Insight
Centre for Data Analytics

DEEP
LEARNING
WORKSHOP

Dublin City University
27-28 April 2017

DCU

Day 2 Lecture 12

Attention Models

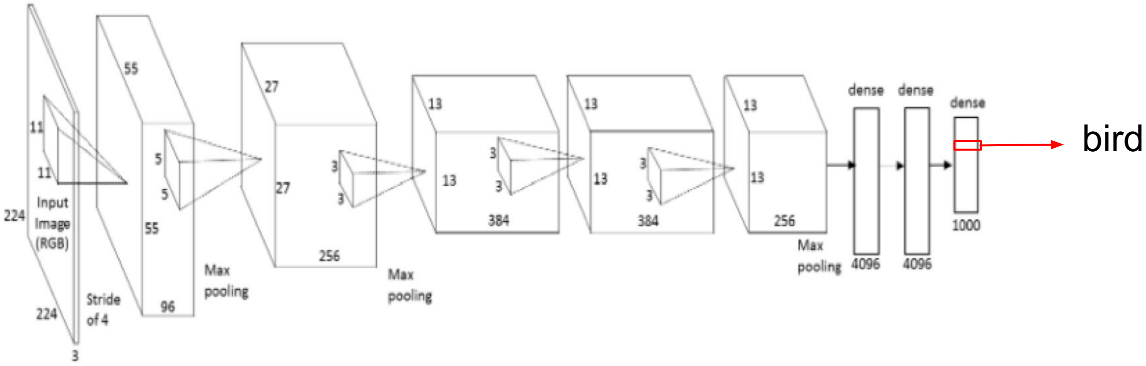
Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

UPC

Attention Models: Motivation

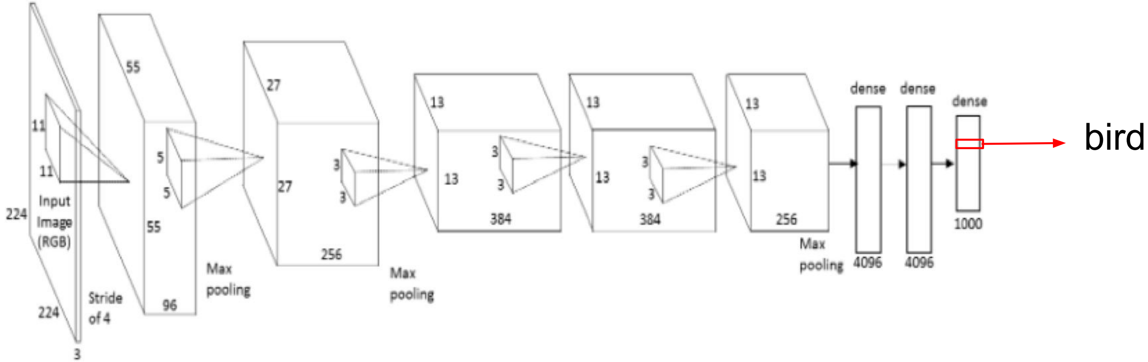
Image:
H x W x 3



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...

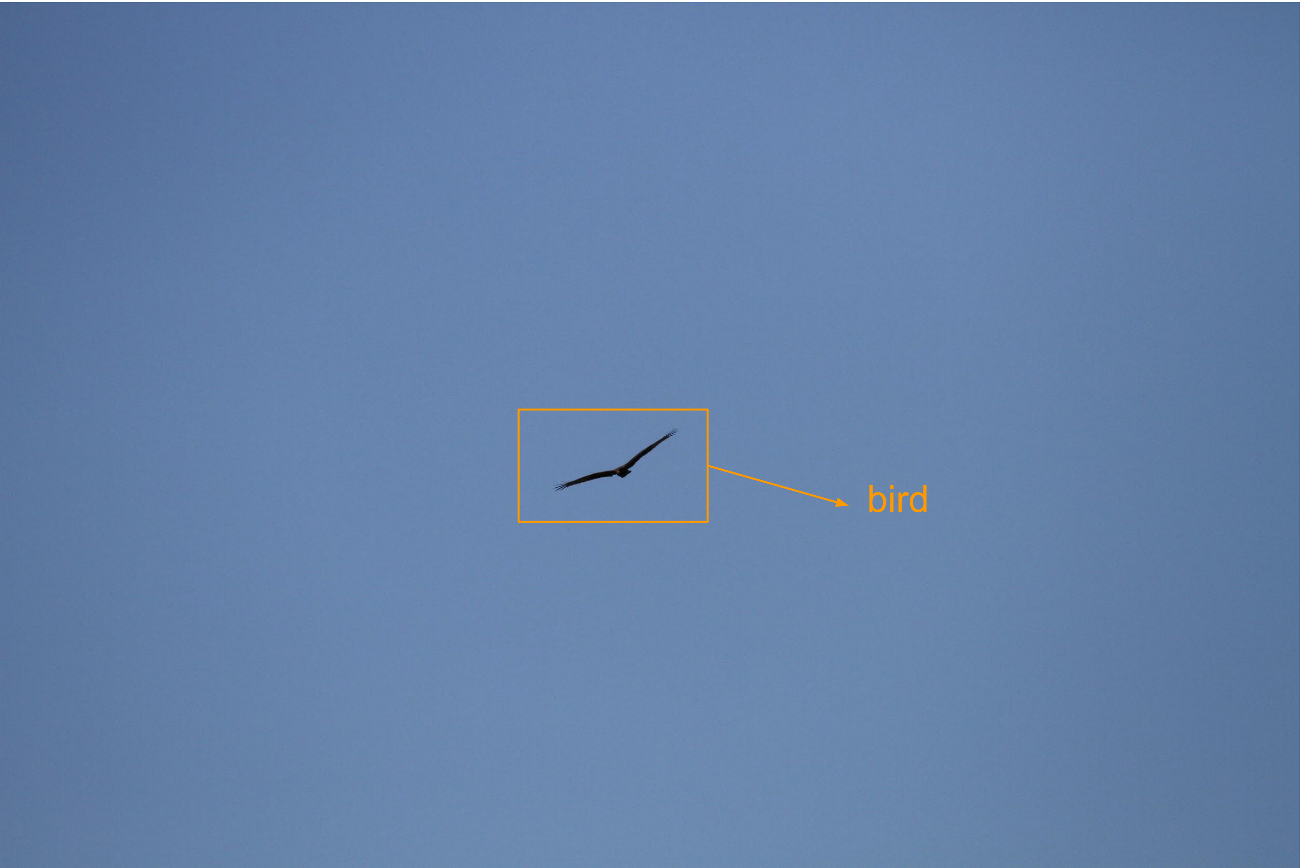
...despite the fact that not all pixels are equally important

Attention Models: Motivation

Attention models can relieve computational burden

Helpful when processing big images !

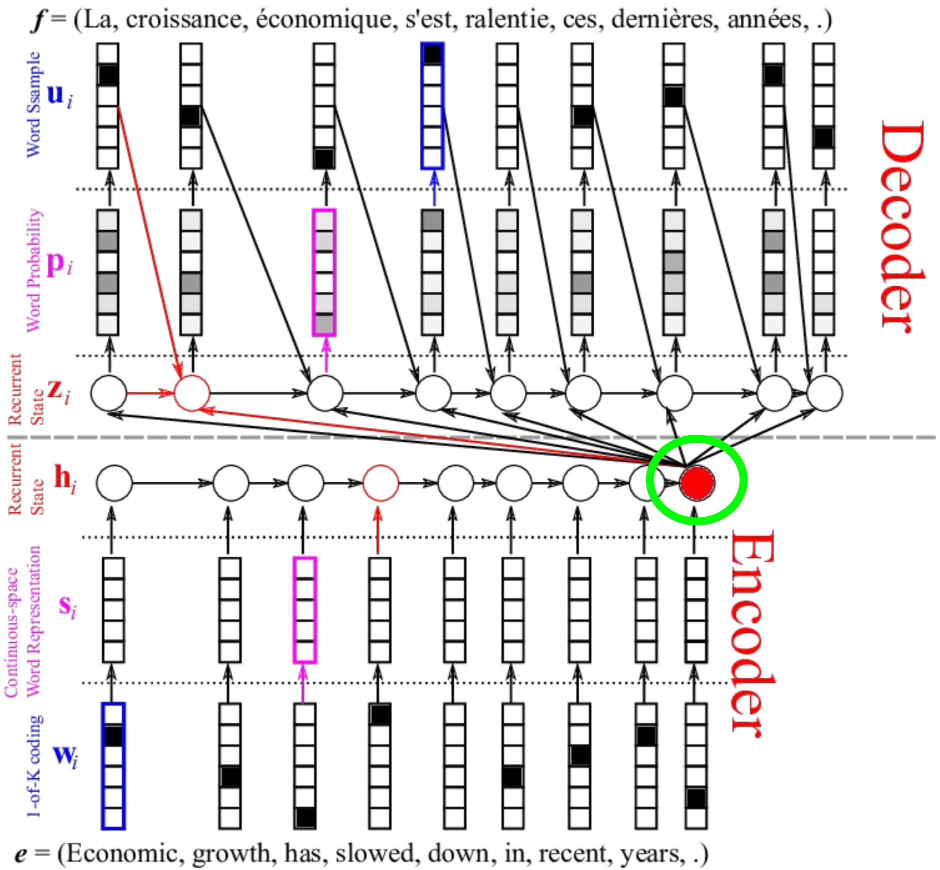
Attention Models: Motivation



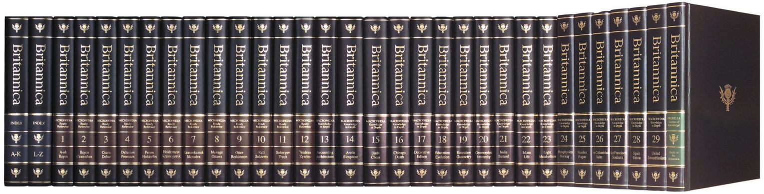
Attention models can relieve computational burden

Helpful when processing big images !

