CS182/282A: Designing, Visualizing and Understanding Deep Neural Networks

John Canny

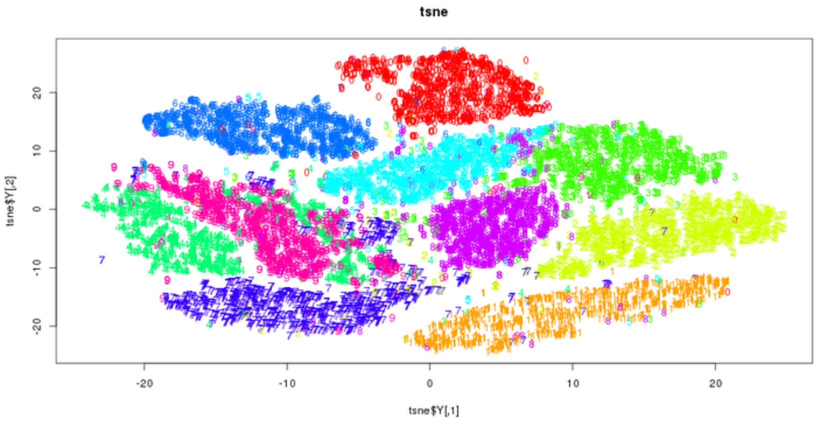
Spring 2020

Lecture 11: Attention

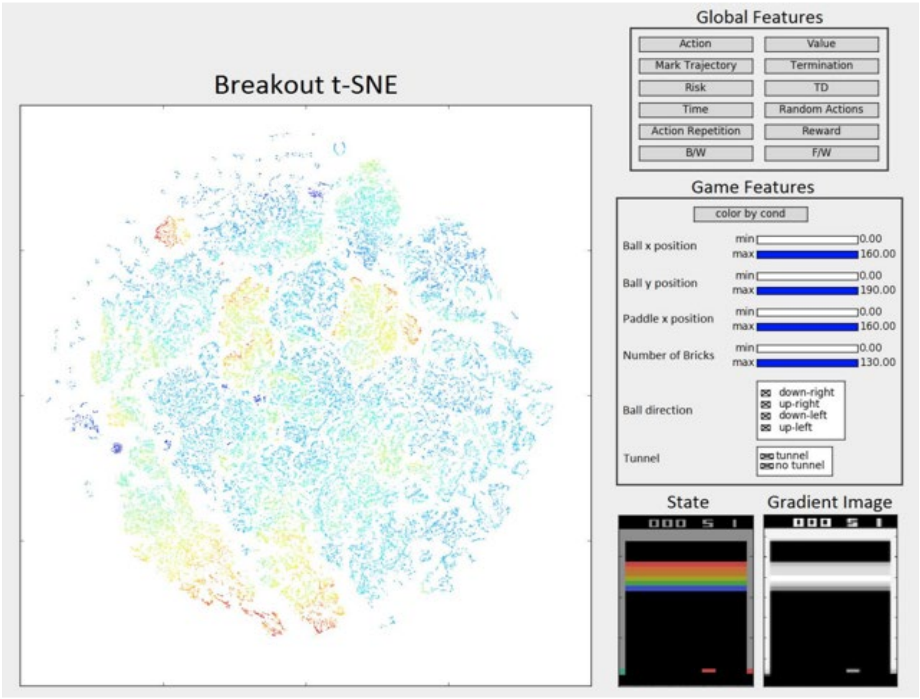
Last Time: t-SNE

Embed high-dimensional points so that locally, pairwise distances are conserved.

Example embedding of MNIST digit images (0-9) in 2D



Graying the black box: Understanding DQNs
Zahavy, Zrihem, Mannor 2016

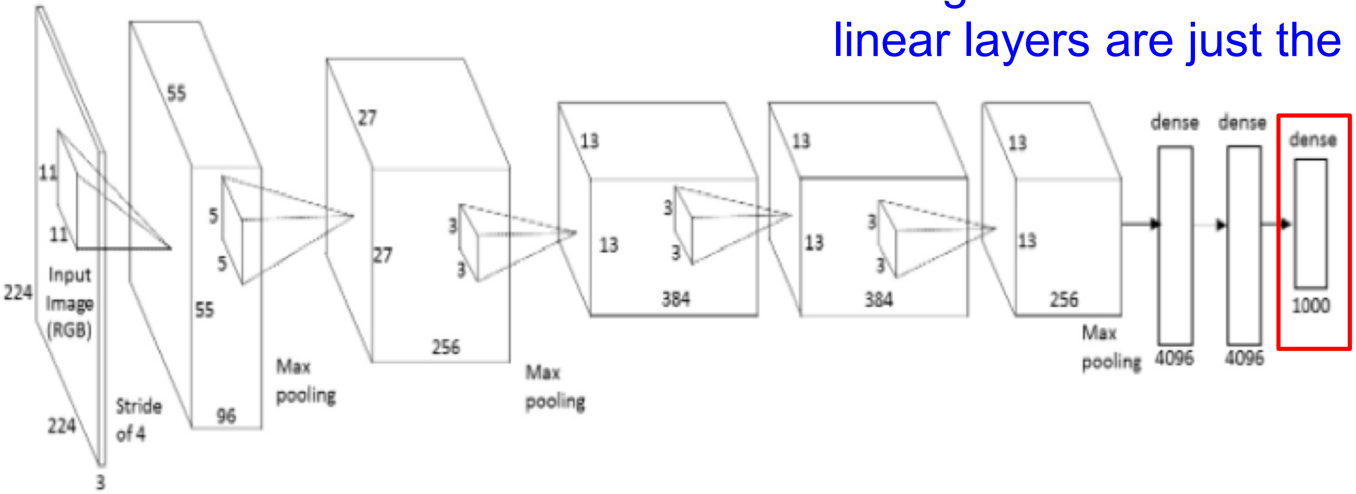


The embedding shows clustering of the *activations of the agent's policy network* for different frames of breakout.

