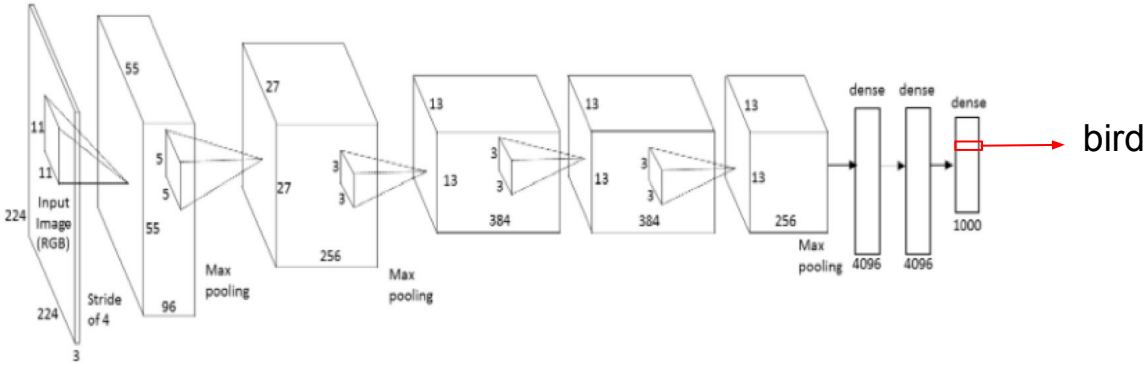
Day 4 Lecture 6

Attention Models

Attention Models: Motivation

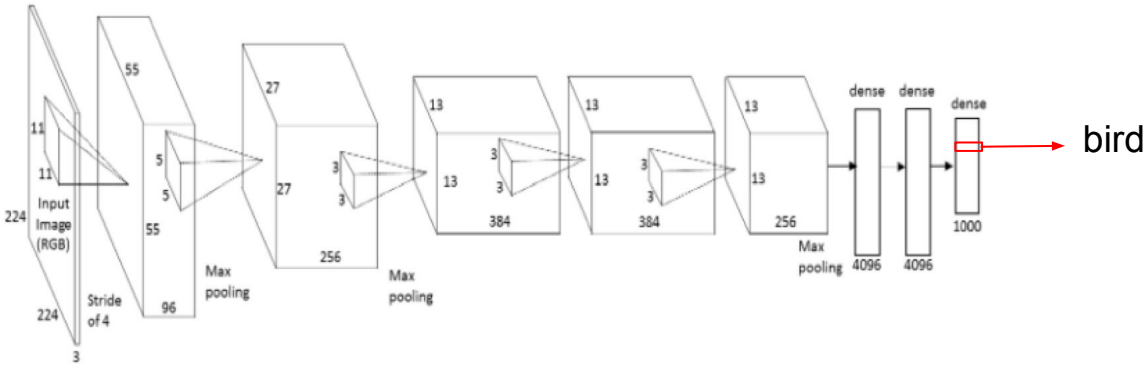
Image:
H x W x 3



The whole input volume is used to predict the output...

Attention Models: Motivation

Image:
H x W x 3



The whole input volume is used to predict the output...

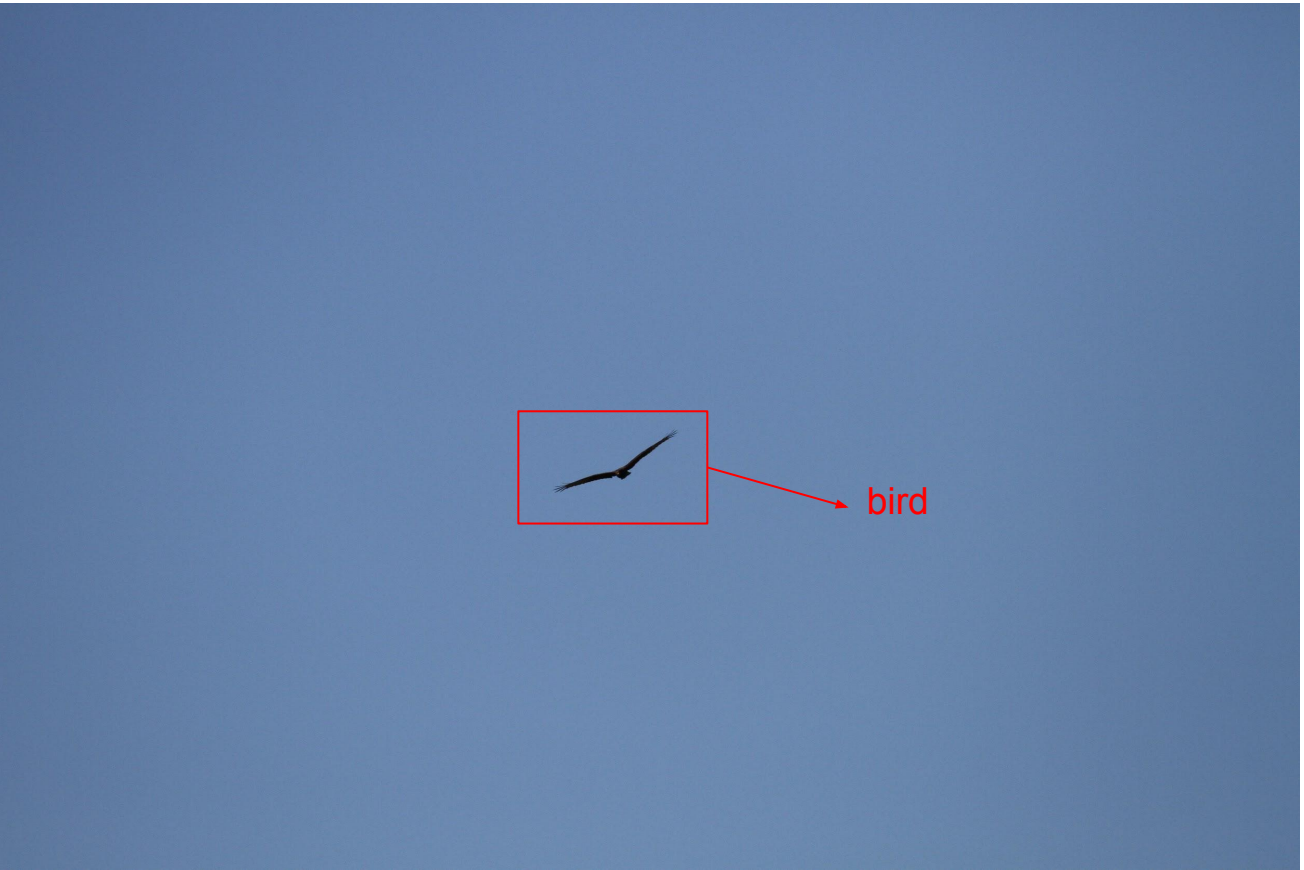
...despite the fact that not all pixels are equally important

Attention Models: Motivation

Attention models can relieve computational burden

Helpful when processing big images !