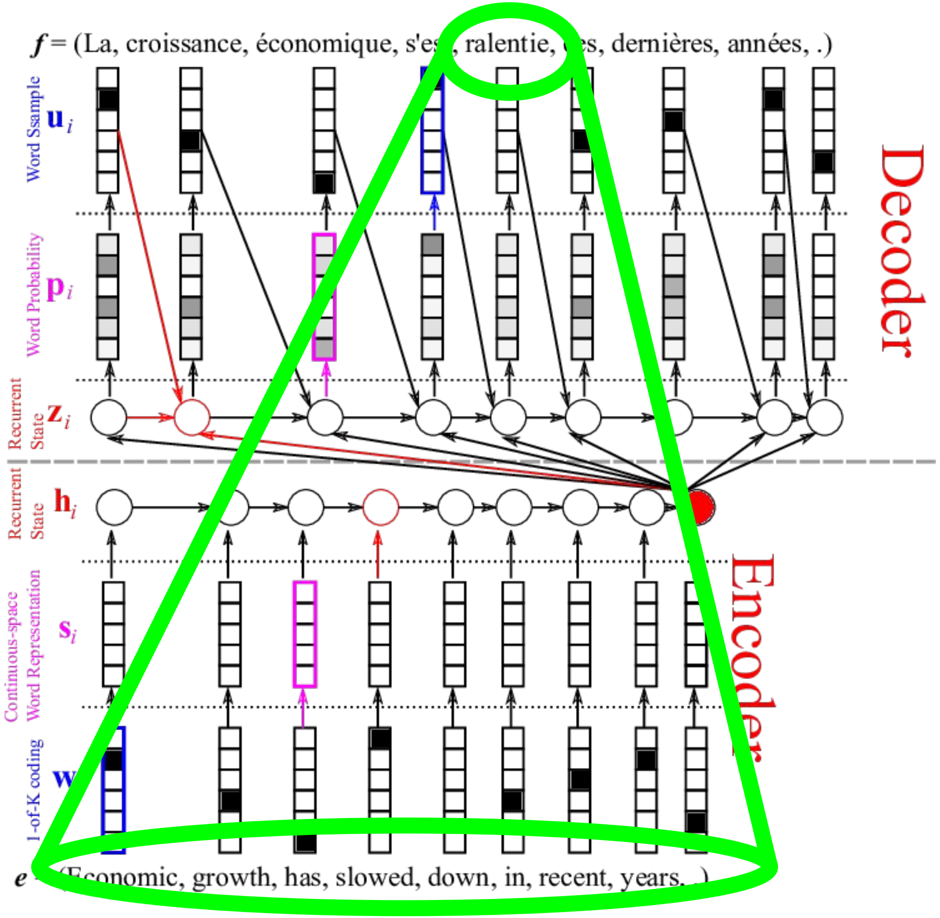
Encoder & Decoder for Machine Translation



<u>Limtation 1</u>: The whole information is encoded in a **<u>fixed-size vector</u>**, no matter the length of the input sentence.



Encoder & Decoder



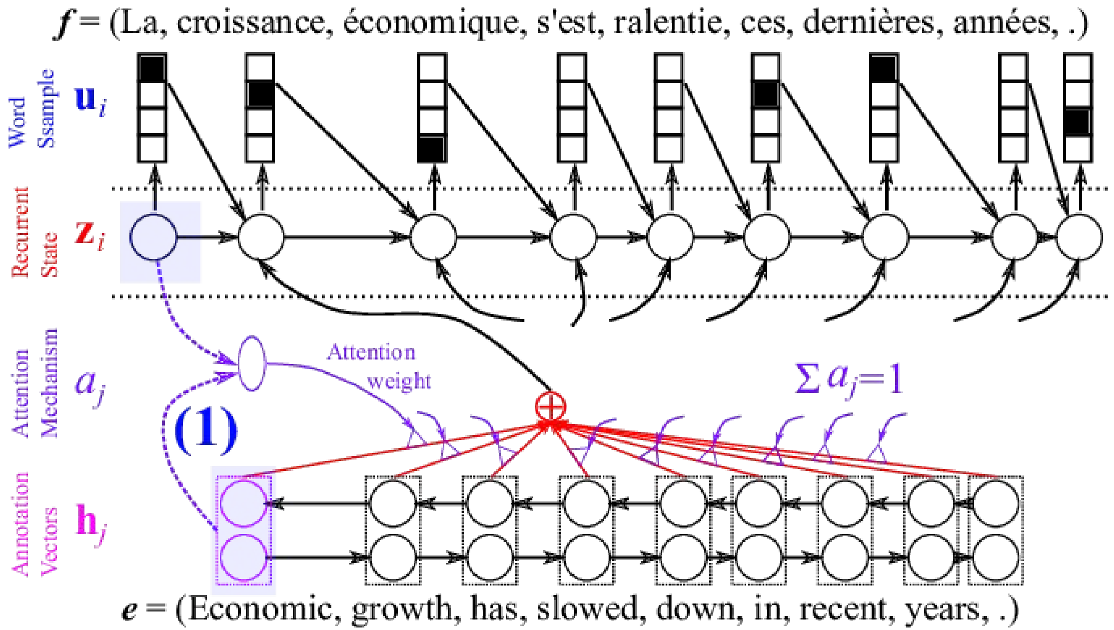
<u>Limitation 2</u>: All output predictions are based on the **final and static** recurrent state of the encoder (h_T)

No matter the output word being predicted at each time step, all input words are considered in an equal way.

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

Attention Mechanism

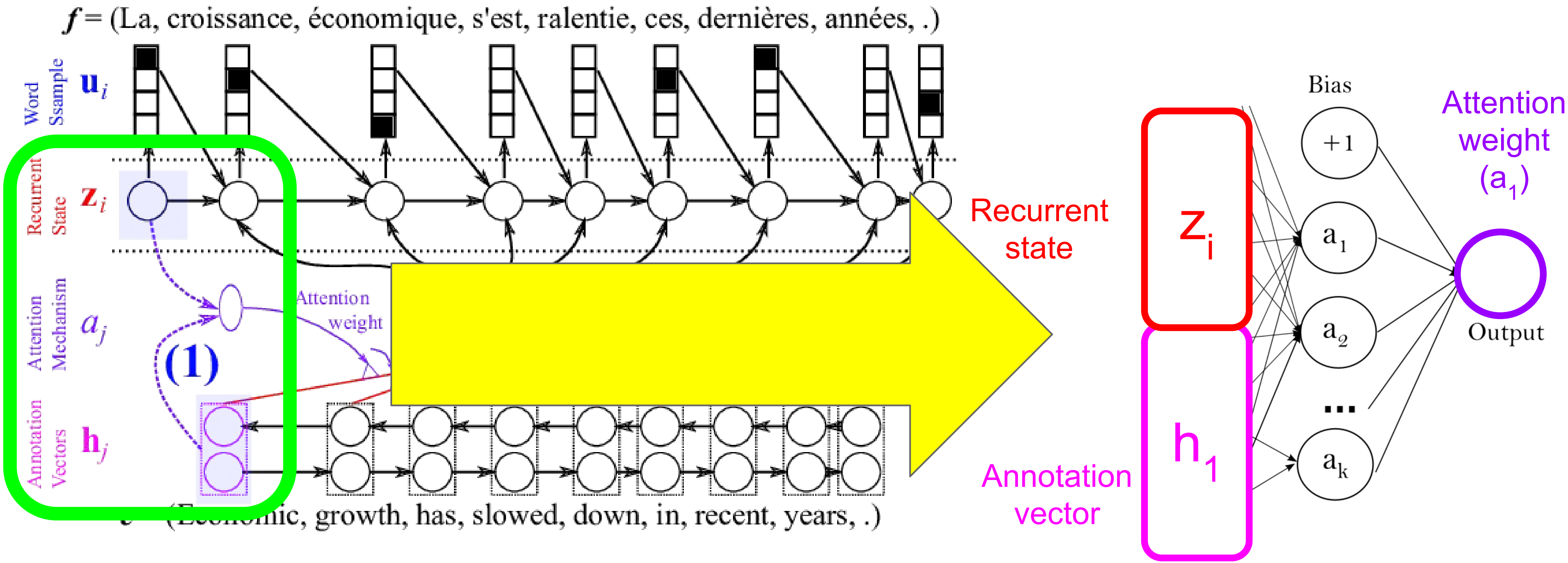
The vector to be fed to the RNN at each timestep is a weighted sum of all the annotation vectors.



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

Attention Mechanism

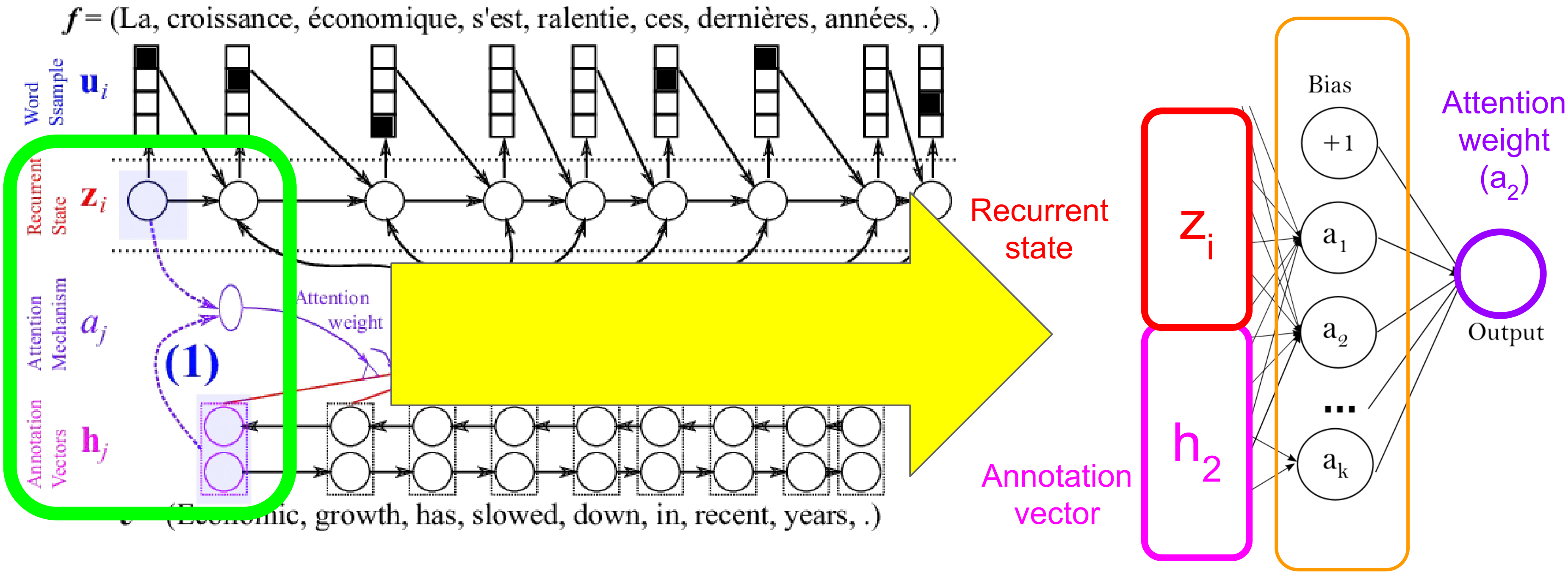
An attention weight (scalar) is predicted at each time-step for each annotation vector h_j with a simple fully connected neural network.



Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

Attention Mechanism

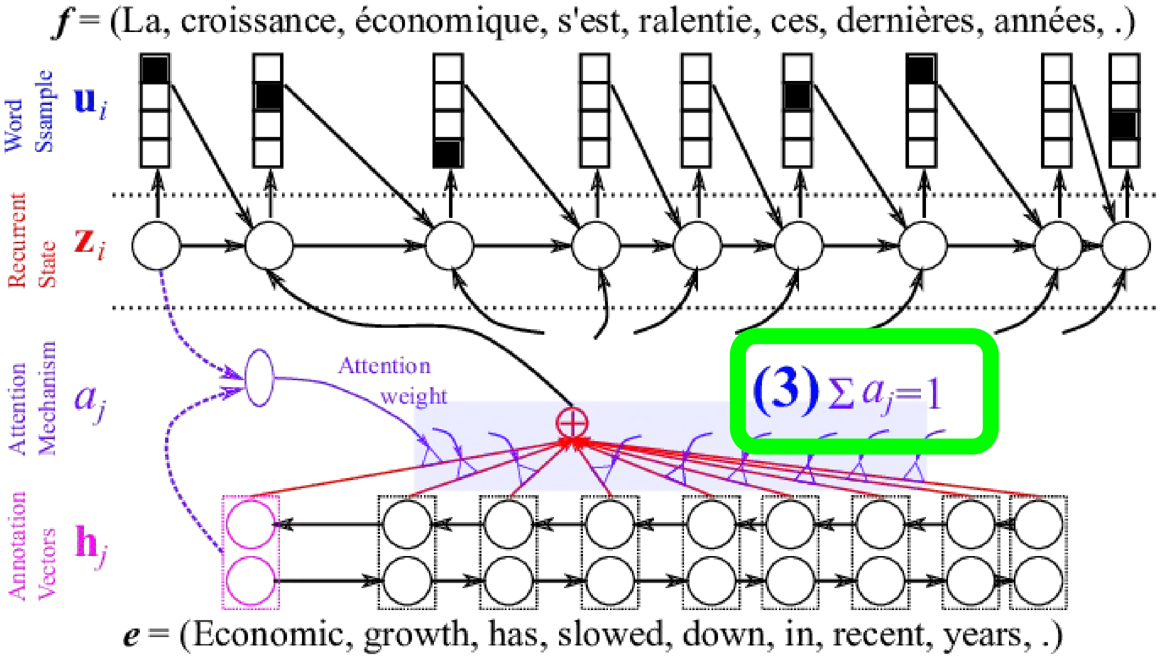
An attention weight (scalar) is predicted at each time-step for each annotation vector h_j with a simple fully connected neural network.



Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

Attention Mechanism

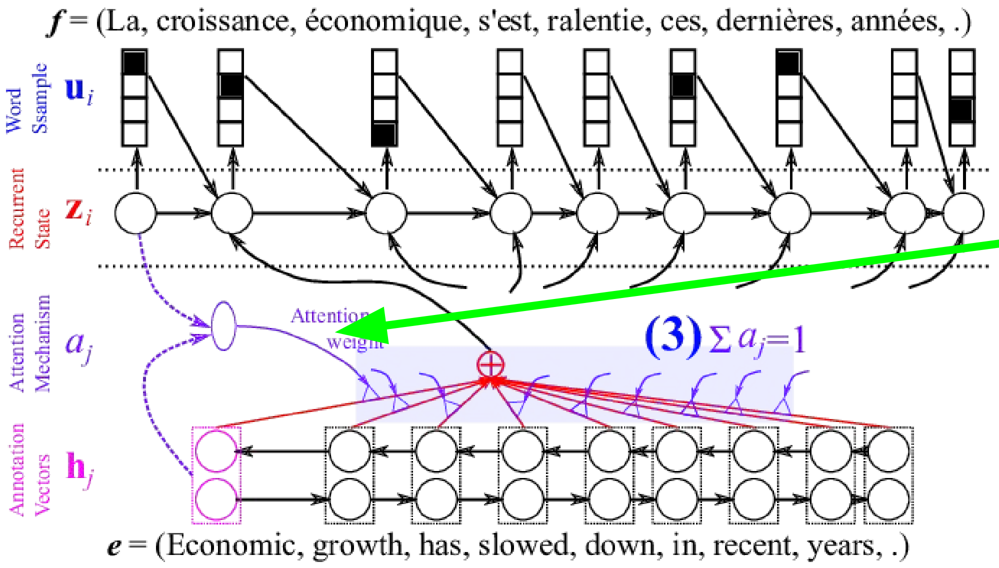
Once a relevance score (weight) is estimated for each word, they are normalized with a softmax function so they sum up to 1.



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

Attention Mechanism

Finally, a context-aware representation c_i for the output word at timestep i can be defined as:

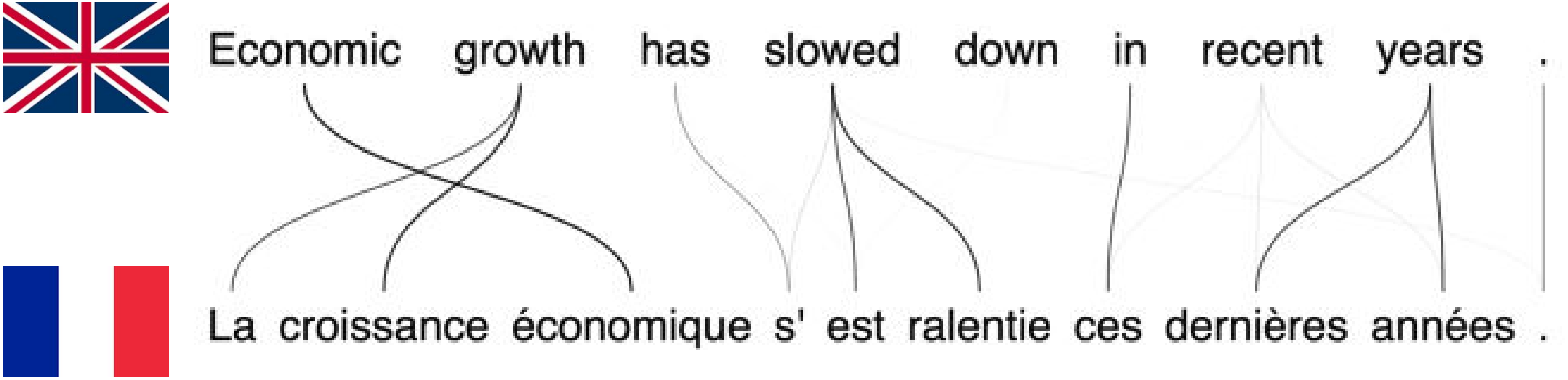


$$c_i = \sum_{j=1}^{T} \alpha_j h_j$$

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

Attention Mechanism

The model automatically finds the correspondence structure between two languages (alignment).

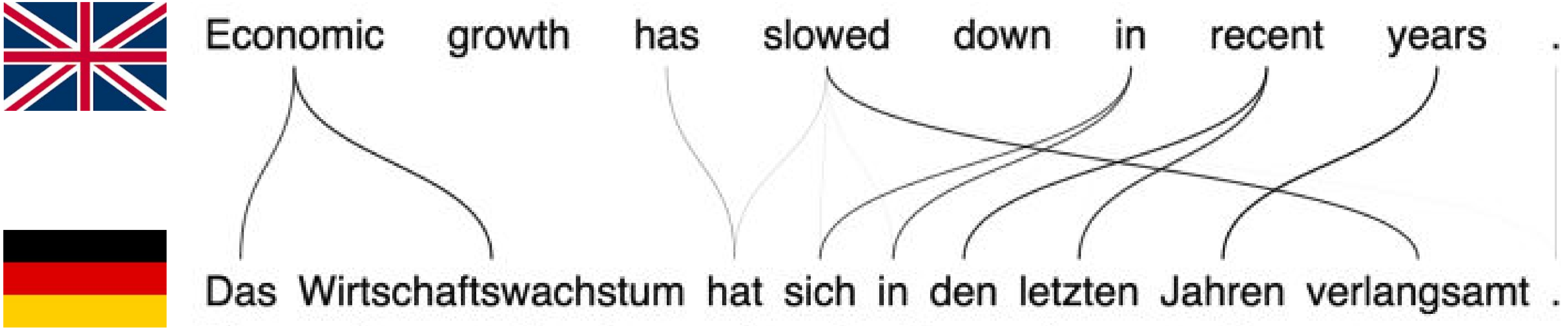


(Edge thicknesses represent the attention weights found by the attention model)

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

Attention Mechanism

The model automatically found the correspondence structure between two languages (alignment).



(Edge thicknesses represent the attention weights found by the attention model)

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

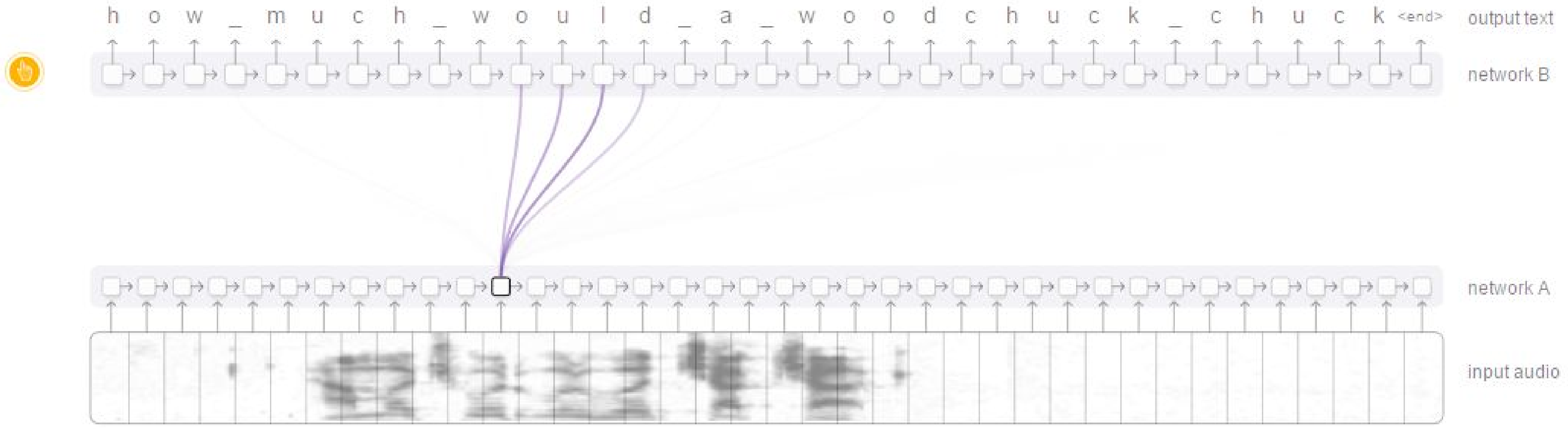
Attention Models

Attend to different parts of the input to optimize a certain output

Attention Models

Attend to different parts of the input to optimize a certain output

Input: Audio features; Output: Text



Chan et al. [Listen, Attend and Spell.](#) ICASSP 2016
Source: [distill.pub](#)