Last Time:
Activation Maximization

$$\arg\max_{I} S_c(I) - \boxed{\lambda \|I\|_2^2}$$

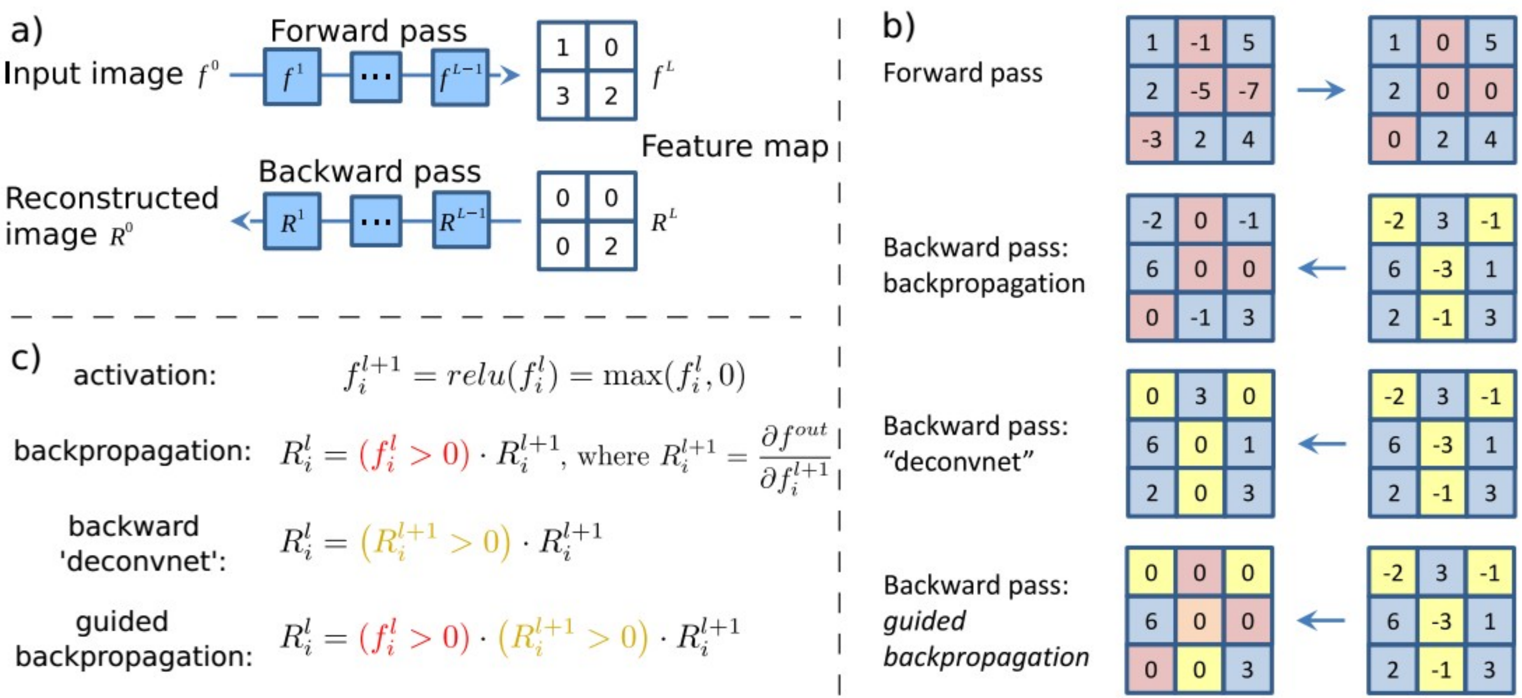
L2 regularization: Saliency maps for linear layers are just the layer weights.



Generate an image that maximizes a class score

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps
Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, 2014

Last Time: Deconv approaches



Last Time:
Neural Style Transfer

[A Neural Algorithm of Artistic Style by Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, 2015]

good implementation by Justin Johnson in Torch:
https://github.com/jcjohnson/neural-style

Midterm 1

Weds February 26th, 7-8:30pm

Genetics & Plant Biology, Room 100, A-K last names
North Gate Hall, Room 105, L-R last names
Latimer Hall, Room 120, S-Z last names

Closed-book, one double-sided sheet of notes

This Time: Attention

Defn: “the regarding of someone or something as interesting or important.”

Attention is one of the most important ideas in deep networks in the last decade…

It cross-cuts computer vision, NLP, speech, RL,…

Early attention models

Larochelle and Hinton, 2010, “Learning to combine foveal glimpses with a third-order Boltzmann machine”

Misha Denil et al, 2011, “Learning where to Attend with Deep Architectures for Image Tracking”

2014: Neural Translation Breakthroughs

- Devlin et al, ACL'2014
- Cho et al EMNLP'2014
- Bahdanau, Cho & Bengio, arXiv sept. 2014
- Jean, Cho, Memisevic & Bengio, arXiv dec. 2014
- Sutskever et al NIPS'2014

Other Applications

- Ba et al 2014, **Visual attention for recognition**
- Chorowski et al, 2014, **Speech recognition**
- Graves et al 2014, **Neural Turing machines**
- Yao et al 2015, **Video description generation**
- Vinyals et al, 2015, **Conversational Agents**
- Xu et al 2015, **Image caption generation**
- Xu et al 2015, **Visual Question Answering**
- Viswani et al, 2017, **Attention Is All You Need**
- Devlin et al, 2018, **BERT: Bidirectional Transformers for Language**

Soft vs Hard Attention Models

Hard attention:

Attend to a single input location.

Can’t use gradient descent.

Need **reinforcement learning.**

Soft attention:

Compute a weighted combination (attention) over some inputs using an attention network.

Can use backpropagation to train end-to-end.

Reinforcement vs. Supervised Learning

Supervised Learning:

Input samples are independent, each sample x receives a label y.

The pair (x, y) is assigned a loss value which is assumed to be *differentiable*.

Reinforcement Learning:

Learner visits a sequence of (correlated) states s_t in an epoch $t = 1, \ldots, T$

At time t, learner performs action a_t and receives reward r_t from the environment.

Agent tries to maximizes the sum of rewards over an epoch.

Reinforcement vs. Supervised Learning

Reinforcement Learning:

Learner visits a sequence of (correlated) states s_t in an epoch $t = 1, \ldots, T$

At time t, learner performs action a_t and receives reward r_t **from the environment**.

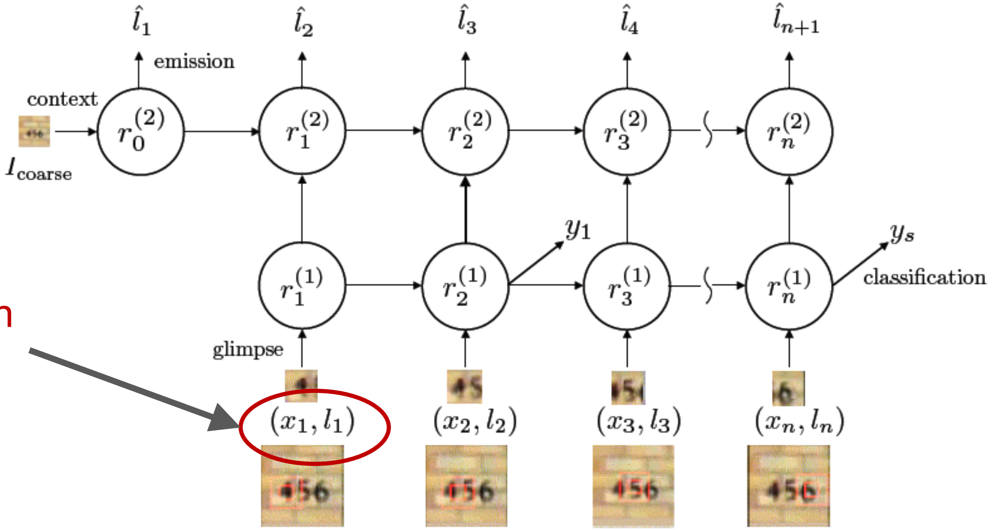
Agent tries to **maximizes the sum of rewards** over an epoch.

Note: the agent cannot differentiate the reward to optimize it (it comes from the environment). This is true also for hard attention.

