MSCOCO

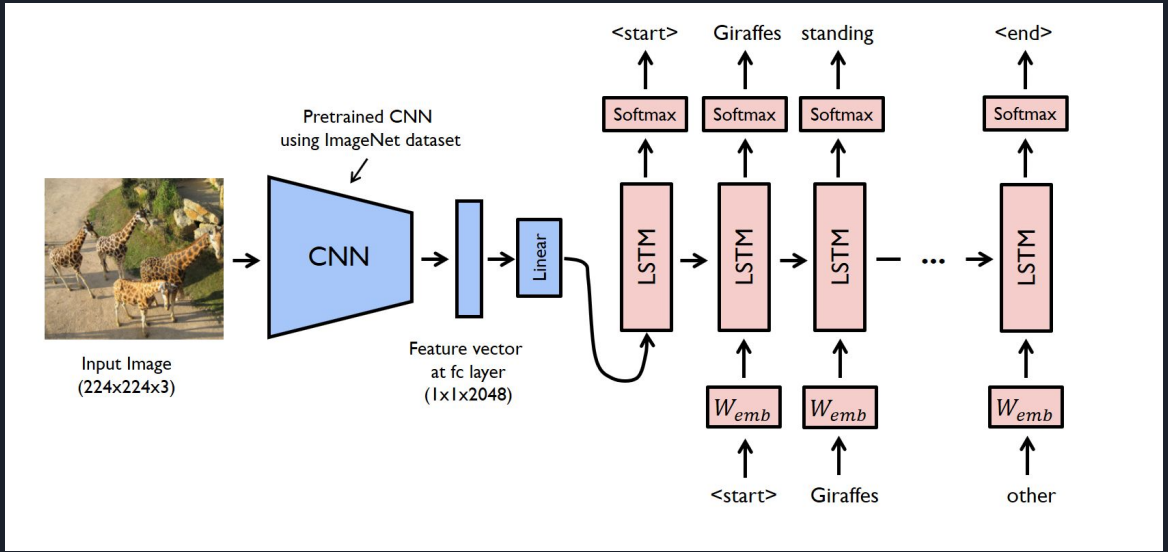
Team 8 Fortune-teller

Ping Zhu, Yang Hu, Zhichao Yang

Model — Overview

Encoder (resnet 152)

Decoder (LSTM)



Model — Knowing when to look

Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning

Jiasen Lu[2*†], Caiming Xiong[1†], Devi Parikh[3], Richard Socher[1]
[1]Salesforce Research, [2]Virginia Tech, [3]Georgia Institute of Technology
jiasenlu@vt.edu, parikh@gatech.edu, {cxiong, rsocher}@salesforce.com

Encoder

Use Resnet 152 to extract the image features V and V_g where v_g is used to form xt.

$$\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}, \boldsymbol{a}_i \in \mathcal{R}^{2048}$$

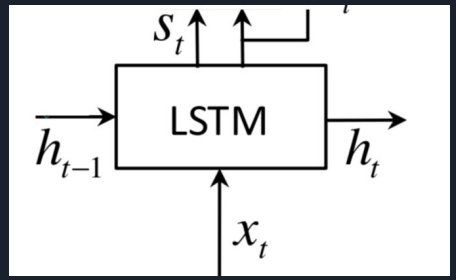
$$\boldsymbol{a}^g = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{a}_i$$