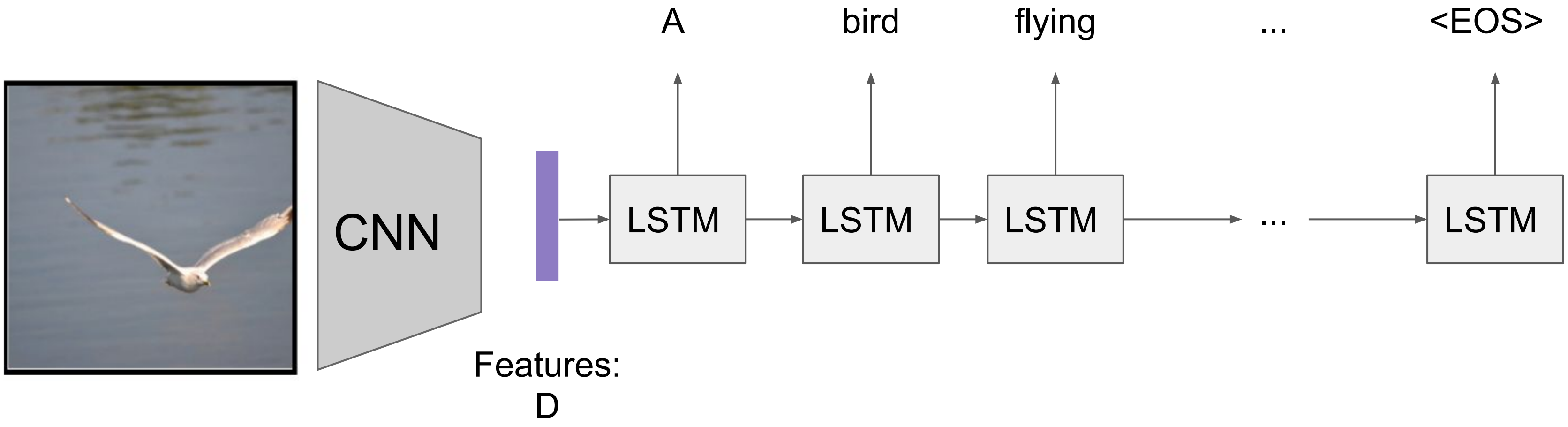
Attention Models

Attend to different parts of the input to optimize a certain output

Input: Image; Output: Text

A bird flying over a body of water

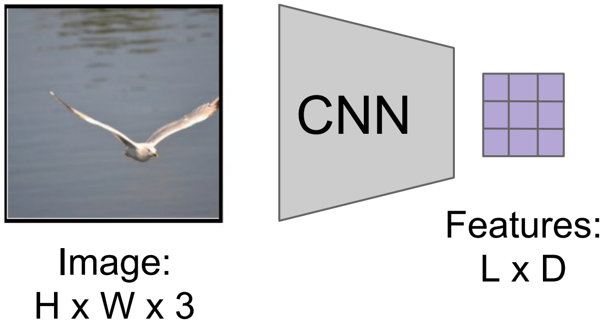
LSTM Decoder for Image Captioning



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

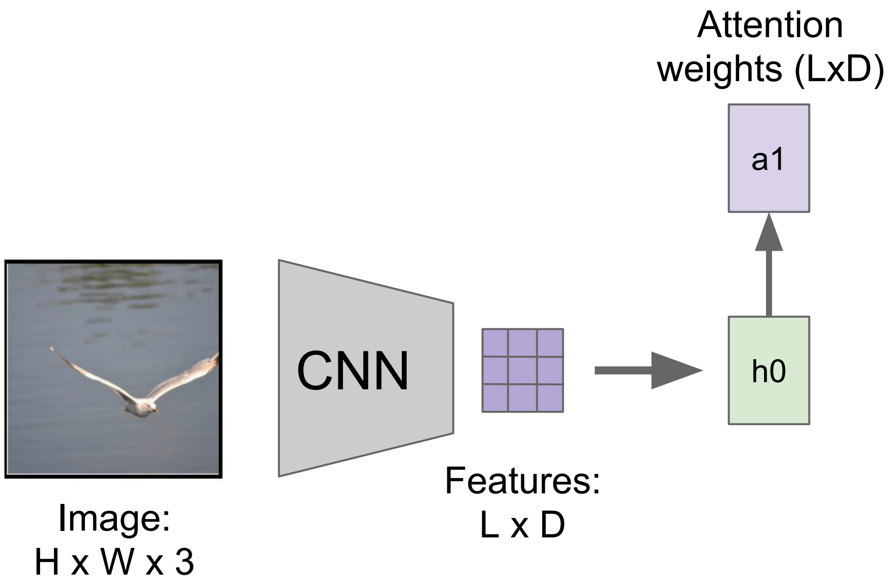
Vinyals et al. Show and tell: A neural image caption generator. CVPR 2015

Attention for Image Captioning



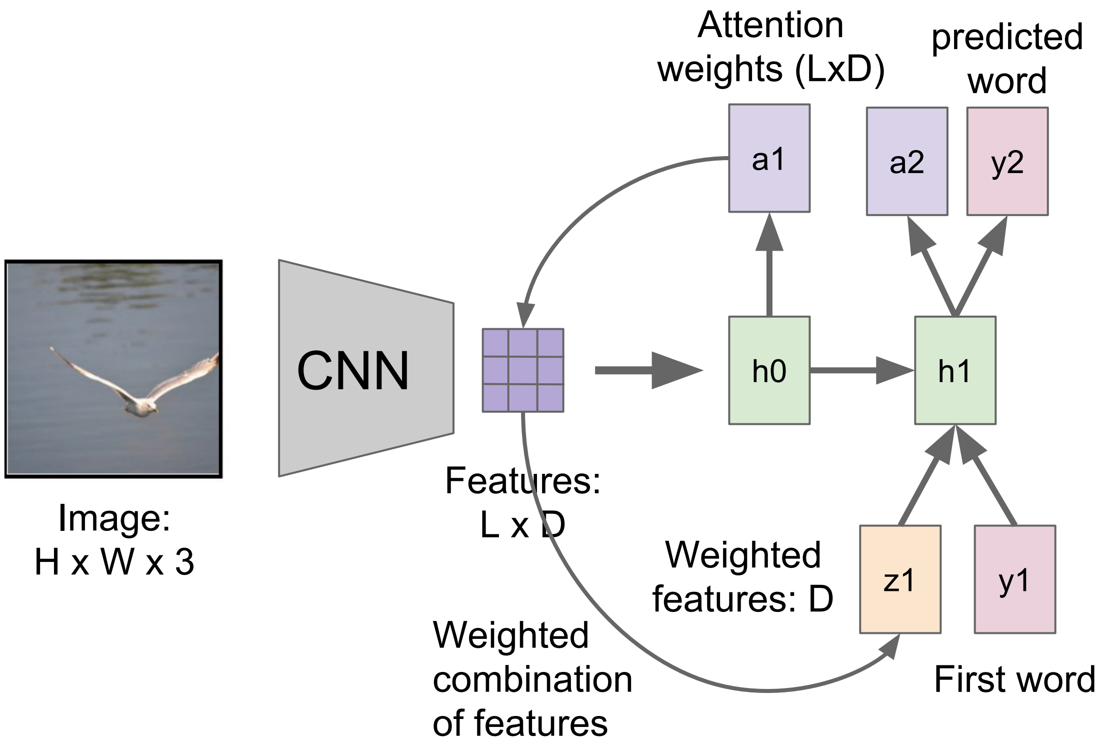
Slide Credit: CS231n

Attention for Image Captioning



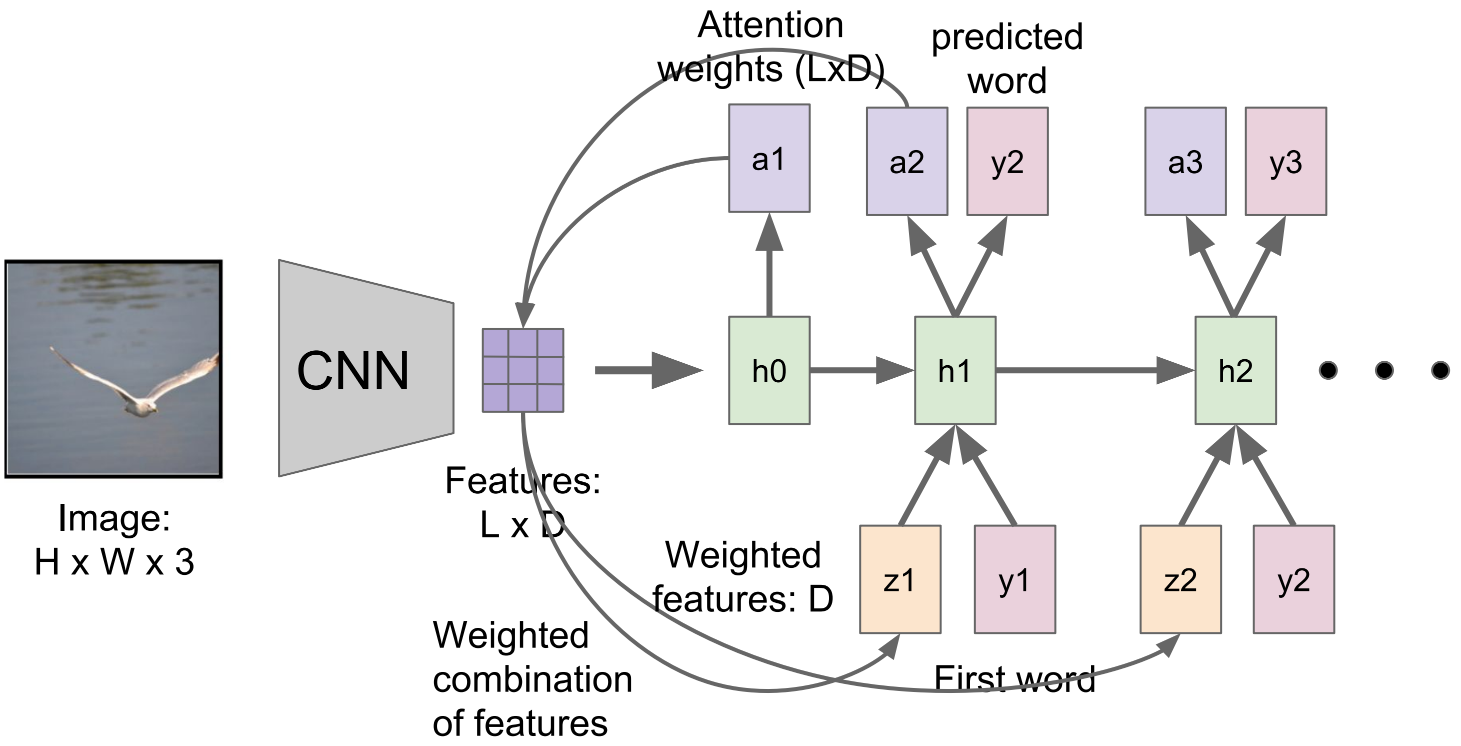
Slide Credit: [CS231n]