Attention for Recognition (Ba et al 2014)

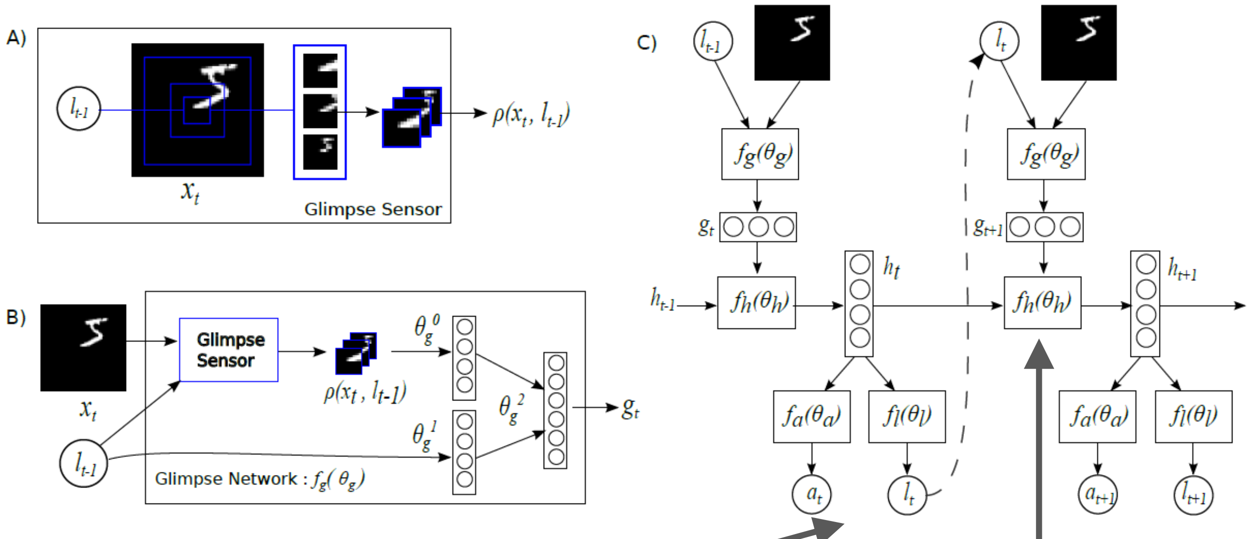
- RNN-based model.
- Hard attention.
- Required reinforcement learning.

l_i = location
x_i = image



Attention for Recognition (Mnih et al 2014)

- Glimpses are retinal (graded resolution) images

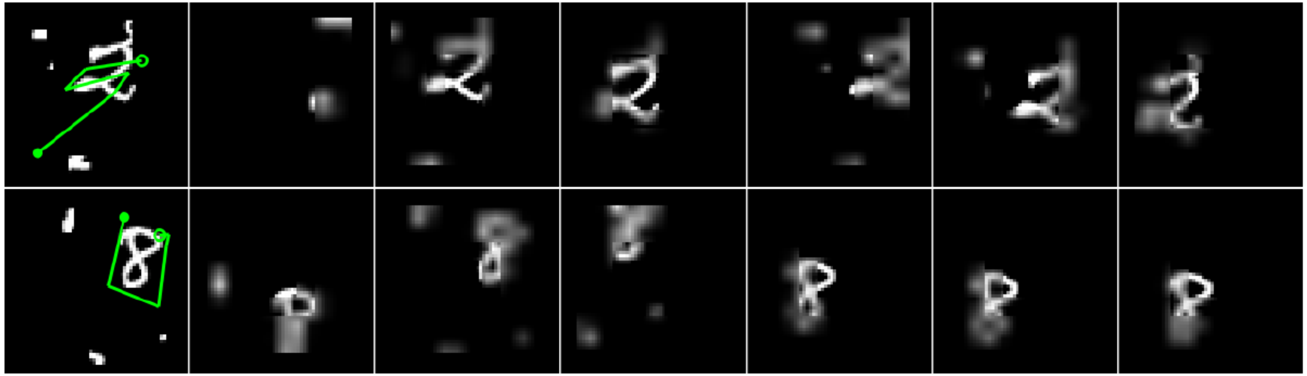


l_i = location
a_i = action (classification)

$f_h()$ = 256-dimensional LSTM

Attention for Recognition (Mnih et al 2014)

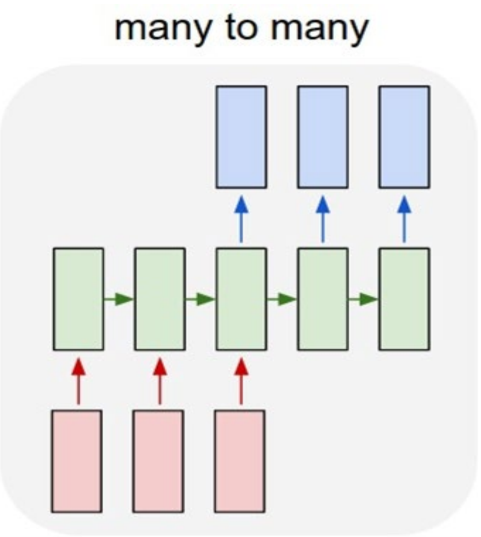
- Glimpse trace on some digit images:
- Green line shows trajectory, other images are the glimpses themselves.



Soft Attention for Translation

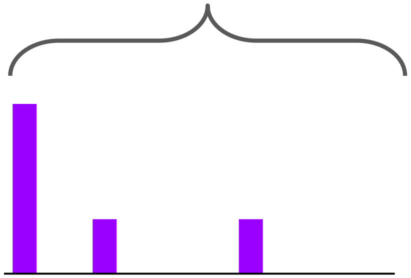
“I love coffee” -> “Me gusta el café”

Bahdanau et al, “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015



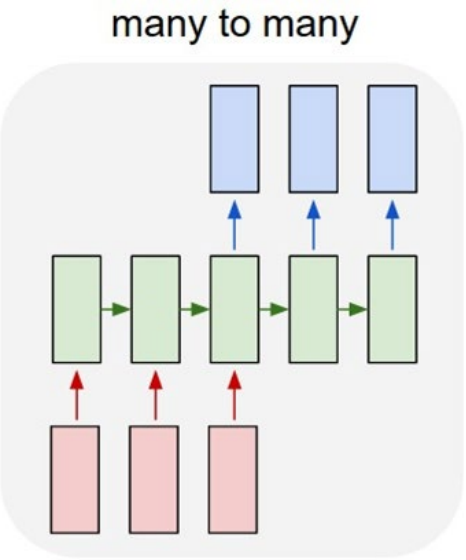
Soft Attention for Translation

Distribution over input words



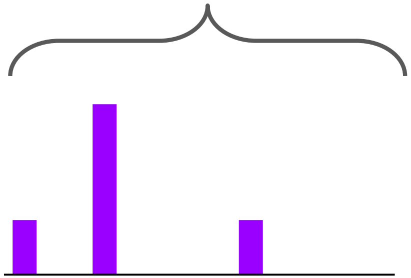
"I love coffee" -> "Me gusta el café"