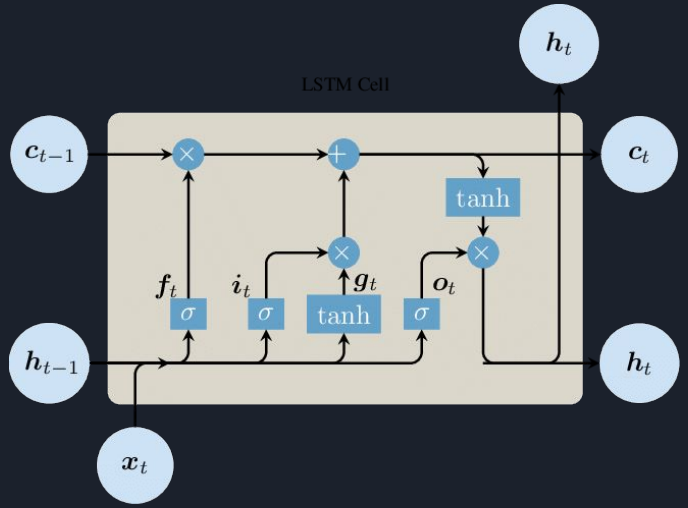
$$\begin{aligned} \boldsymbol{v}_i &= \text{ReLU}(\boldsymbol{W}_a \boldsymbol{a}_i) \\ \boldsymbol{v}^g &= \text{ReLU}(\boldsymbol{W}_b \boldsymbol{a}^g) \end{aligned}$$

Knowing when to look — Decoder

Modified LSTM cell to extract sentinel

$$
\begin{aligned}
\boldsymbol{g}_t &= \sigma\left(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1}\right) \\
\boldsymbol{s}_t &= \boldsymbol{g}_t \odot \tanh\left(\boldsymbol{m}_t\right)
\end{aligned}
$$





Knowing when to look — Decoder

Attention

$$
\begin{aligned}
\boldsymbol{z}_t &= \boldsymbol{w}_h^T \tanh(\boldsymbol{W}_v \boldsymbol{V} + (\boldsymbol{W}_g \boldsymbol{h}_t)\mathbb{1}^T) \\
\boldsymbol{\alpha}_t &= \text{softmax}(\boldsymbol{z}_t)
\end{aligned}
$$

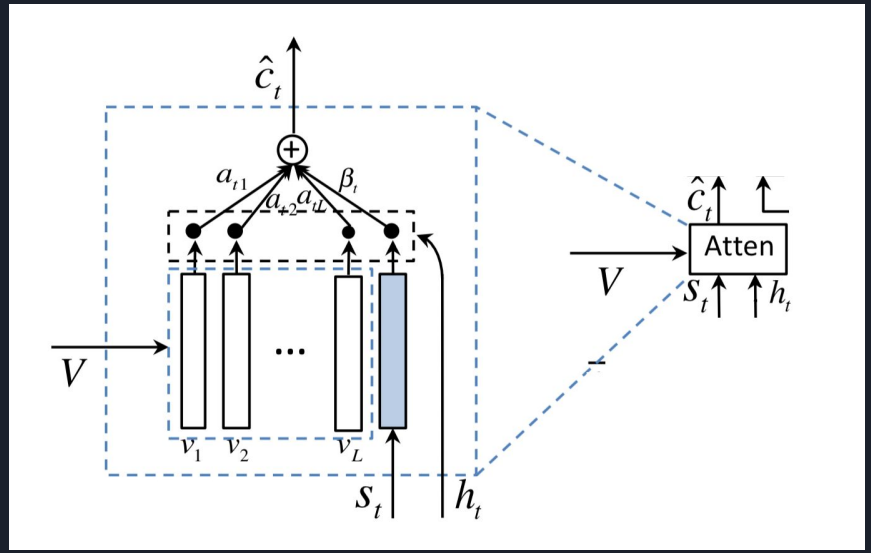
$$
\hat{\boldsymbol{\alpha}}_t = \text{softmax}([\boldsymbol{z}_t; \boldsymbol{w}_h^T \tanh(\boldsymbol{W}_s \boldsymbol{s}_t + (\boldsymbol{W}_g \boldsymbol{h}_t))])
$$

$$
\boldsymbol{c}_t = \sum_{i=1}^{k} \boldsymbol{\alpha}_{ti} \boldsymbol{v}_{ti}
$$

$$
\hat{\boldsymbol{\alpha}}_t \in \mathcal{R}^{k+1}
$$

$$
\beta_t = \boldsymbol{\alpha}_t[k+1]
$$

$$
\hat{\boldsymbol{c}}_t = \beta_t \boldsymbol{s}_t + (1 - \beta_t)\boldsymbol{c}_t
$$