Attention for Image Captioning



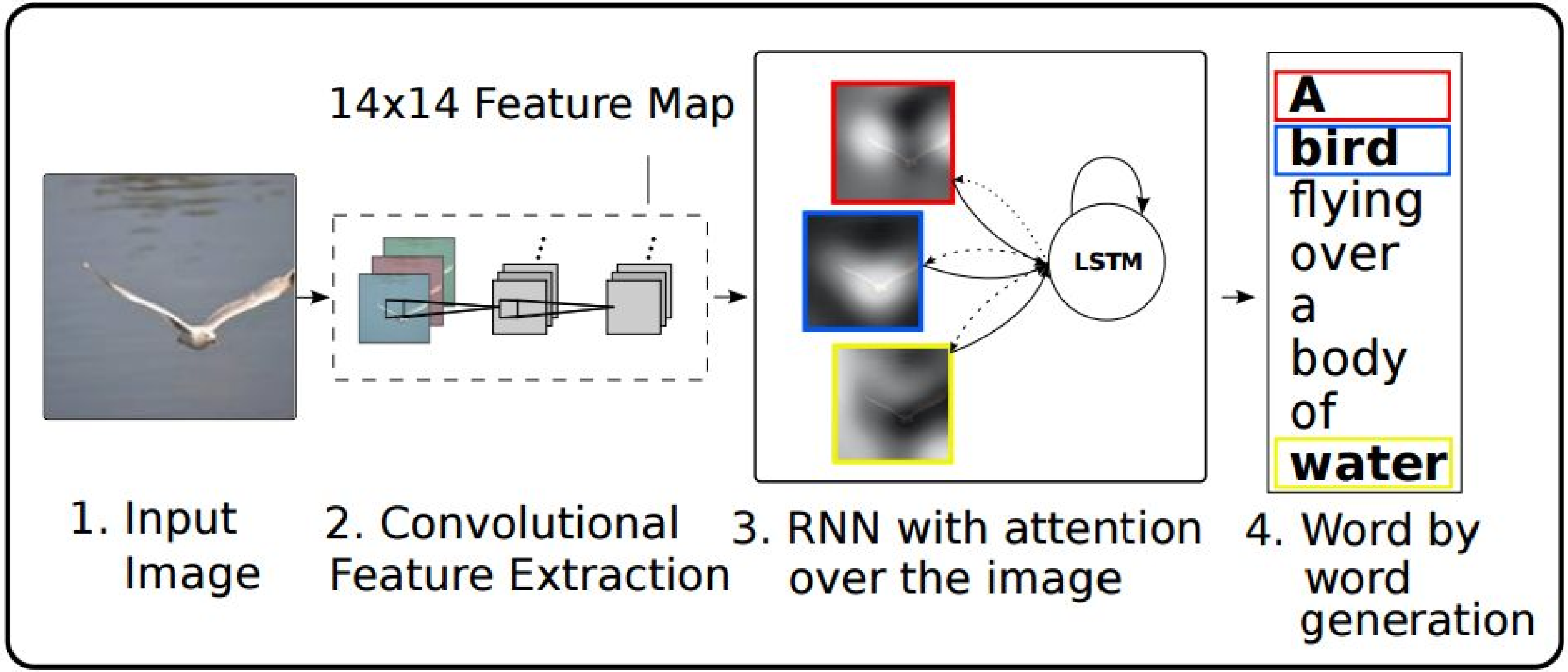
Slide Credit: [CS231n]

Attention for Image Captioning



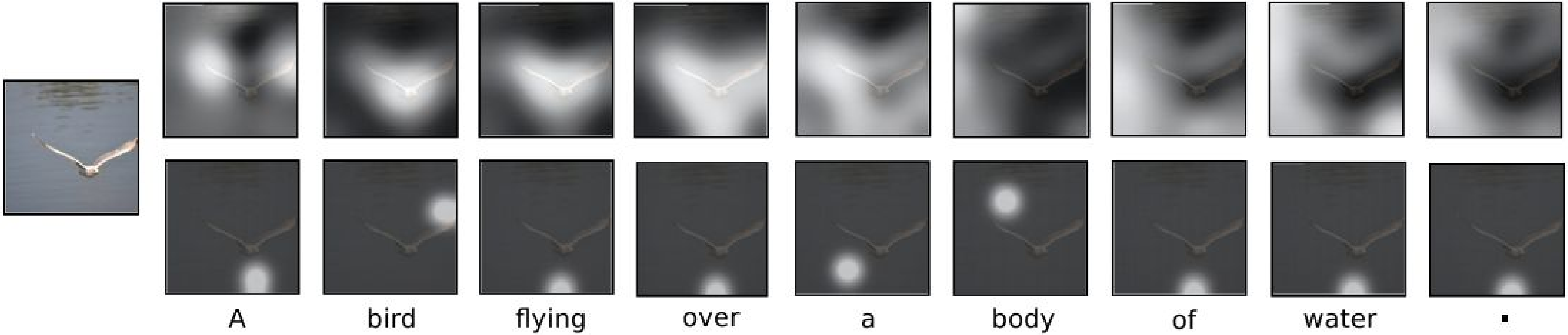
Slide Credit: [CS231n](CS231n)

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

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

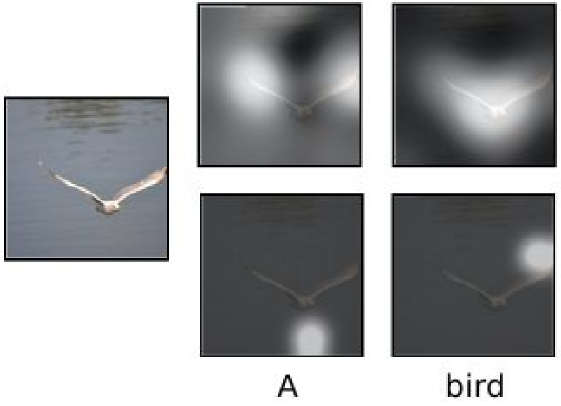
A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

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

Attention for Image Captioning

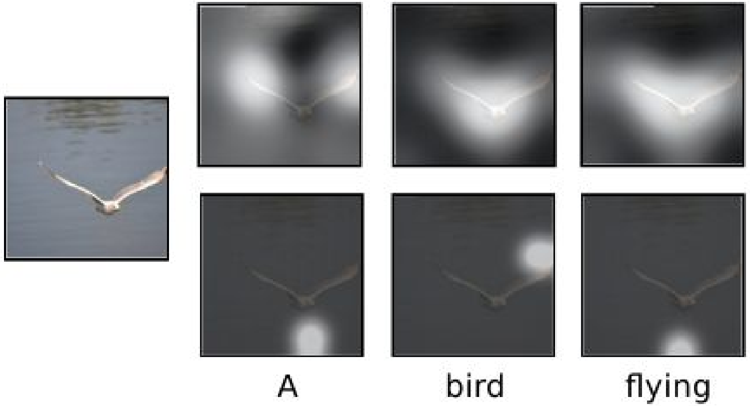
Some outputs can probably be predicted without looking at the image...



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

Attention for Image Captioning

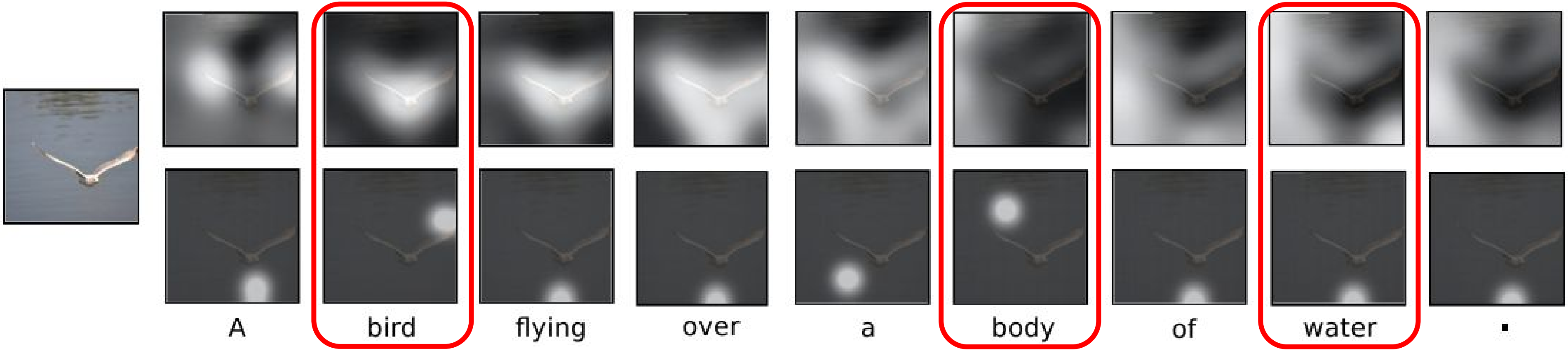
Some outputs can probably be predicted without looking at the image...



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

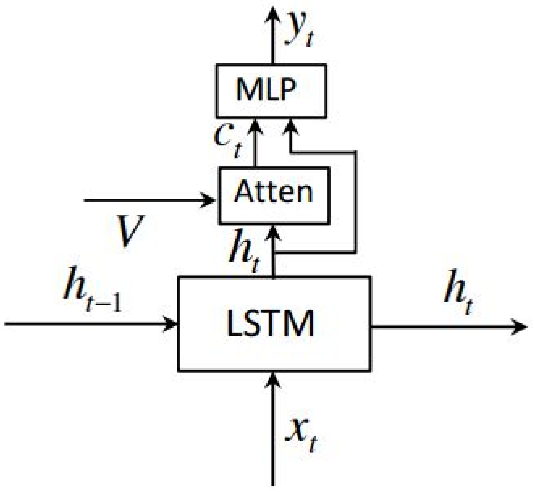
Attention for Image Captioning

Can we focus on the image only when necessary?