Bahdanau et al, "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015



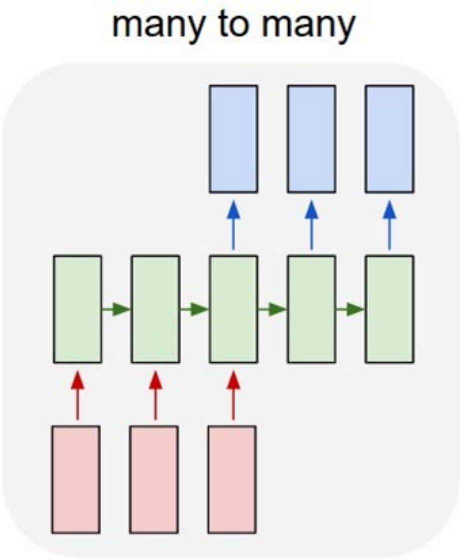
Soft Attention for Translation

Distribution over input words

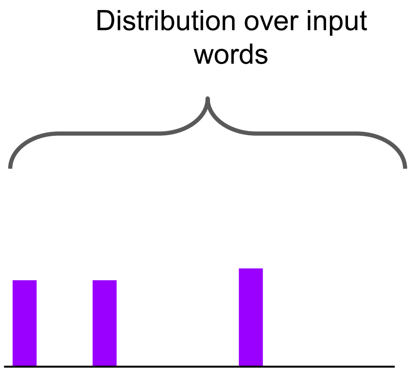


"I love coffee” -> “Me gusta el café”

Bahdanau et al, “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015

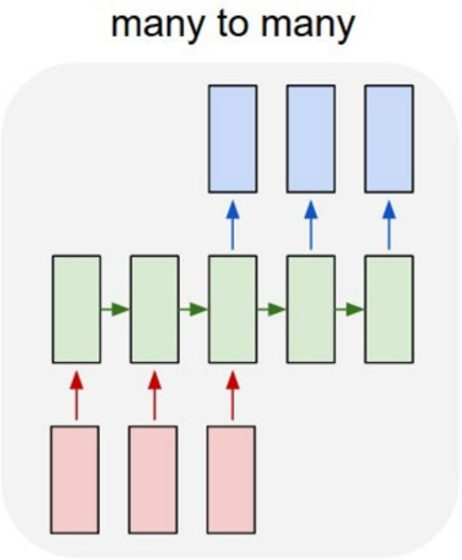


Soft Attention for Translation



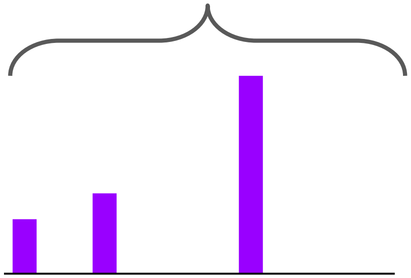
"I love coffee" -> "Me gusta el café"

Bahdanau et al, "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015



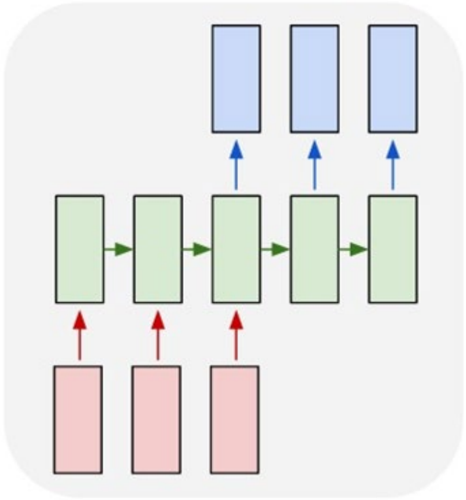
Soft Attention for Translation

Distribution over input words



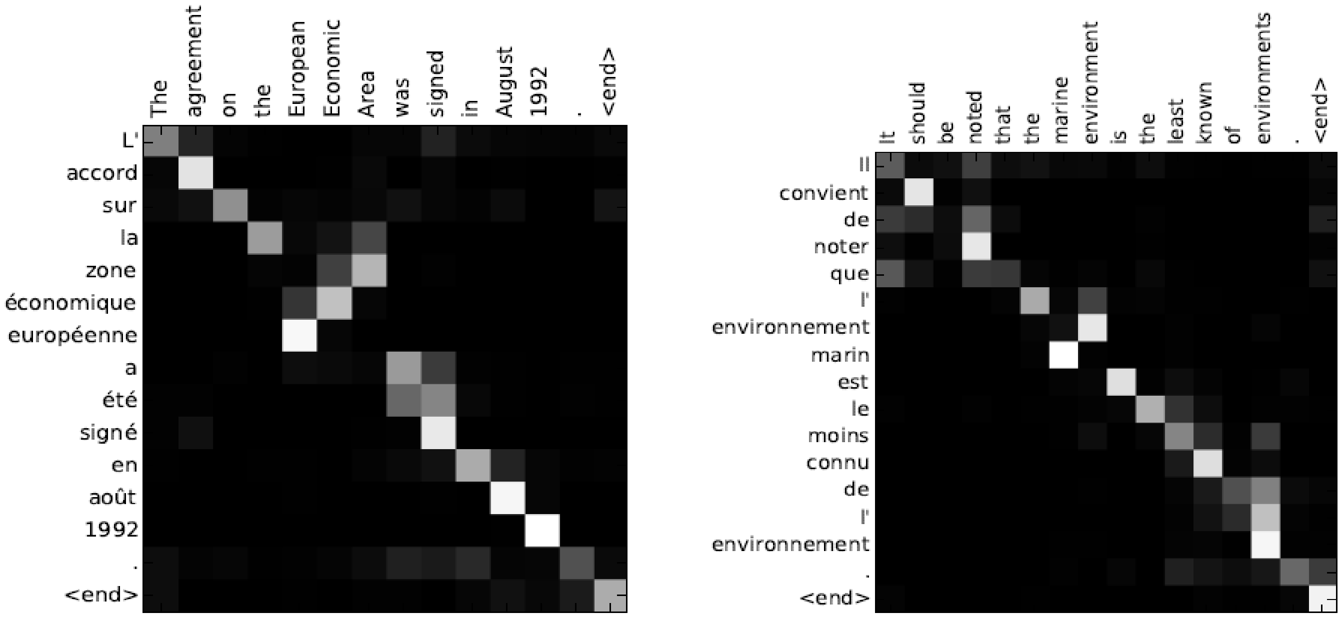
"I love coffee” -> “Me gusta el café”

many to many



Bahdanau et al, “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015

Soft Attention for Translation



(a) (b)

Bahdanau et al, "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015

Soft Attention for Translation

Reached State of the art in one year:

(a) **English→French (WMT-14)**

	NMT(A)	Google	P-SMT
NMT	32.68	30.6*	**37.03•**
+Cand	33.28	–	
+UNK	33.99	32.7°	
+Ens	36.71	**36.9°**	

(b) **English→German (WMT-15)**

Model	Note
24.8	Neural MT
24.0	U.Edinburgh, Syntactic SMT
23.6	LIMSI/KIT
22.8	U.Edinburgh, Phrase SMT
22.7	KIT, Phrase SMT

(c) **English→Czech (WMT-15)**

Model	Note
18.3	Neural MT
18.2	JHU, SMT+LM+OSM+Sparse
17.6	CU, Phrase SMT
17.4	U.Edinburgh, Phrase SMT
16.1	U.Edinburgh, Syntactic SMT