Attention Models: Motivation



Attention models can relieve computational burden

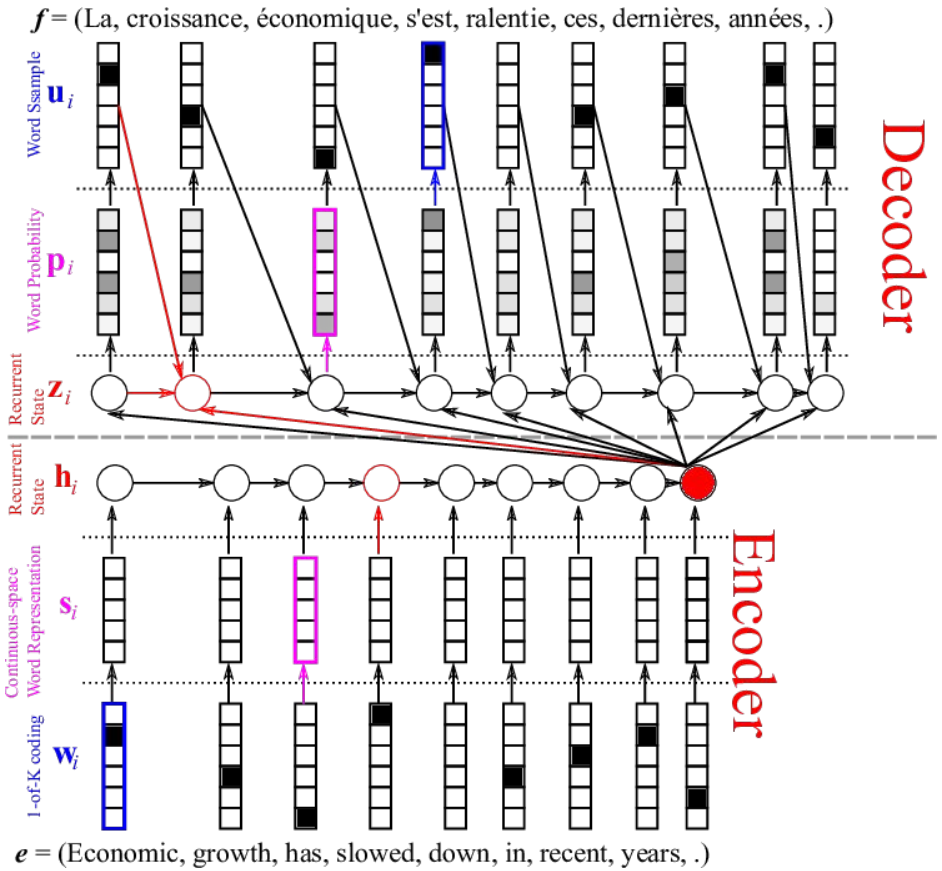
Helpful when processing big images !

Encoder & Decoder

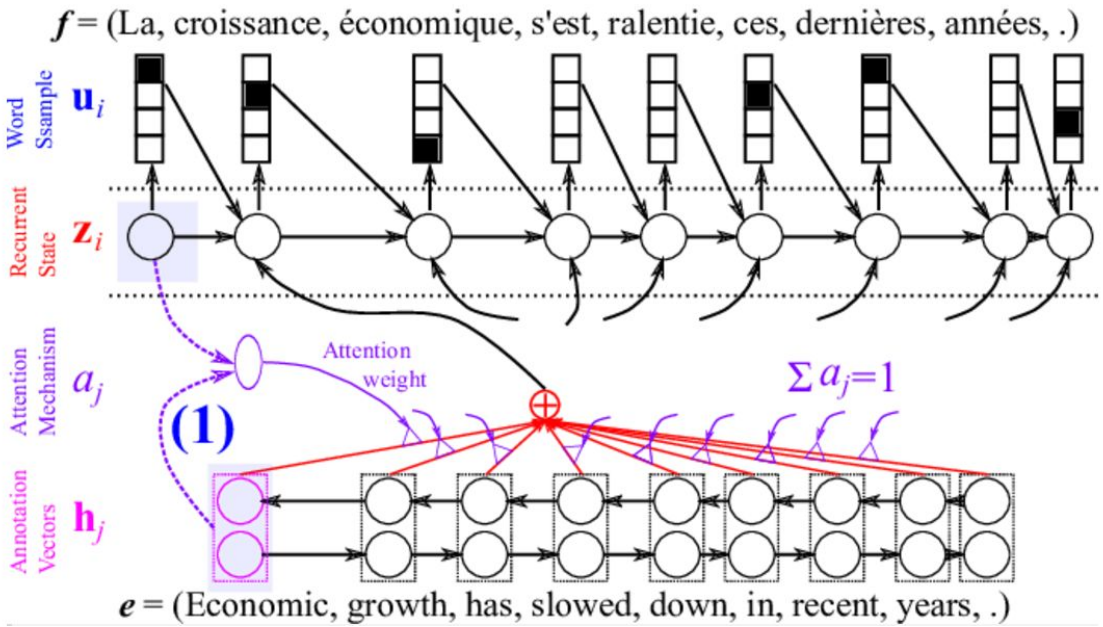
From previous lecture...

The whole input sentence
is used to produce the translation

Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”](#) (2015)



Attention Models



Bahdanau et al. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015
Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

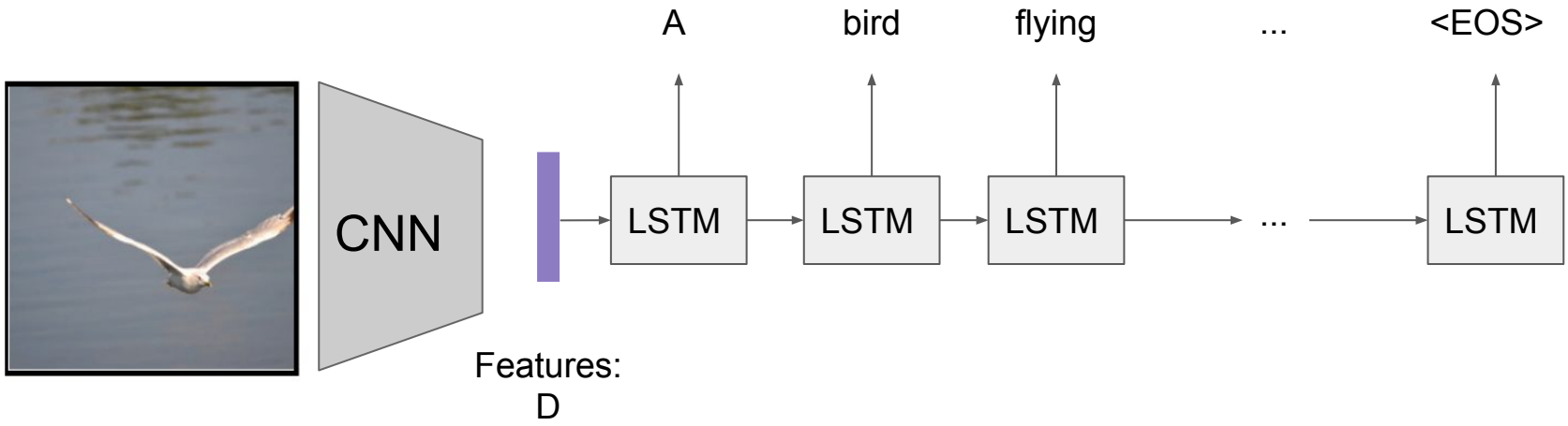
Attention Models

Idea: Focus in different parts of the input as you make/refine predictions in time

E.g.: Image Captioning

A bird flying over a body of water

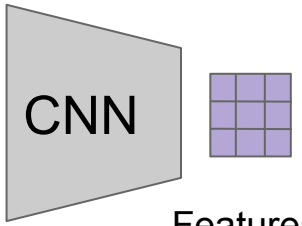
LSTM Decoder



The LSTM decoder “sees” the input only at the beginning !