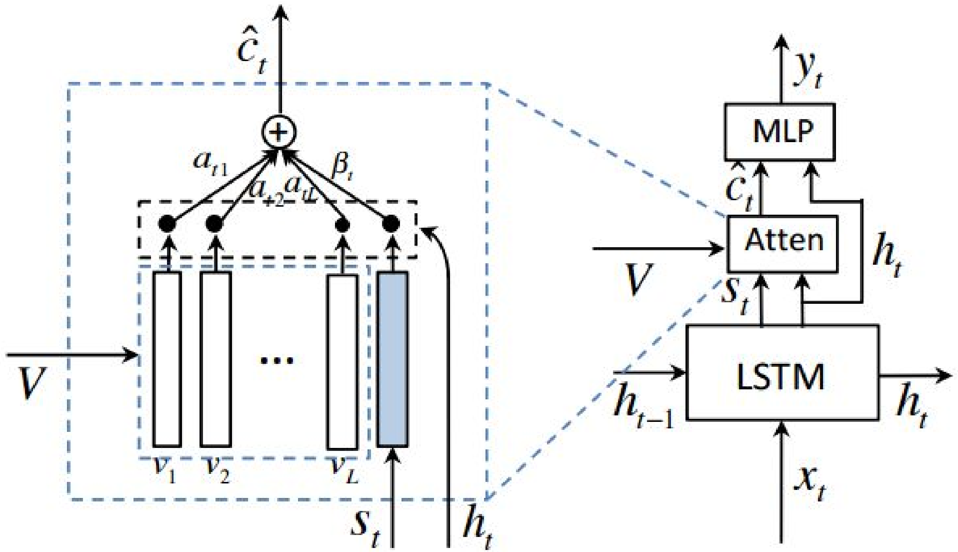


Xu et al. [Show, Attend and Tell: Neural Image Caption Generation with Visual Attention](). ICML 2015

Attention for Image Captioning



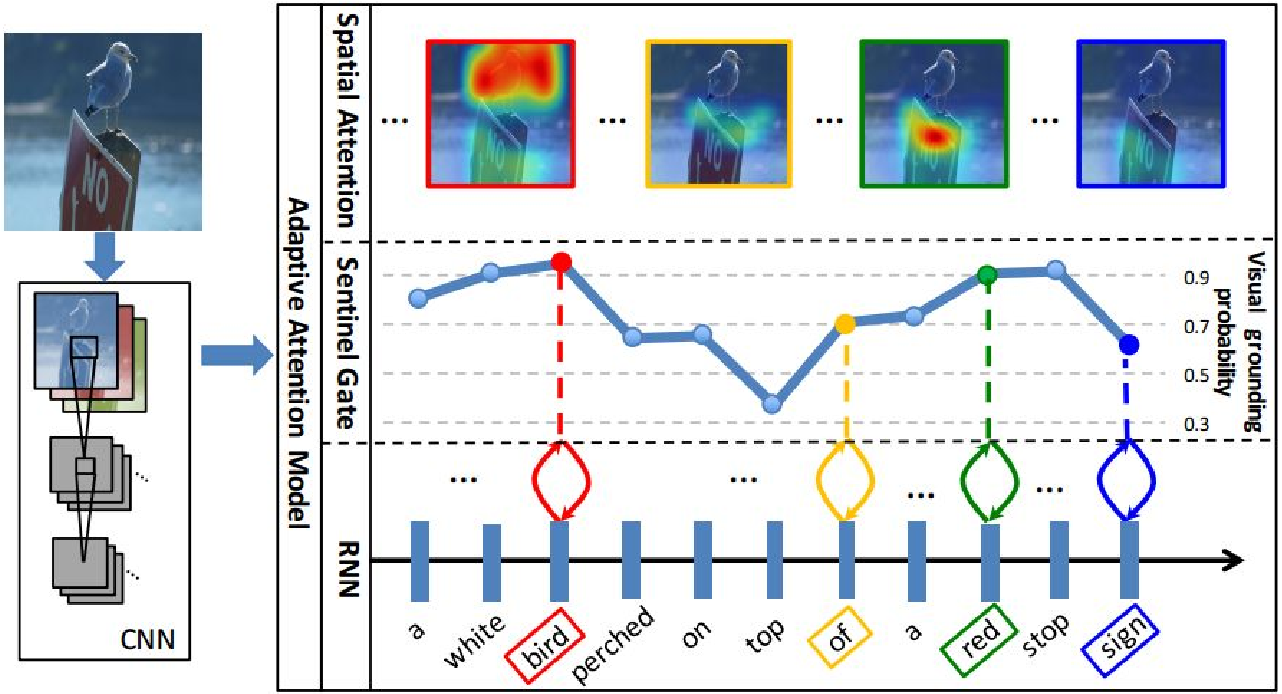
"Regular" Attention



Attention with Sentinel

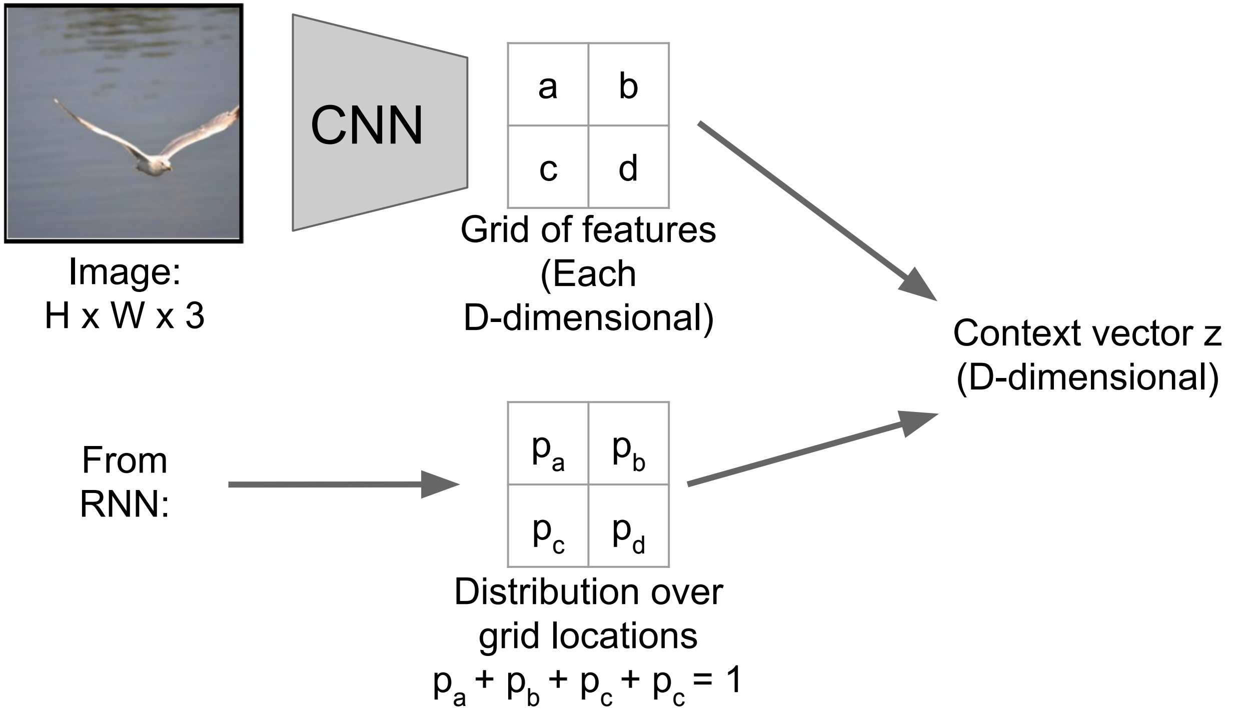
Lu et al. [Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning.] CVPR 2017

Attention for Image Captioning



Lu et al. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. CVPR 2017

Soft Attention



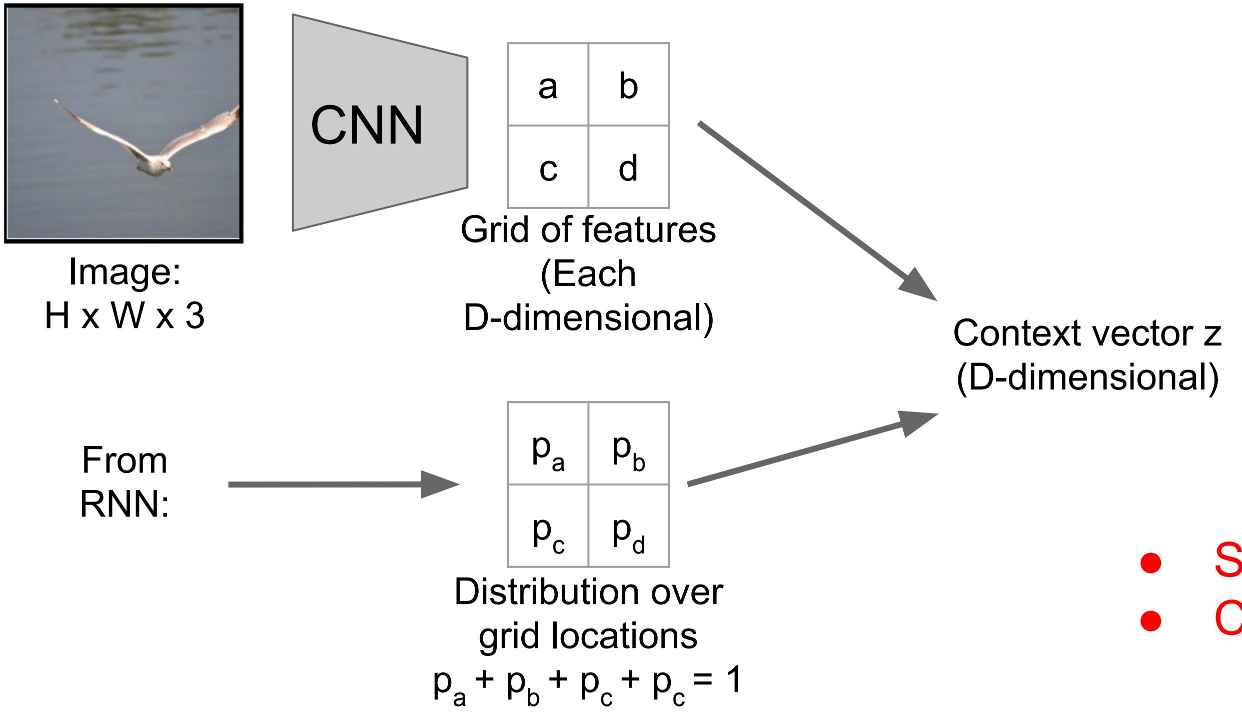
Soft attention:
Summarize ALL locations
$z = p_a a + p_b b + p_c c + p_d d$

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



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

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

Cropped and
rescaled image:
X x Y x 3

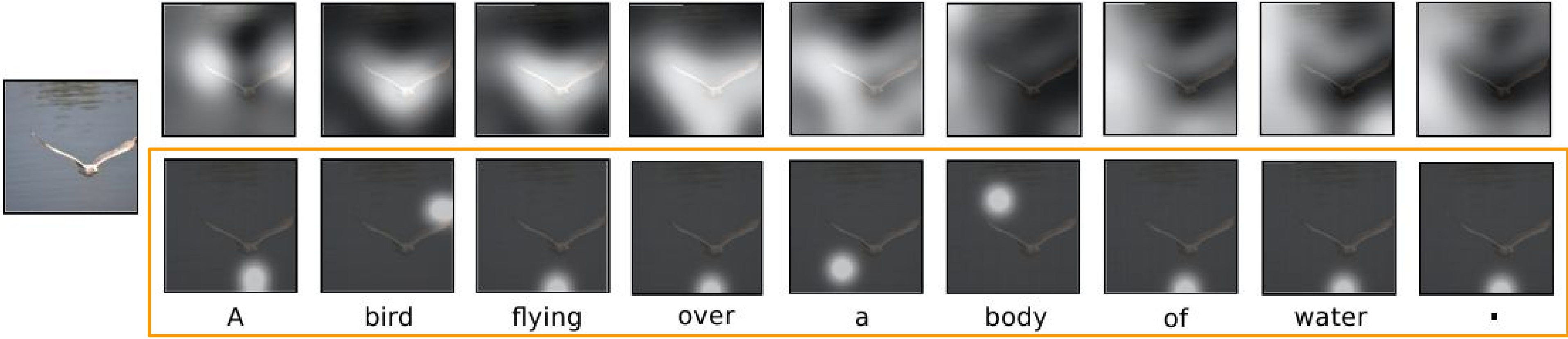
Hard attention:
Sample a subset
of the input

Not a differentiable function !

