0.433)	(0.27/0.301)	(/0.209)	(/0.491)	(/0.462)	(/0.13586)	(/0.199)
Show,Attend and Tell	(0.67/0.652)	(0.448/0.473)	(0.299/0.332)	(0.195/0.228)	(/0.600)	(/0.483)	(0.189/0.157)	(/0.216)
Top Down and Bottom Up Attention	(0.772/0.657)	(/0.477)	(/0.334)	(0.36/0.23)	(113.5/0.59)	(0.56/0.48)	(0.20/0.157)	(0.27/0.21)

Ground Truth:
tourists are standing a mountain viewpoint beneath a clear blue sky

Show and Tell:
a group of people are sitting on a bench in front of the ocean

Show, Attend, and Tell:
a group of people are sitting on a bench

Top Down and Bottom Up Attention:
a group of people are sitting on a rocky hill

Ground Truth:
a blonde boy in a white and orange t-shirt is playing on a swing

Show and Tell:
a young boy wearing a blue shirt is swinging on a swing

Show, Attend, and Tell:
a young girl in a white shirt and blue shirt is swinging on a swing

Top Down and Bottom Up Attention:
a young girl in a blue shirt is swinging on a swing

Ground Truth:
A football player in red and white is holding both hands up

Show and Tell:
a football player in a red jersey is running

Show, Attend, and Tell:
a football player in red is being tackled

Top Down and Bottom Up Attention:
a group of football players in red and white uniforms

Ground Truth:
A woman wading through a pool in front of a waterfall

Show and Tell:
a woman in a bikini is standing on a rock overlooking the ocean

Show, Attend, and Tell:
a girl in a swimsuit is standing in the water

Top Down and Bottom Up Attention:
a girl in a swimsuit is splashing in the water

Ground Truth:
A little girl is walking along a line of logs on a sandy beach

Show and Tell:
a woman and a dog are standing on a beach

Show, Attend, and Tell:
a young boy is standing on a beach

Top Down and Bottom Up Attention:
a woman and a woman are walking on a bridge

Ground Truth:
A person riding a snowboard jumps high over the snowy hill

Show and Tell:
a snowboarder is jumping over a hill

Show, Attend, and Tell:
a snowboarder is jumping over a snowy hill

Top Down and Bottom Up Attention:
a snowboarder is jumping over a snowy hill

Ground Truth:
A group of friends play in a lake

Show and Tell:
a group of people are in a canoe on a lake

Show, Attend, and Tell:
a group of people are standing on a beach

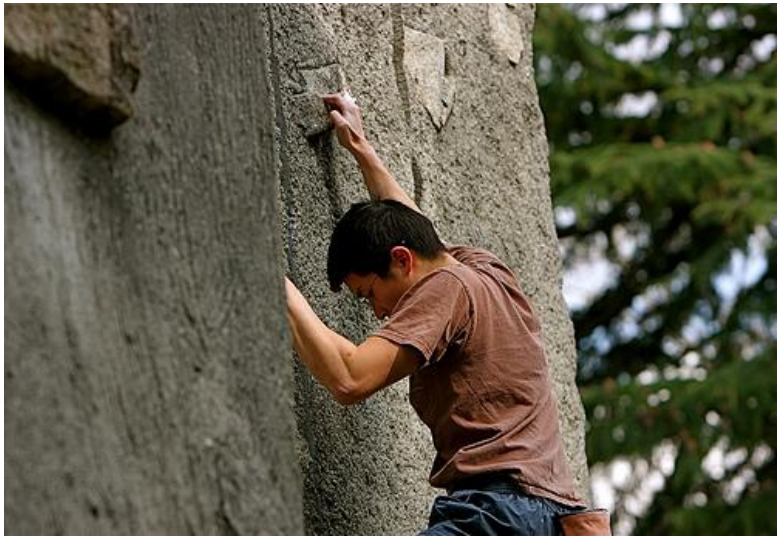
Top Down and Bottom Up Attention:
a group of people are walking along a lake

Ground Truth:
A dark-haired man wearing a brown shirt is free-climbing a grey stone wall

Show and Tell:
a girl in a pink shirt is climbing a rock wall

Show, Attend, and Tell:
a man in a white shirt and jeans is climbing a rock wall

Top Down and Bottom Up Attention:
a man climbing a rock



THANK YOU