Attention for Image Captioning

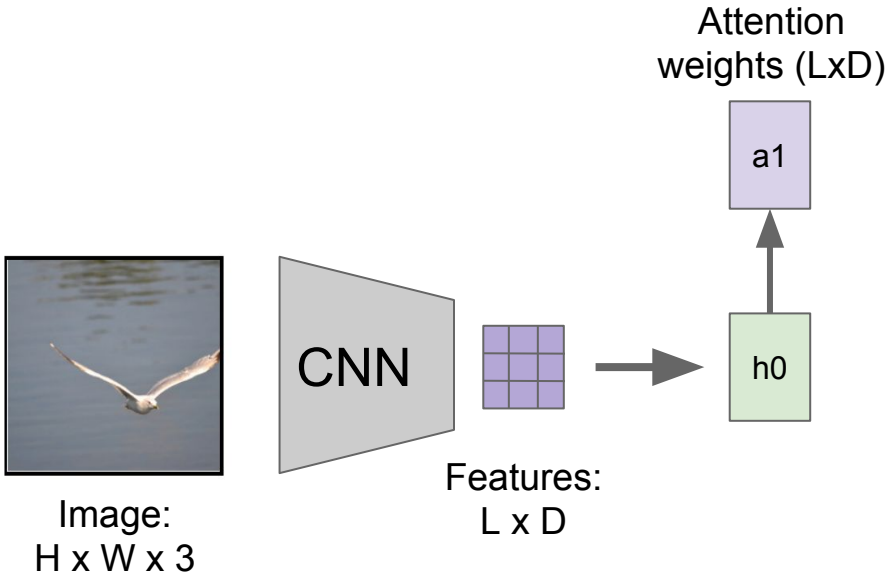
Image:
H x W x 3



Features:
L x D

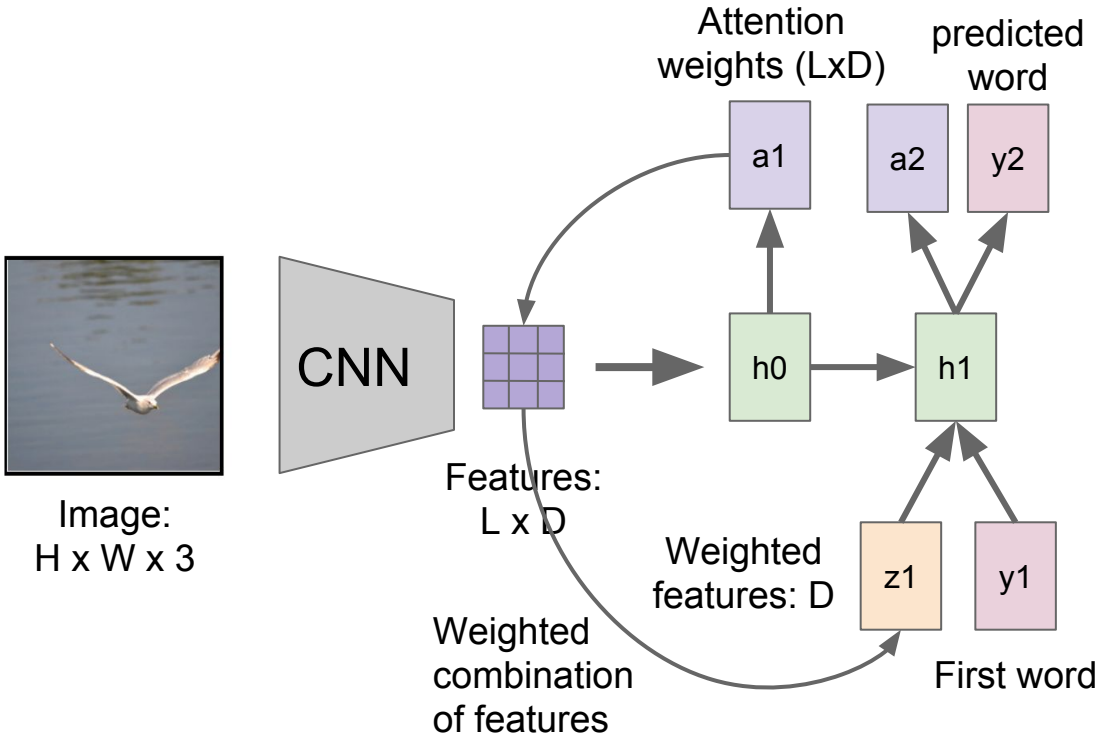
Slide Credit: CS231n

Attention for Image Captioning



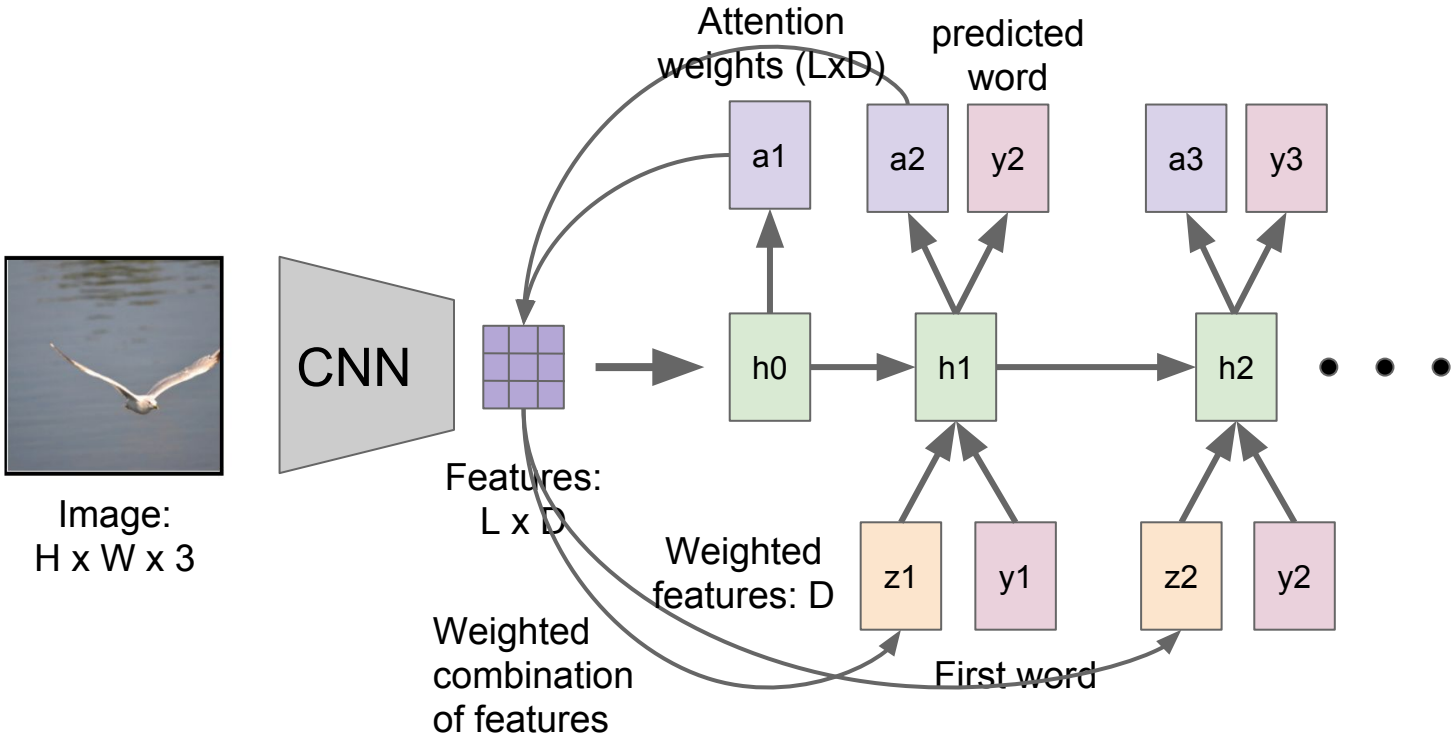
Slide Credit: CS231n

Attention for Image Captioning



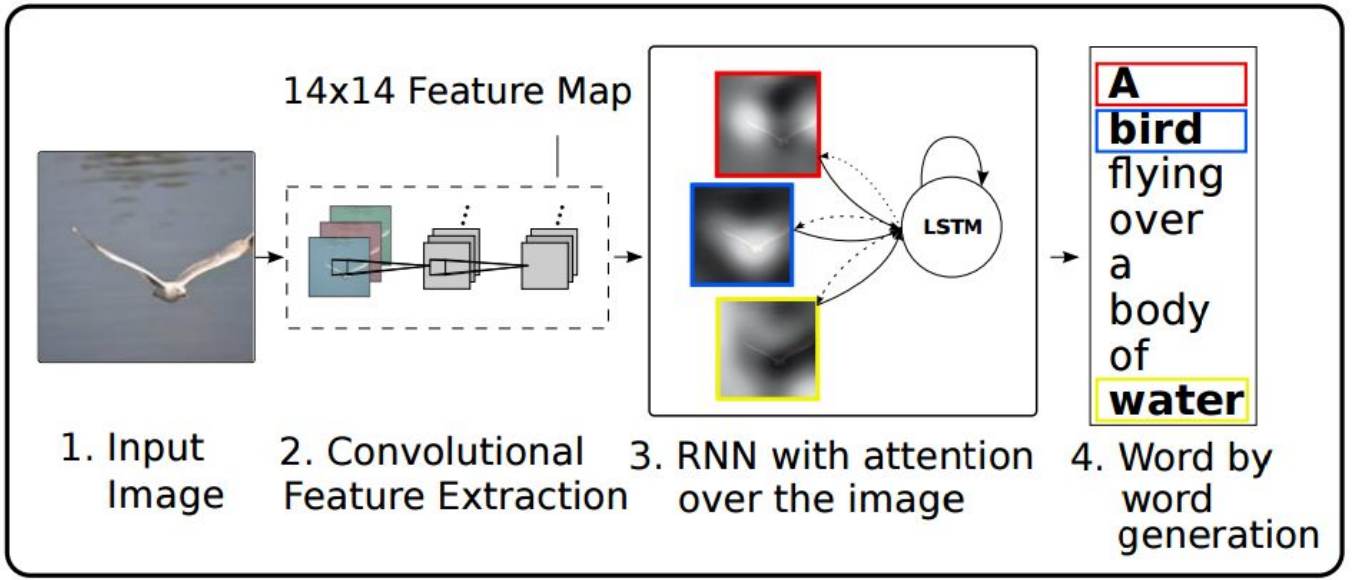
Slide Credit: CS231n

Attention for Image Captioning



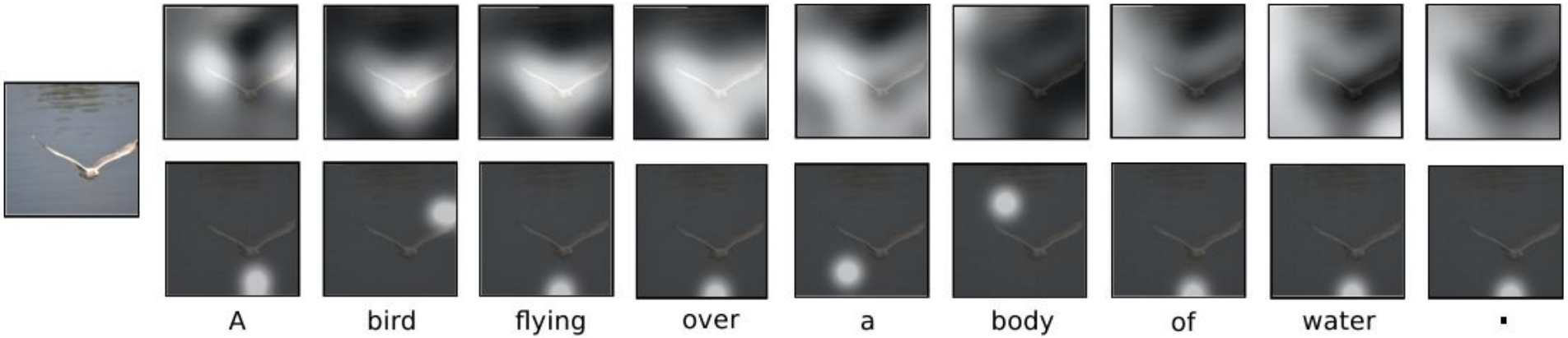
Slide Credit: CS231n

Attention for Image Captioning



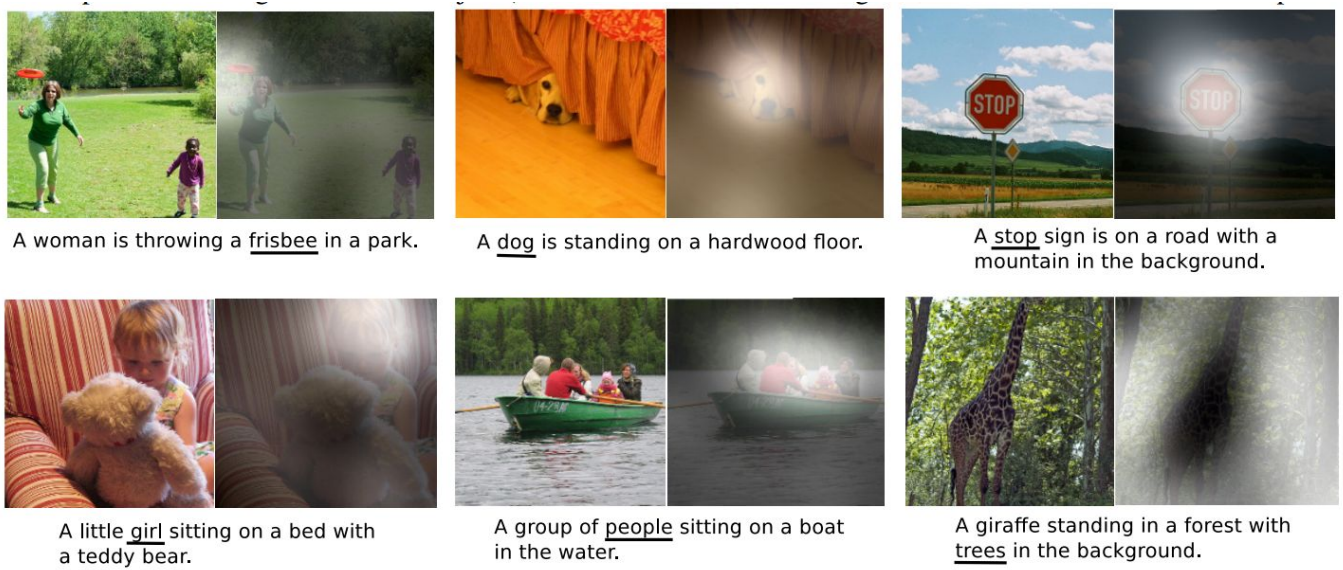
Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Attention for Image Captioning



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