Yoshua Bengio, NIPS RAM workshop 2015

RNN for Captioning

Image:
H x W x 3

RNN for Captioning

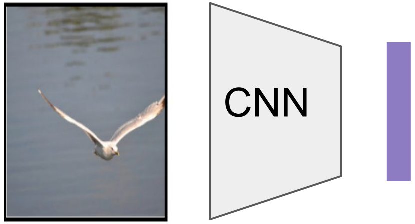
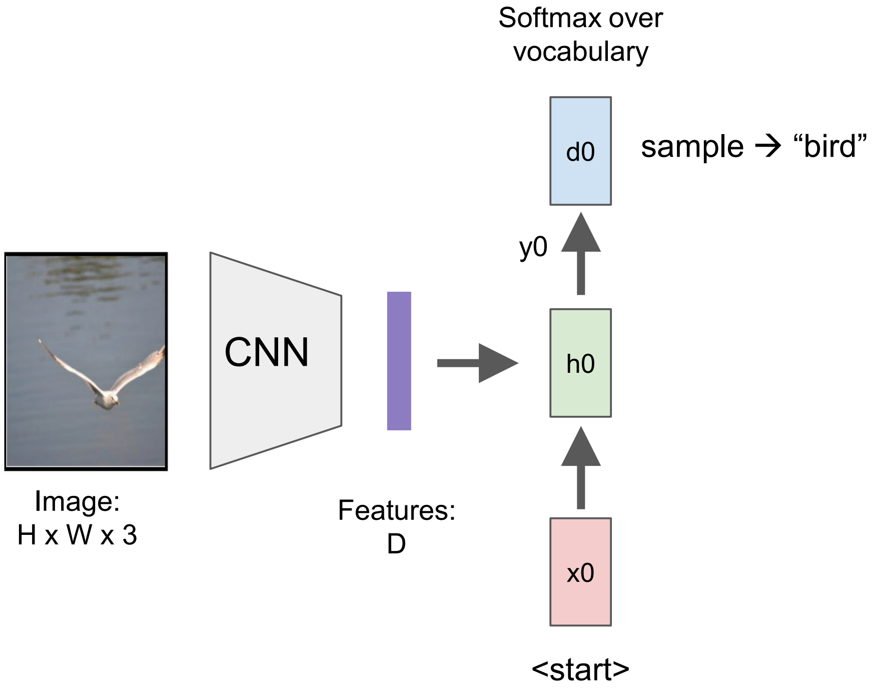