Can't train with backprop :(

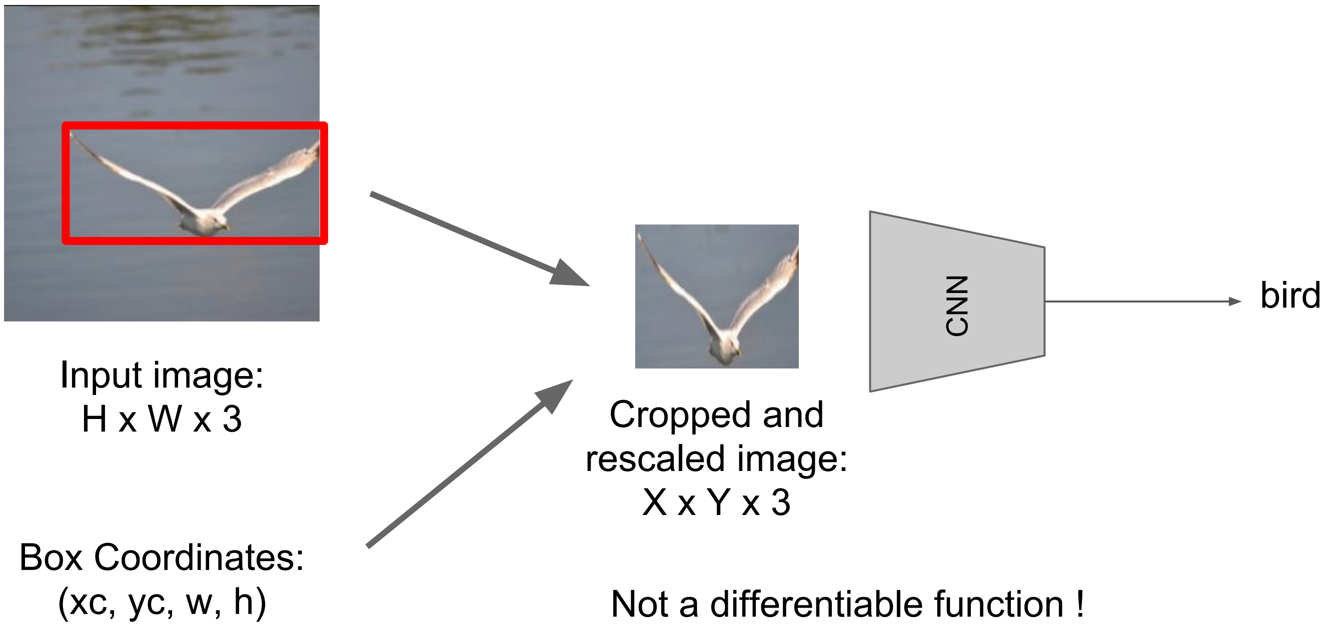
Need other optimization strategies
e.g.: [reinforcement learning]

Attention for Image Captioning



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

Hard Attention



Not a differentiable function !

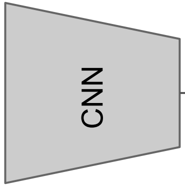
Can't train with backprop :(

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



bird

Not a differentiable function !

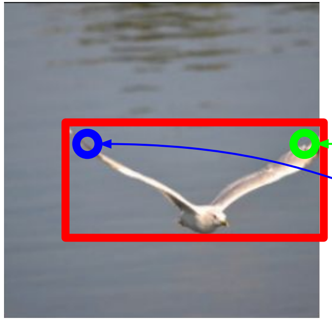
Can't train with backprop :(

Make it differentiable

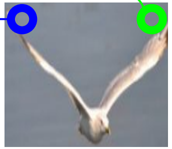
Train with backprop :)

Jaderberg et al. [Spatial Transformer Networks](). NIPS 2015

Spatial Transformer Networks



Can we make this function differentiable?



Input image:
H x W x 3

Cropped and rescaled image:
X x Y x 3

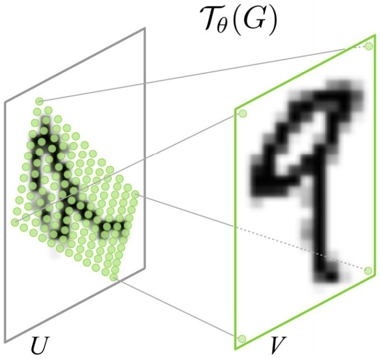
Box Coordinates:
(xc, yc, w, h)

Mapping given by box coordinates (translation + scale)

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

Network attends to input by predicting θ

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



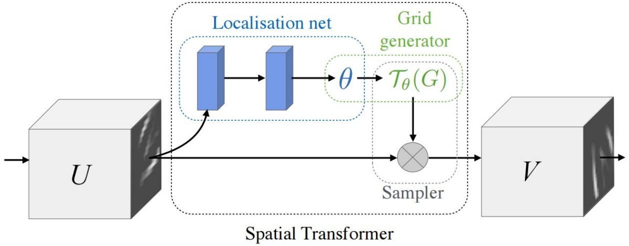
Repeat for all pixels in *output*

Jaderberg et al. Spatial Transformer Networks. NIPS 2015

Slide Credit: CS231n

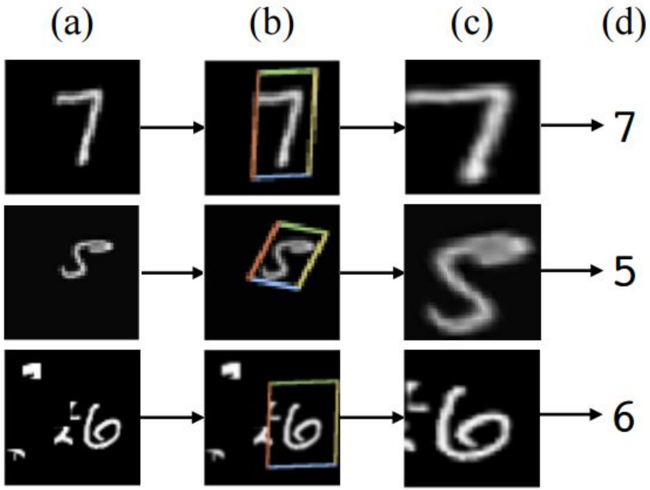
Spatial Transformer Networks

Differentiable module



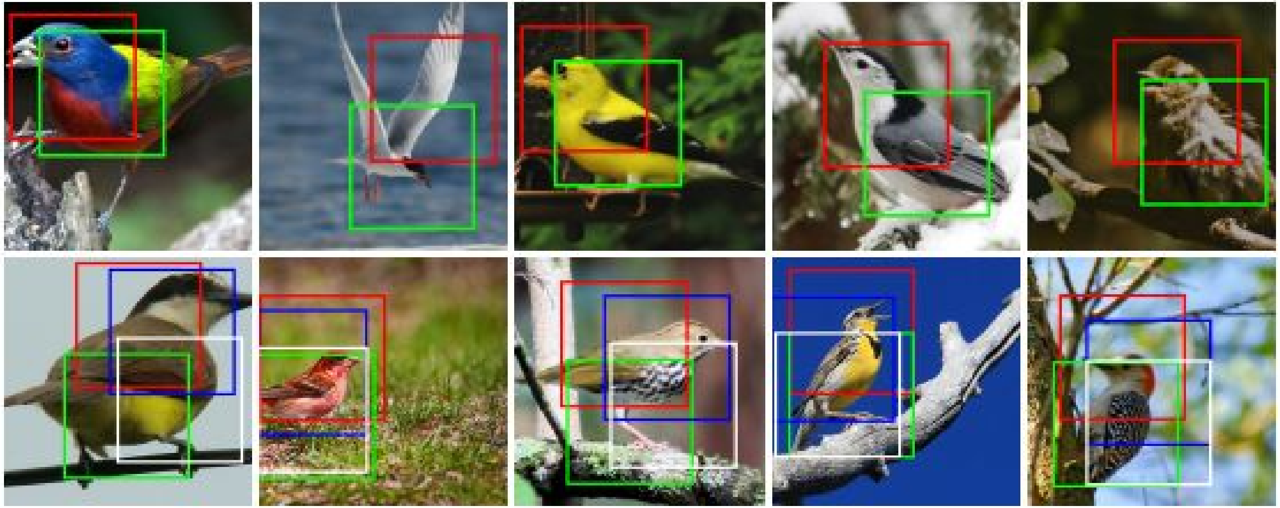
Easy to incorporate in any network, anywhere !

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

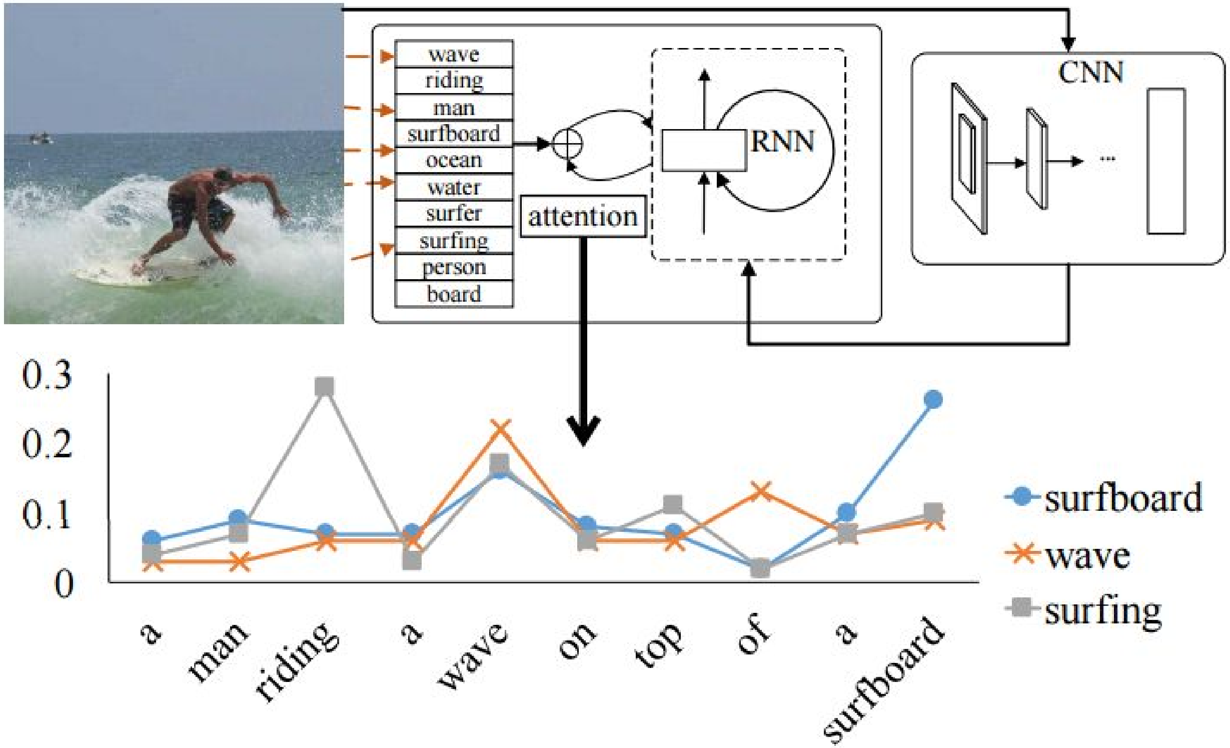


Fine-grained classification

Also used as an alternative to RoI pooling in proposal-based detection & segmentation pipelines