Attention for Image Captioning



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

Soft Attention

Image:
H x W x 3

CNN

a	b
c	d

Grid of features
(Each D-dimensional)

Context vector z
(D-dimensional)

From
RNN:

p_a	p_b
p_c	p_d

Distribution over grid locations
$p_a + p_b + p_c + p_c = 1$

Soft attention:
Summarize ALL locations
$z = p_a a + p_b b + p_c c + p_d d$

Derivative dz/dp is nice!
Train with gradient descent

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

Slide Credit: CS231n

Soft Attention

Image:
H x W x 3

CNN

a	b
c	d

Grid of features
(Each D-dimensional)

Context vector z
(D-dimensional)

From
RNN:

p_a	p_b
p_c	p_d

Distribution over grid locations
$p_a + p_b + p_c + p_c = 1$

Soft attention:
Summarize ALL locations
$z = p_a a + p_b b + p_c c + p_d d$

Differentiable function
Train with gradient descent

- Still uses the whole input !
- Constrained to fix grid

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

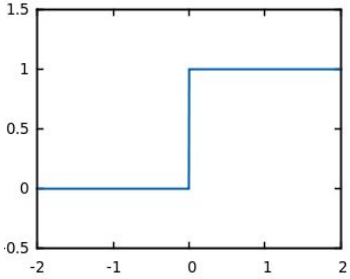
Slide Credit: CS231n

Hard attention

Input image:
H x W x 3

Cropped and
rescaled image:
X x Y x 3

Box Coordinates:
(xc, yc, w, h)



Gradient is 0 almost everywhere
Gradient is undefined at x = 0

Hard attention:
Sample a subset
of the input

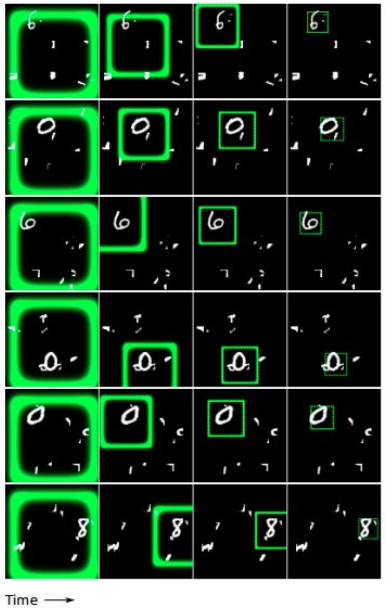
Not a differentiable function !