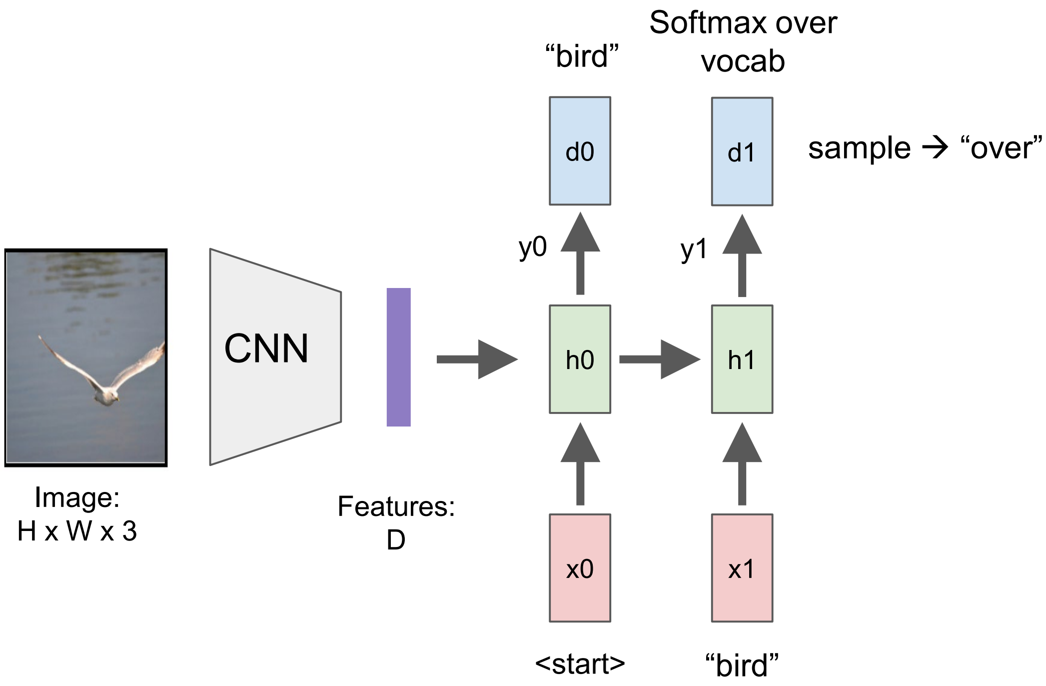
Image:
H x W x 3

Features:
D

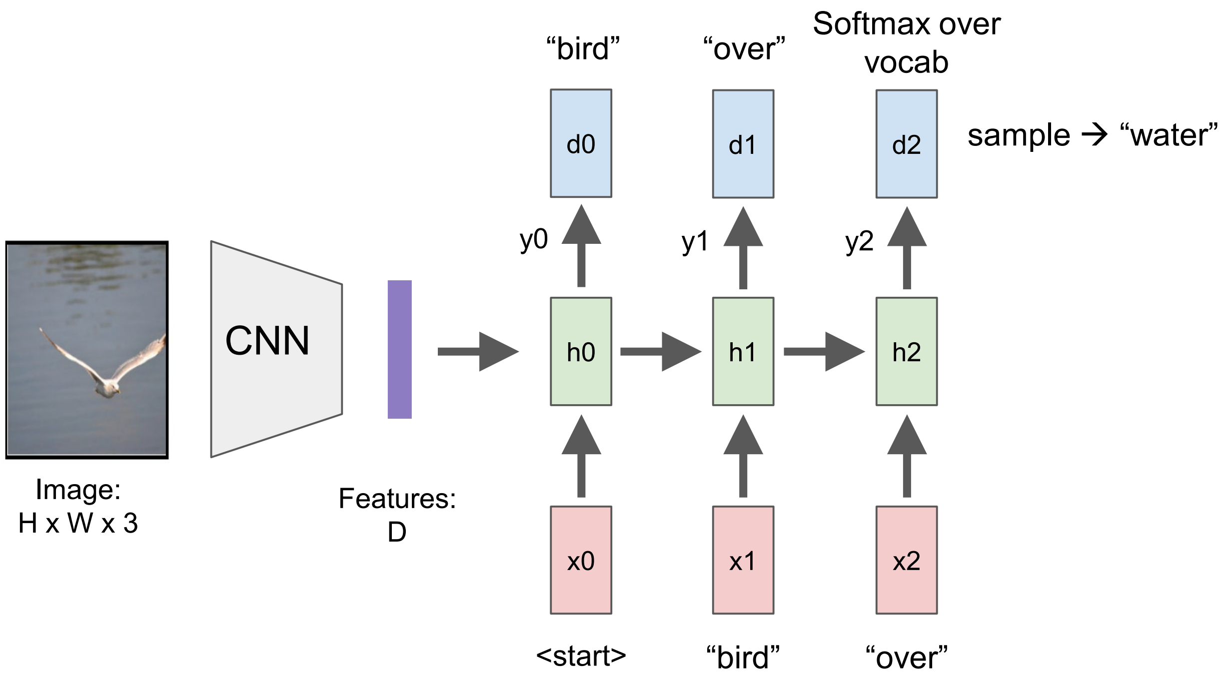
RNN for Captioning



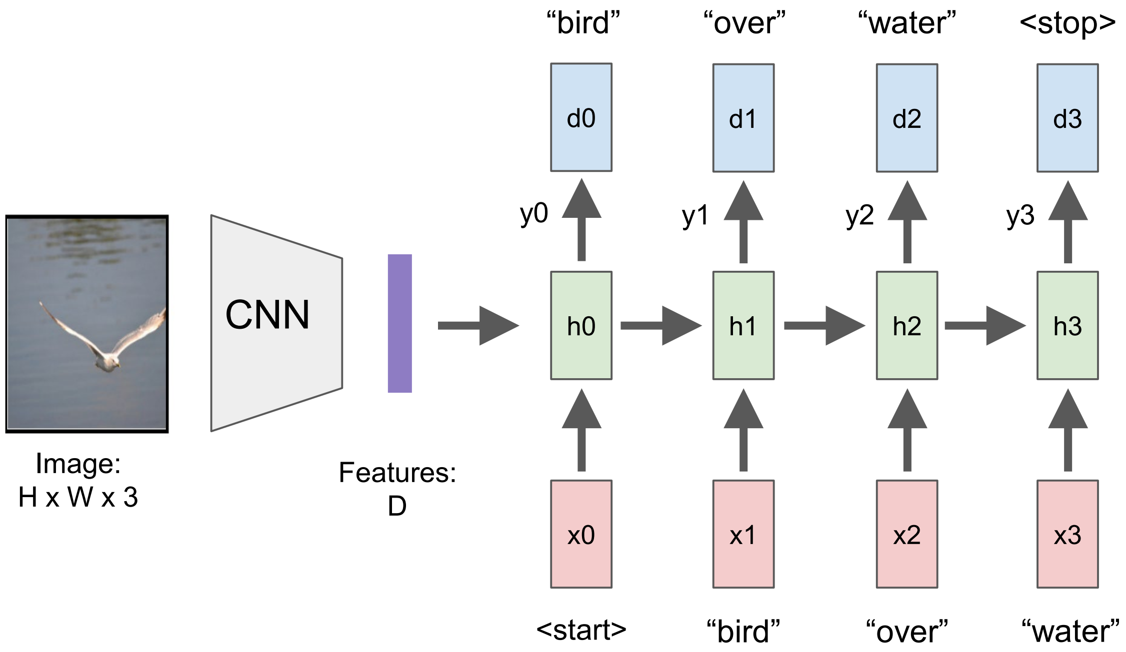
RNN for Captioning



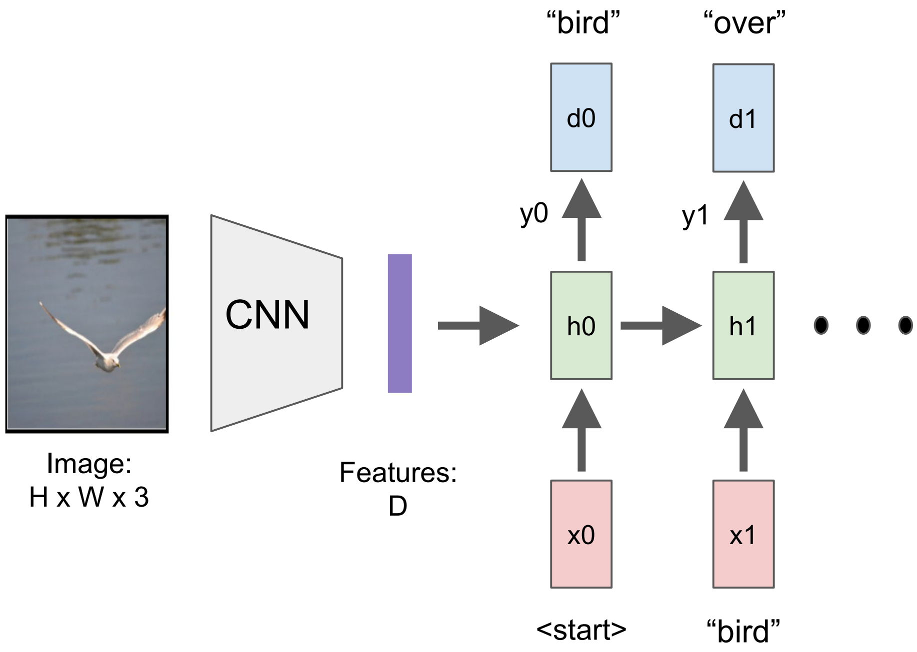
RNN for Captioning



RNN for Captioning

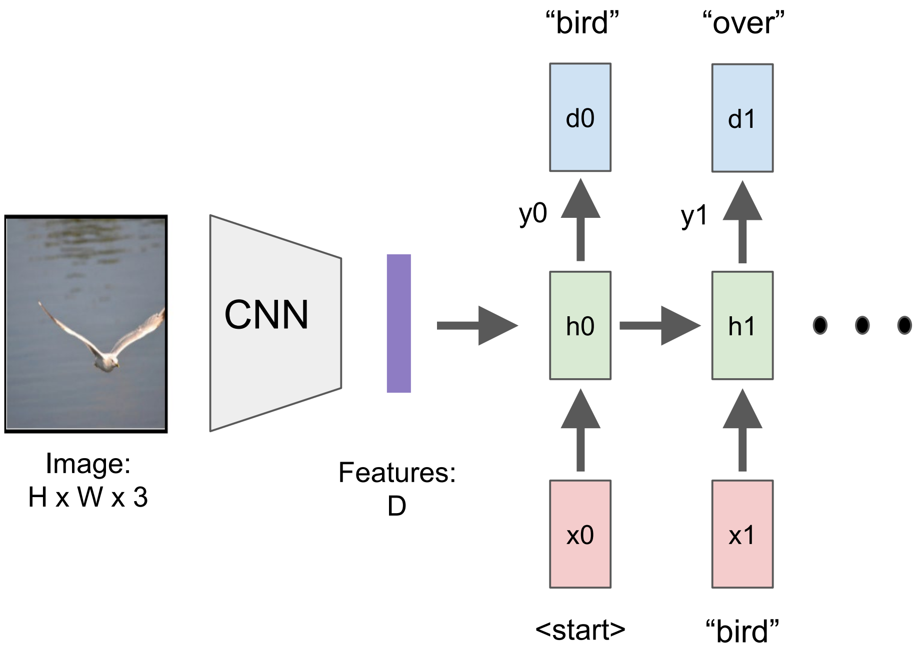


RNN for Captioning



RNN only looks at whole image, once

RNN for Captioning



RNN only looks at whole image, once

What if the RNN looked at different parts of the image at each time (word position)?