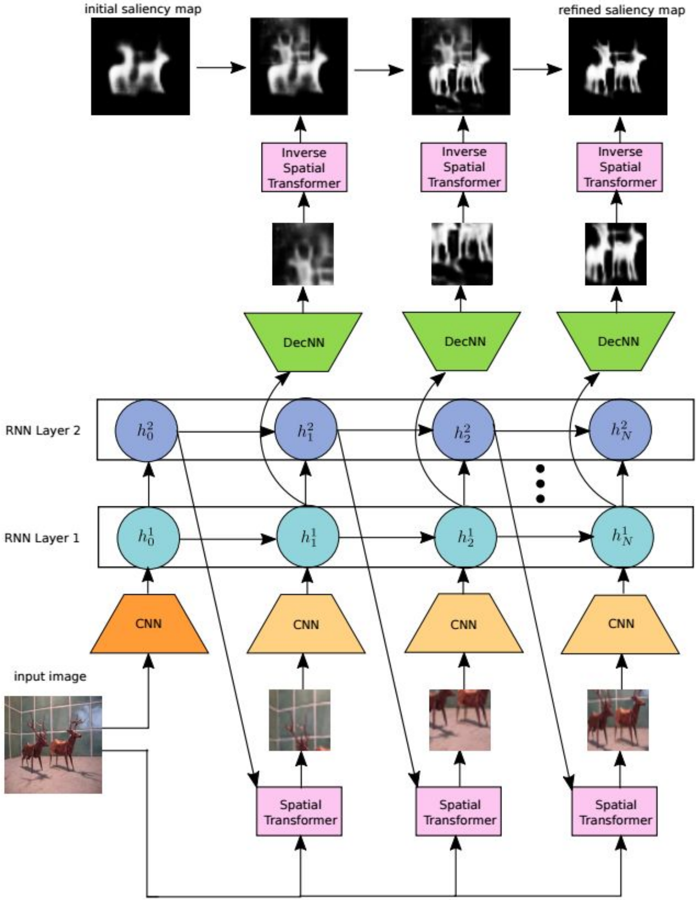
Jaderberg et al. [Spatial Transformer Networks](). NIPS 2015

Semantic Attention: Image Captioning



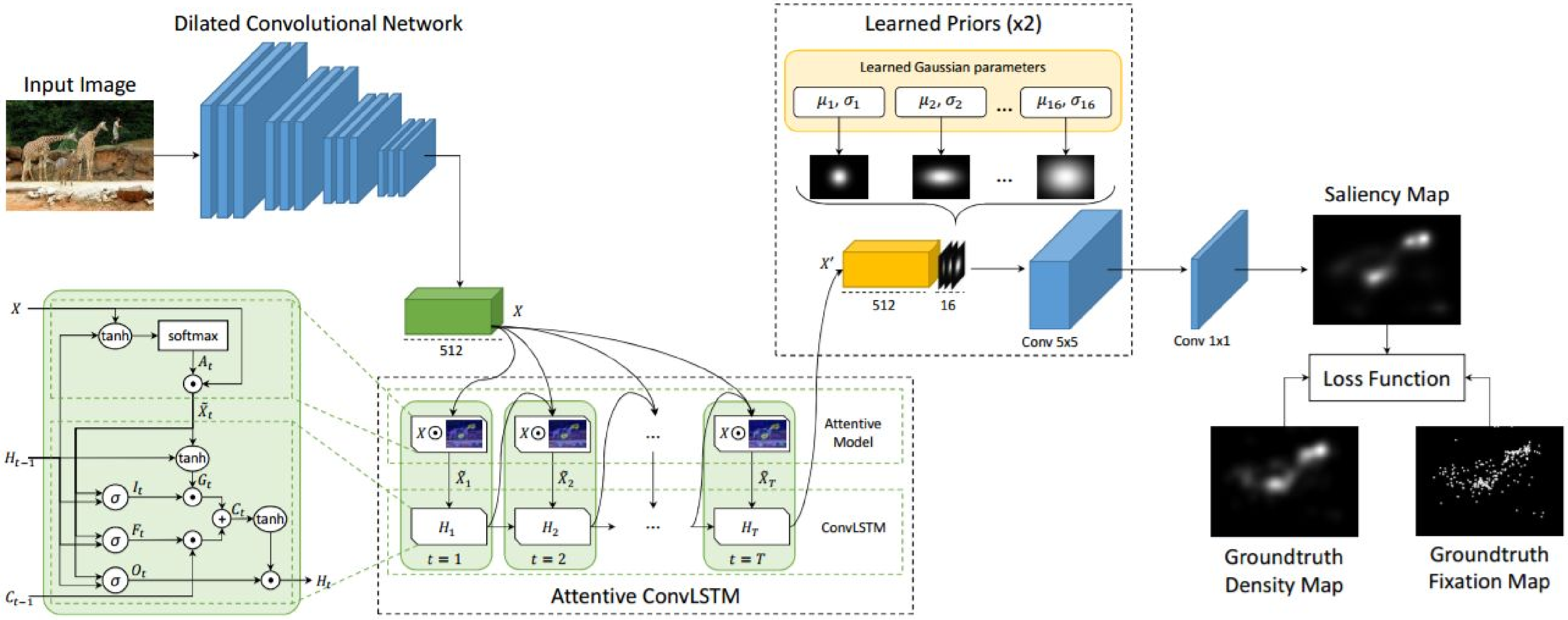
You et al. [Image Captioning with Semantic Attention](). CVPR 2016

Visual Attention: Saliency Detection



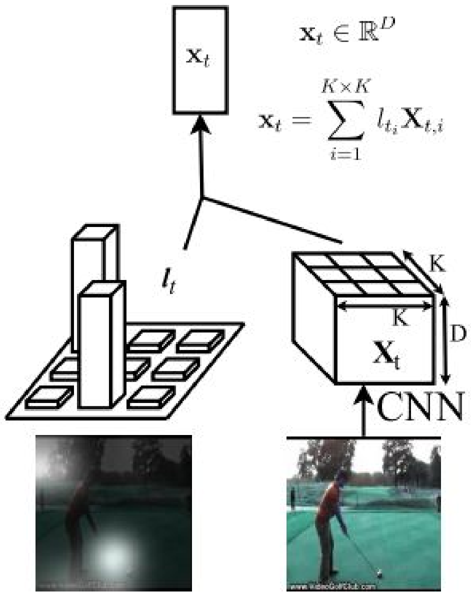
Kuen et al. [Recurrent Attentional Networks for Saliency Detection](). CVPR 2016

Visual Attention: Fixation Prediction

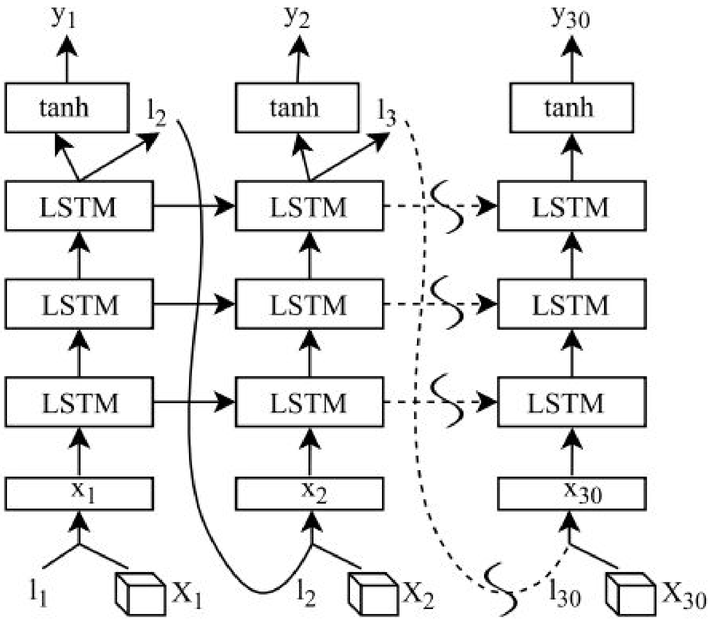


Cornia et al. [Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model.]

Visual Attention: Action Recognition



(a) The soft attention mechanism

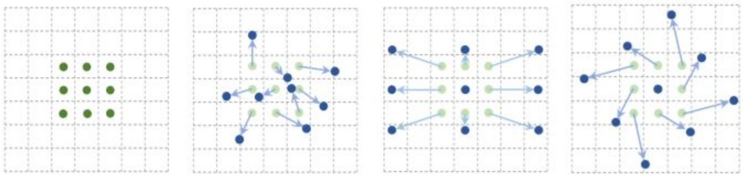


(b) Our recurrent model

Sharma et al. Action Recognition Using Visual Attention. arXiv 2016

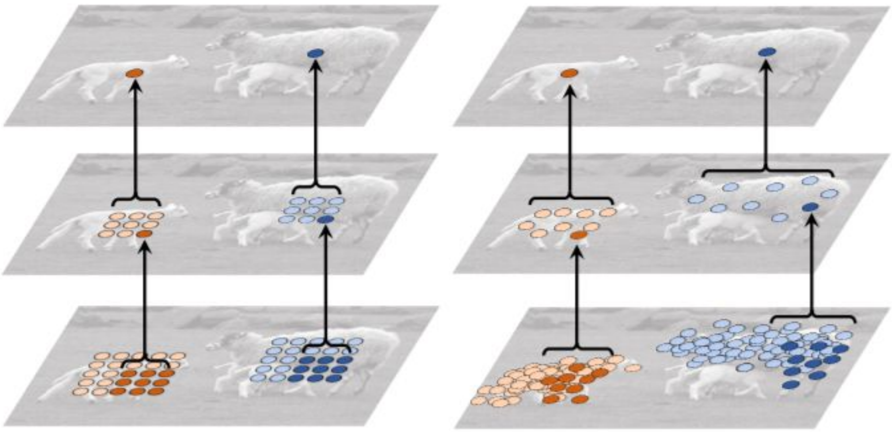
Deformable Convolutions

Dynamic & learnable receptive field



(a) (b) (c) (d)

Figure 1: Illustration of sampling grids in 3×3 regular and deformable convolutions. (a) regular sampling grid (green points) of standard convolution. (b) deformed sampling grid with augmented offsets (blue arrows) in deformable convolution. (c)(d) are special cases of (b), showing that the deformable convolution generalizes scale, aspect ratio and rotation transformations.



(a) standard convolution (b) deformable convolution

Dai, Qi, Xiong, Li, Zhang et al. [Deformable Convolutional Networks](). arXiv Mar 2017

Questions?