Can't train with backprop :(

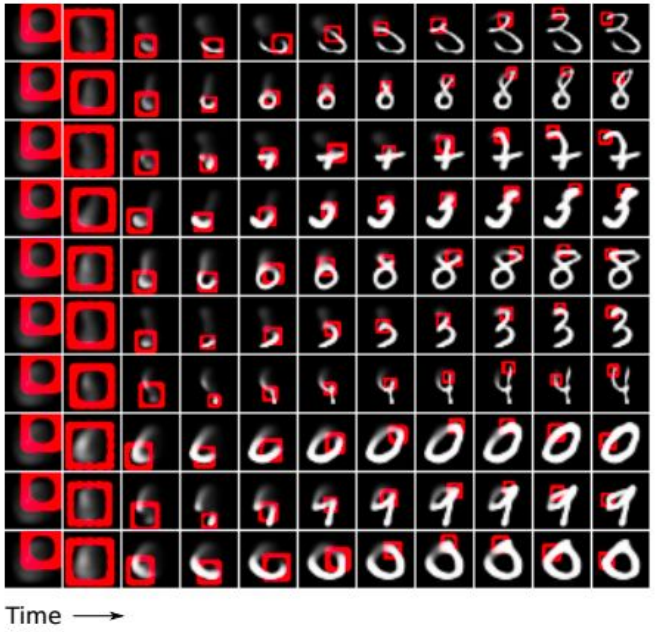
need reinforcement learning

Hard attention

Classify images by attending to arbitrary regions of the *input*

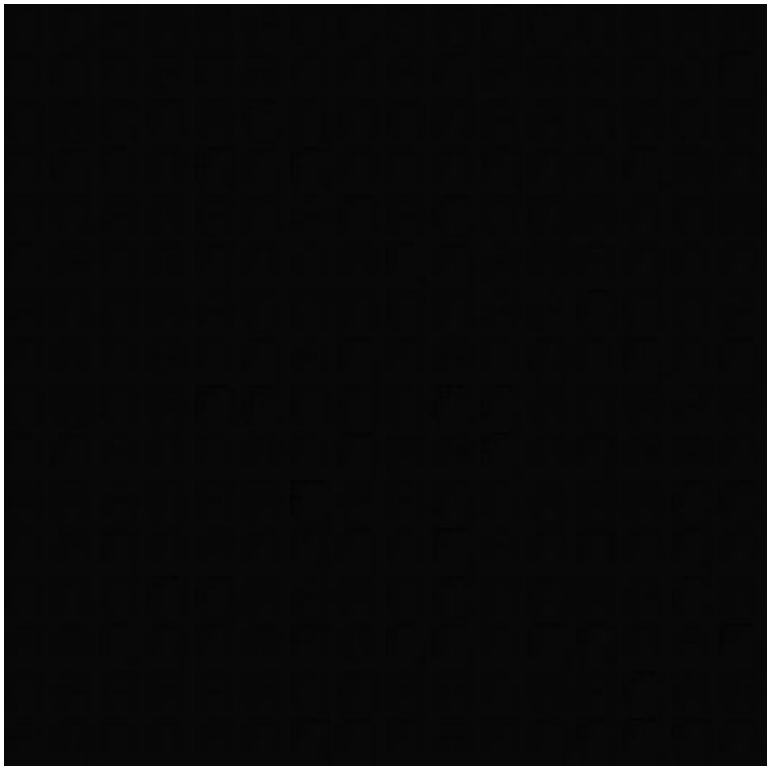


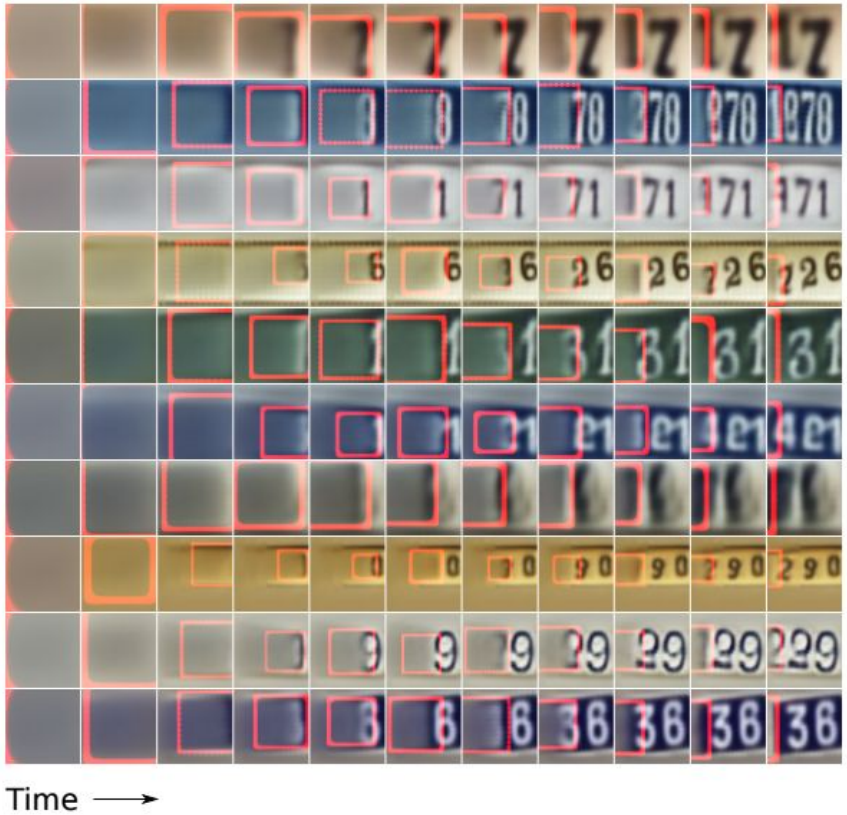
Generate images by attending to arbitrary regions of the *output*



Gregor et al. DRAW: A Recurrent Neural Network For Image Generation. ICML 2015

Hard attention

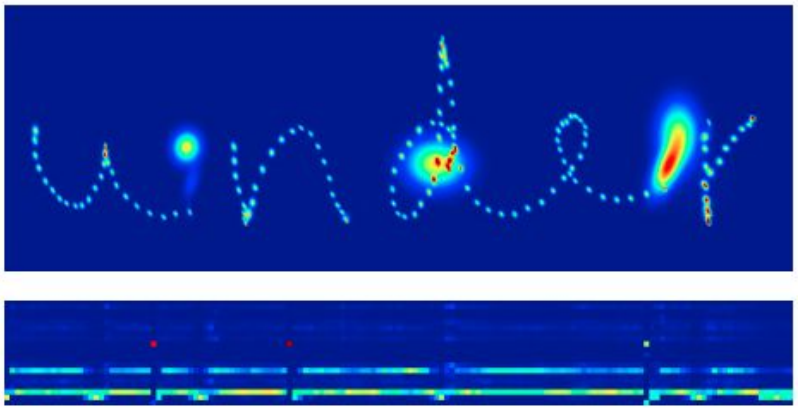


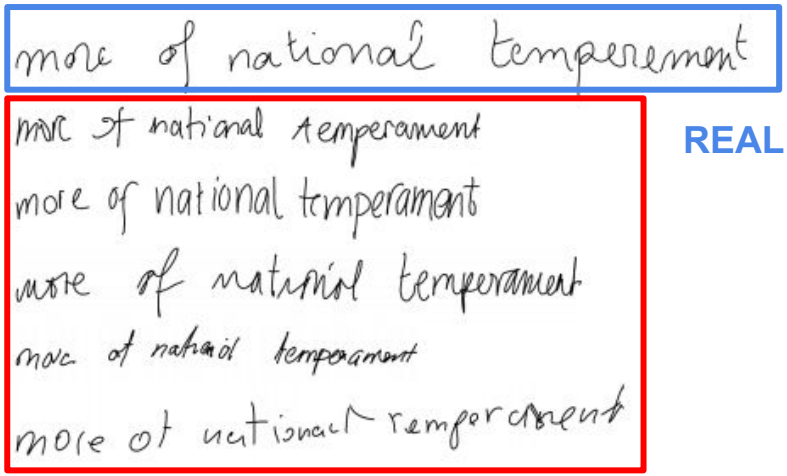


Gregor et al. DRAW: A Recurrent Neural Network For Image Generation. ICML 2015

Hard attention

Read text, generate handwriting using an RNN that attends at different arbitrary regions over time





Graves. Generating Sequences with Recurrent Neural Networks. arXiv 2013

Hard attention

Input image:
H x W x 3

Box Coordinates:
(xc, yc, w, h)

Cropped and
rescaled image:
X x Y x 3

CNN

bird

Not a differentiable function !

Can’t train with backprop :(

Spatial Transformer Networks

Input image:
H x W x 3

Box Coordinates:
(xc, yc, w, h)

Cropped and
rescaled image:
X x Y x 3

CNN

bird

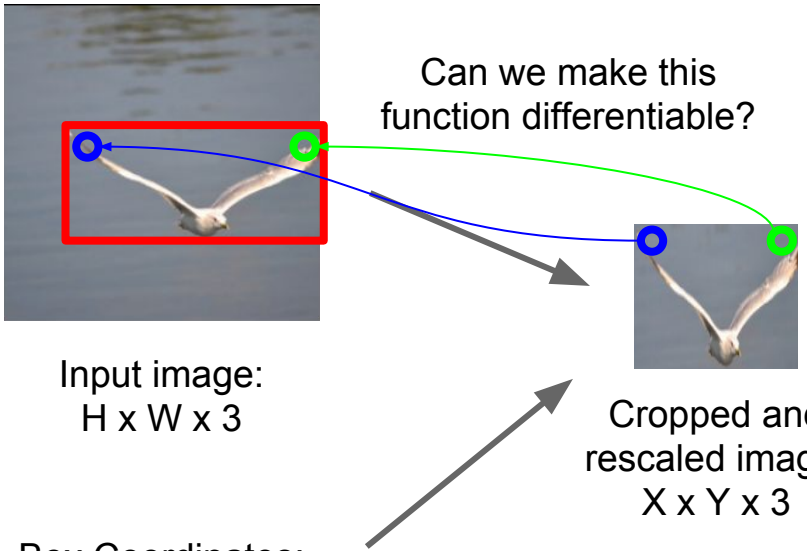
Not a differentiable function ! → Make it differentiable