RNN for Captioning

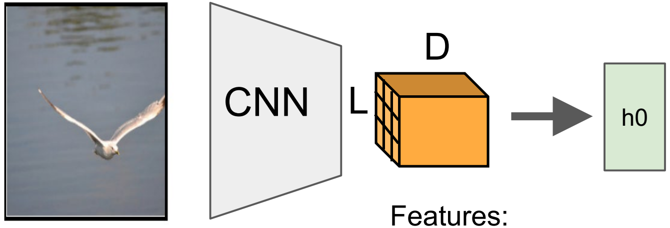
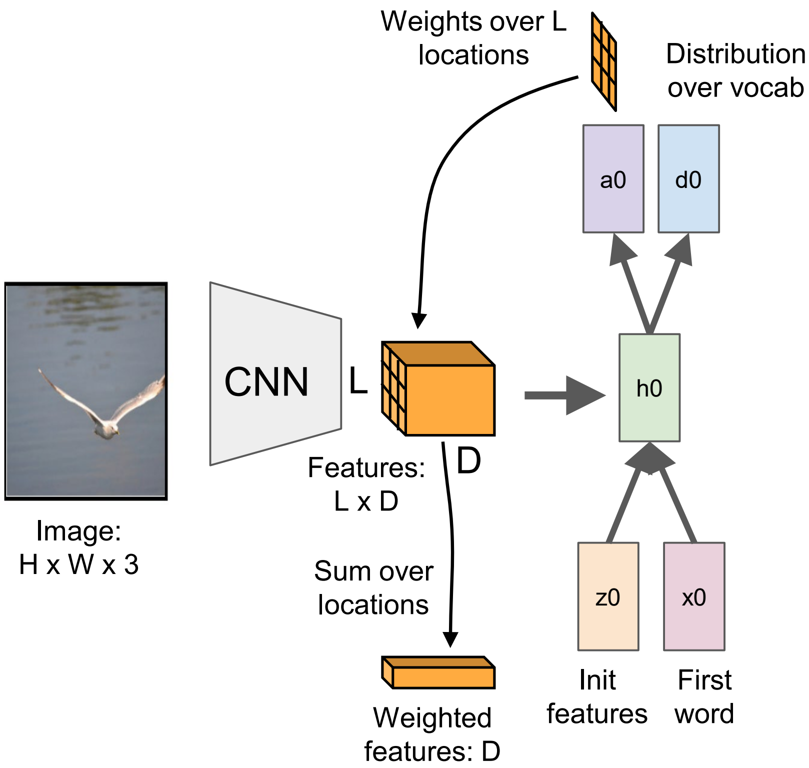


Image:
H x W x 3

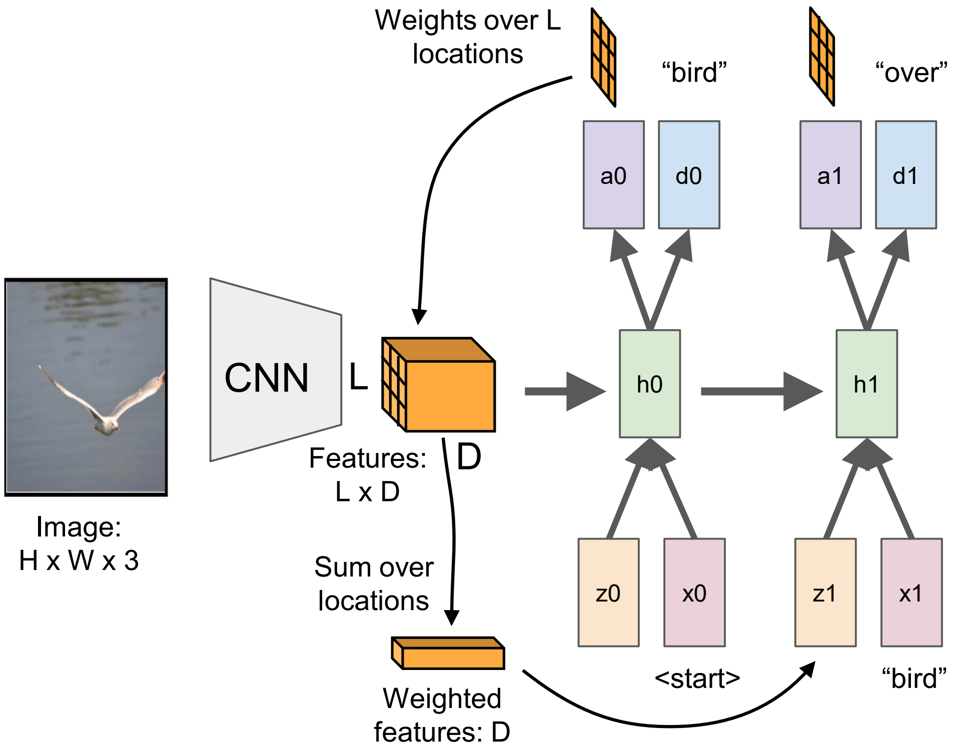
Features:
L x D

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

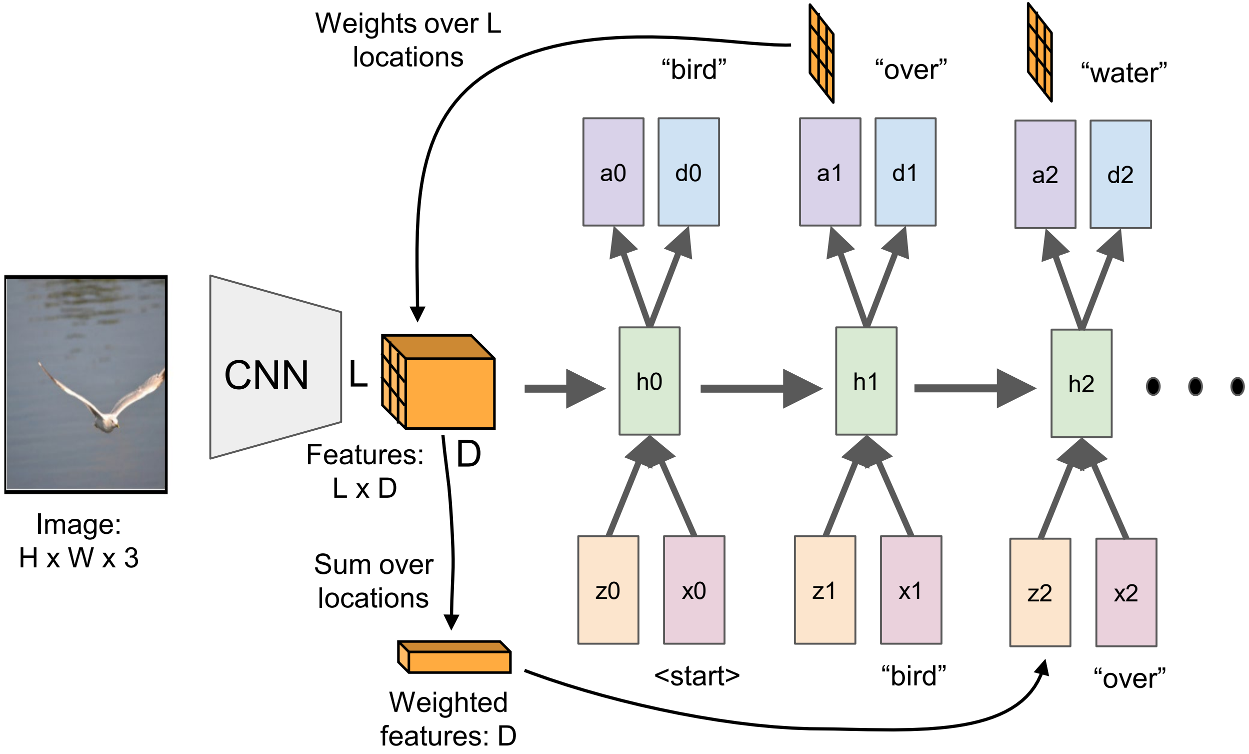
RNN for Captioning



RNN for Captioning



RNN for Captioning



Soft vs. Hard Attention

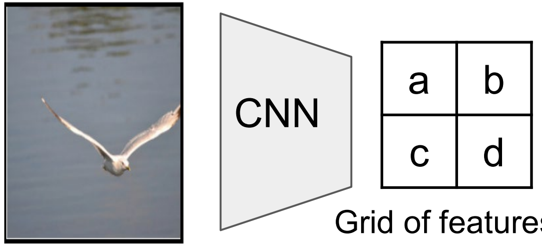
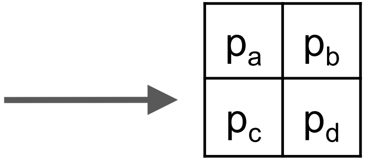