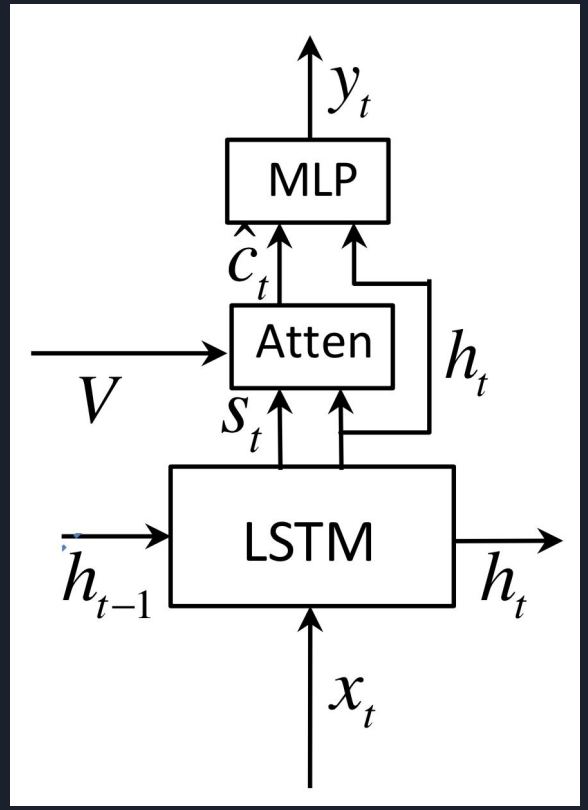


Knowing when to look — Decoder

Predict & Overview

$$\boldsymbol{p}_t = \text{softmax}\left(\boldsymbol{W}_p(\hat{\boldsymbol{c}}_t + \boldsymbol{h}_t)\right)$$

batch_size x # x len(vocab)



Performance

	CIDEr	Bleu 1	Bleu 2	Bleu 3	Bleu 4	ROUGE_L
C5	0.67169	0.66128	0.4769	0.3254	0.21609	0.48193
C40	0.7045	0.8432	0.7246	0.5873	0.4528	0.61058

Good Predictions

GROUND TRUTH CAPTION:
A man flying a kite on the beach.
An older man watches a kite fly from across a body of water.
An older gentleman flies a kite on the beach.
A man that is standing in the sand flying a kite.
An old man flying a kite high up in the air.

PREDICTED CAPTION:
a man is flying a kite on a beach .

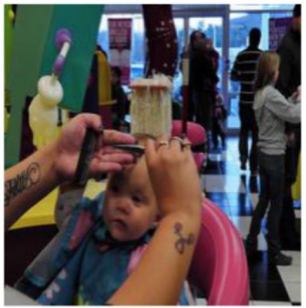
GROUND TRUTH CAPTION:
a man riding a big wave on his surfboard
A person surfing under a wave in the ocean.
A surfer catches and rides a large wave.
A person on a surfboard riding a wave.
A person a surf board riding the water waves.

PREDICTED CAPTION:
a man is riding a wave on the water .

Bad Predictions

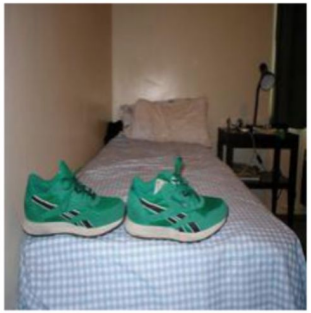
GROUND TRUTH CAPTION:
A little baby is getting a haircut in a pink chair.
A baby sitting in a chair getting a haircut at a salon.
A barber shop with a young child getting a haircut.
A baby sitting in a chair getting a haircut.
A baby getting a hair cut in the salon sitting in the chair

PREDICTED CAPTION:
a man holding a banana in a banana .



GROUND TRUTH CAPTION:
Two green shoes lined up on a bed.
A pair of sneakers lined up on a bed.
A pair of running shoes sit on the end of a bed.
Bright green sneakers on a bed with a gingham bedspread.
A pair of green sneakers on a single bed with a nightstand next to it

PREDICTED CAPTION:
a bed with a bed and a bed on it .



Thank you!