Can't train with backprop :(

Train with backprop :)

Jaderberg et al. Spatial Transformer Networks. NIPS 2015

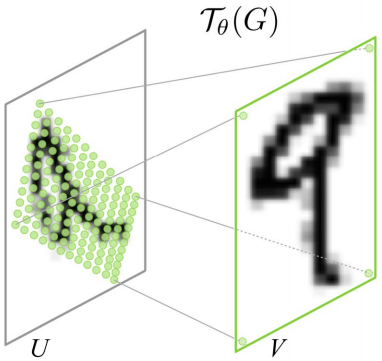
Spatial Transformer Networks



Idea: Function mapping *pixel coordinates* (xt, yt) of output to *pixel coordinates* (xs, ys) of input

Network attends to input by predicting θ

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$



Repeat for all pixels in *output* to get a **sampling grid**

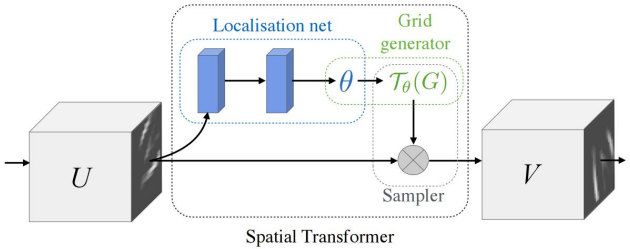
Then use **bilinear interpolation** to compute output

Jaderberg et al. Spatial Transformer Networks. NIPS 2015

Slide Credit: CS231n

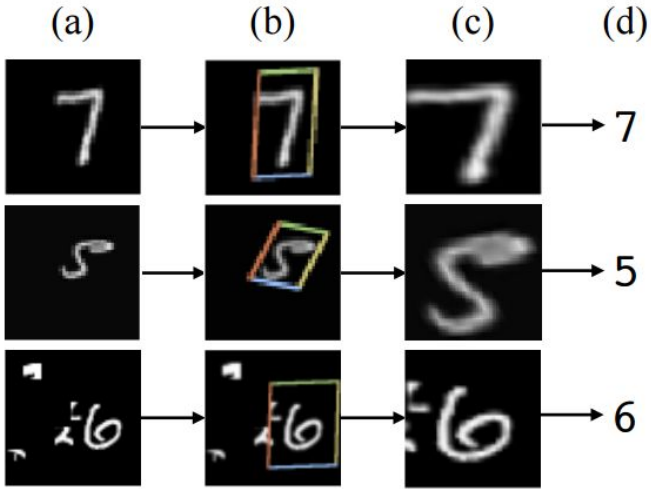
Spatial Transformer Networks

Differentiable module



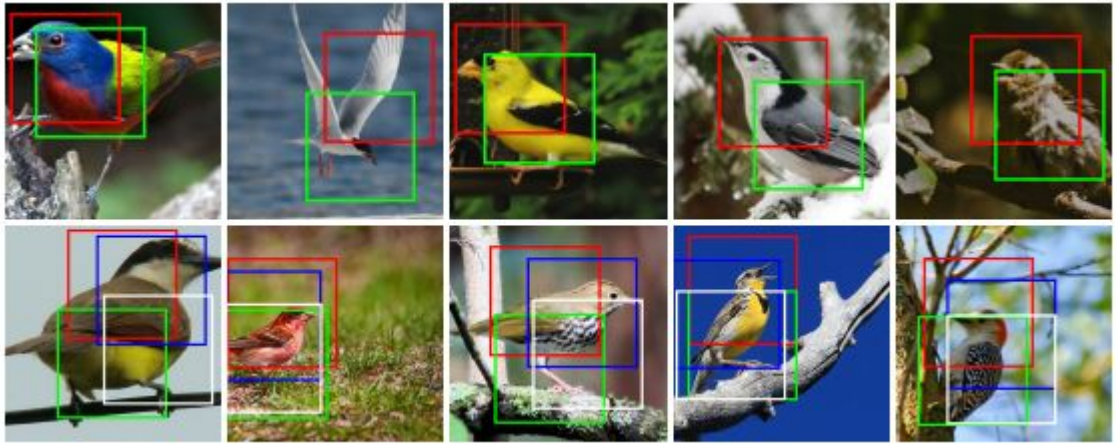
Easy to incorporate in any network, anywhere !

Insert spatial transformers into a classification network and it learns to attend and transform the input



Jaderberg et al. Spatial Transformer Networks. NIPS 2015

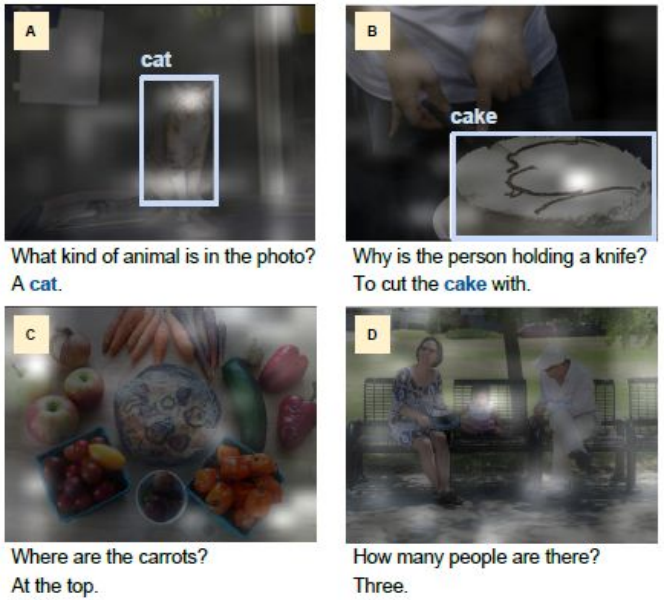
Spatial Transformer Networks

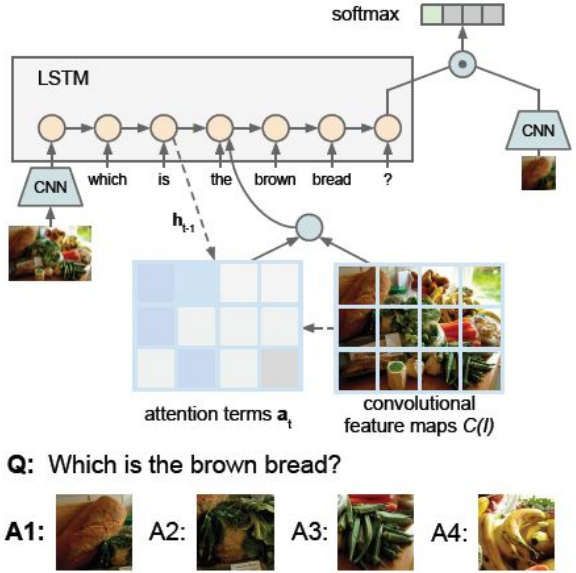


Fine-grained classification

Jaderberg et al. Spatial Transformer Networks. NIPS 2015

Visual Attention

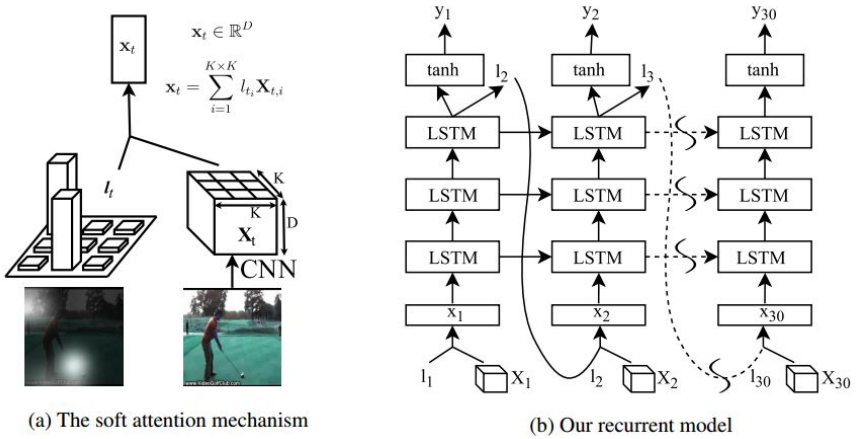




Visual Question Answering

Zhu et al. [Visual7w: Grounded Question Answering in Images.](#) arXiv 2016

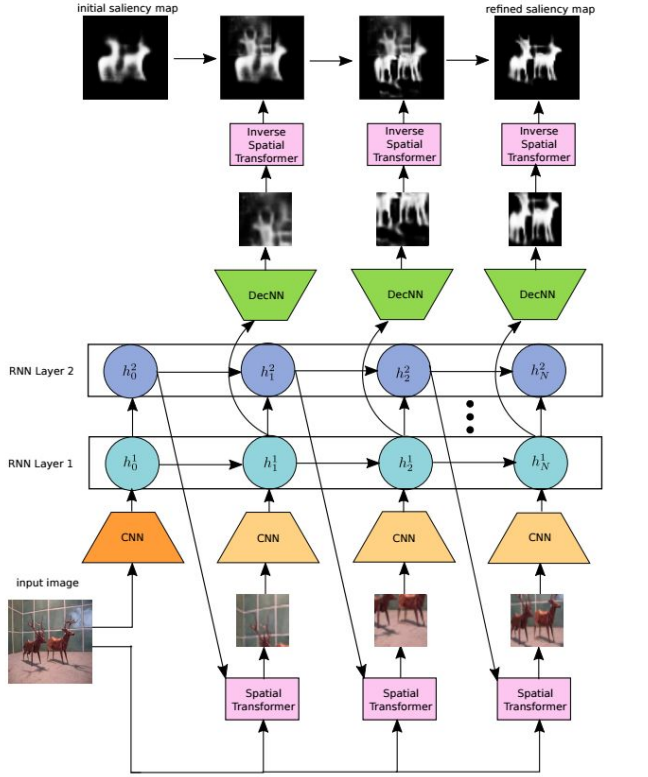
Visual Attention



(a) The soft attention mechanism

(b) Our recurrent model

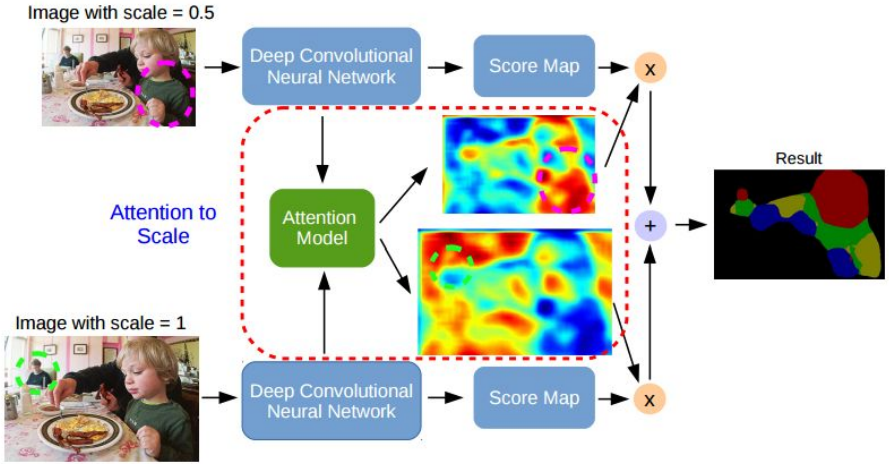
Action Recognition in Videos



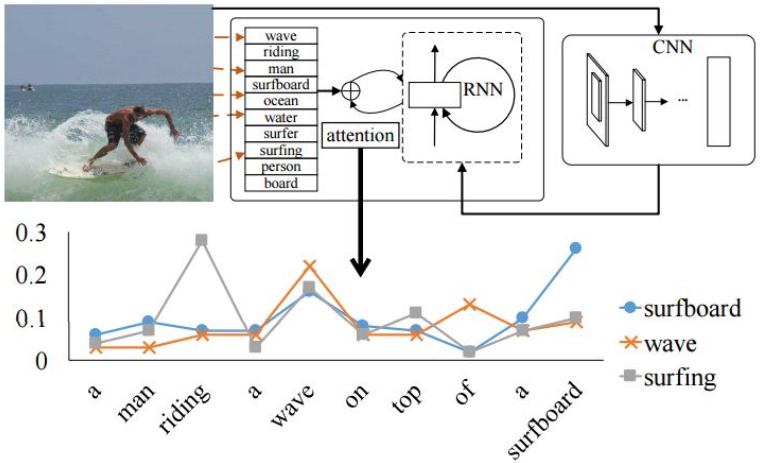
Salient Object Detection

Sharma et al. Action Recognition Using Visual Attention. arXiv 2016
Kuen et al. Recurrent Attentional Networks for Saliency Detection. CVPR 2016

Other examples



Attention to scale for semantic segmentation



Semantic attention For image captioning

Chen et al. Attention to Scale: Scale-aware Semantic Image Segmentation. CVPR 2016
You et al. Image Captioning with Semantic Attention. CVPR 2016

Resources

- CS231n Lecture @ Stanford [slides][video]
- More on Reinforcement Learning
- Soft vs Hard attention
- Handwriting generation demo
- Spatial Transformer Networks - Slides & Video by Victor Campos
- Attention implementations:
  - Seq2seq in Keras
  - DRAW & Spatial Transformers in Keras
  - DRAW in Lasagne
  - DRAW in Tensorflow