Image:
H x W x 3

Grid of features
(Each D-dimensional)

From
RNN:

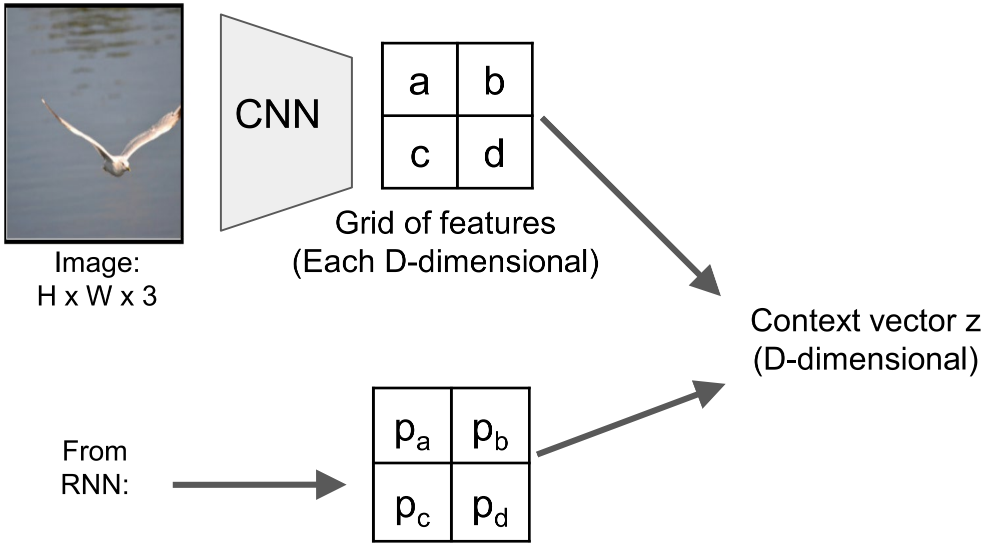


Distribution over grid locations

$p_a + p_b + p_c + p_c = 1$

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

Soft vs. Hard Attention



Distribution over grid locations

$p_a + p_b + p_c + p_c = 1$

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

Soft vs. Hard Attention

Image:
H x W x 3

CNN

a	b
c	d

Grid of features
(Each D-dimensional)

From
RNN:

p_a	p_b
p_c	p_d

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

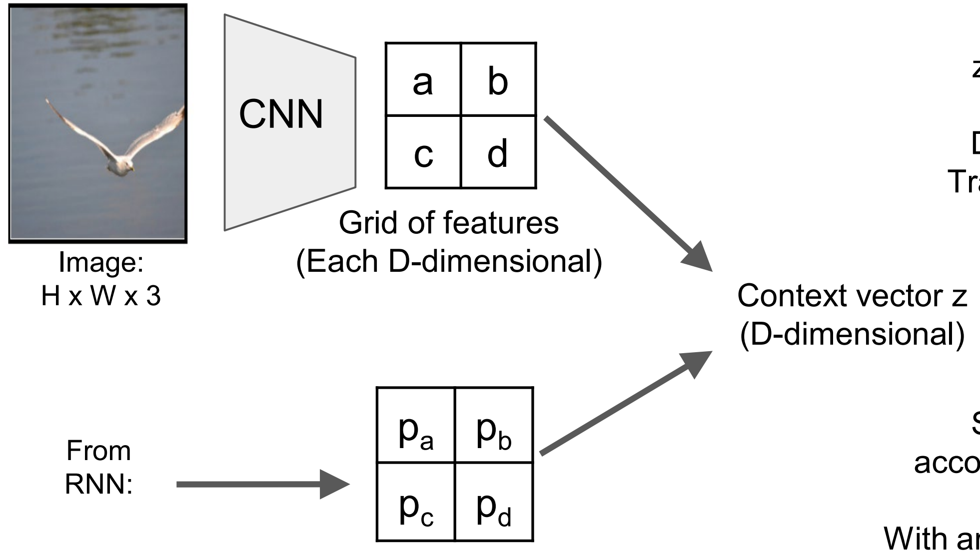
Context vector z
(D-dimensional)

Soft attention:
Sum 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

Soft vs. Hard Attention



Distribution over grid locations

$p_a + p_b + p_c + p_c = 1$

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

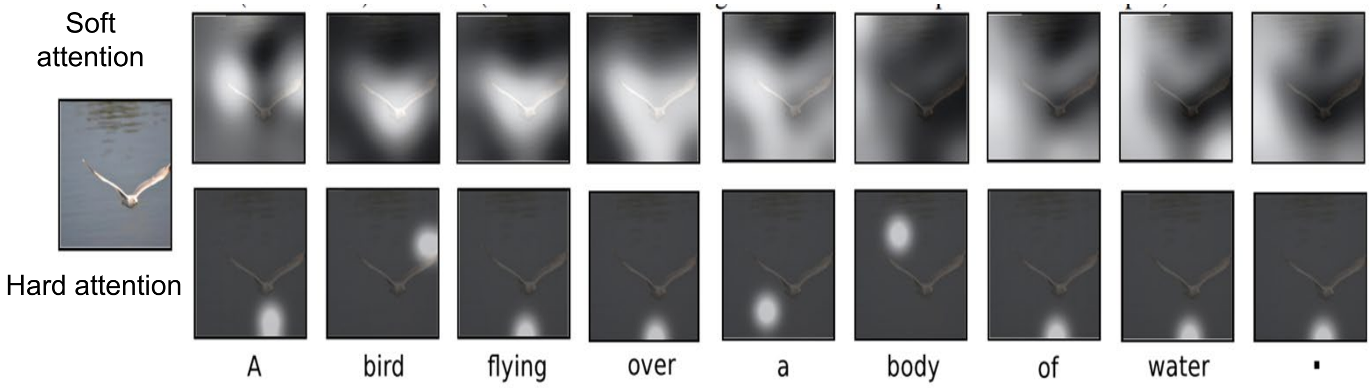
Soft attention:
Sum ALL locations
$z = p_a a + p_b b + p_c c + p_d d$

Derivative dz/dp is nice!
Train with gradient descent

Hard attention:
Sample ONE location
according to p, z = that vector

With argmax, dz/dp is zero almost everywhere …
Can't use gradient descent;
need reinforcement learning

Soft Attention for Captioning



Soft Attention for 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.

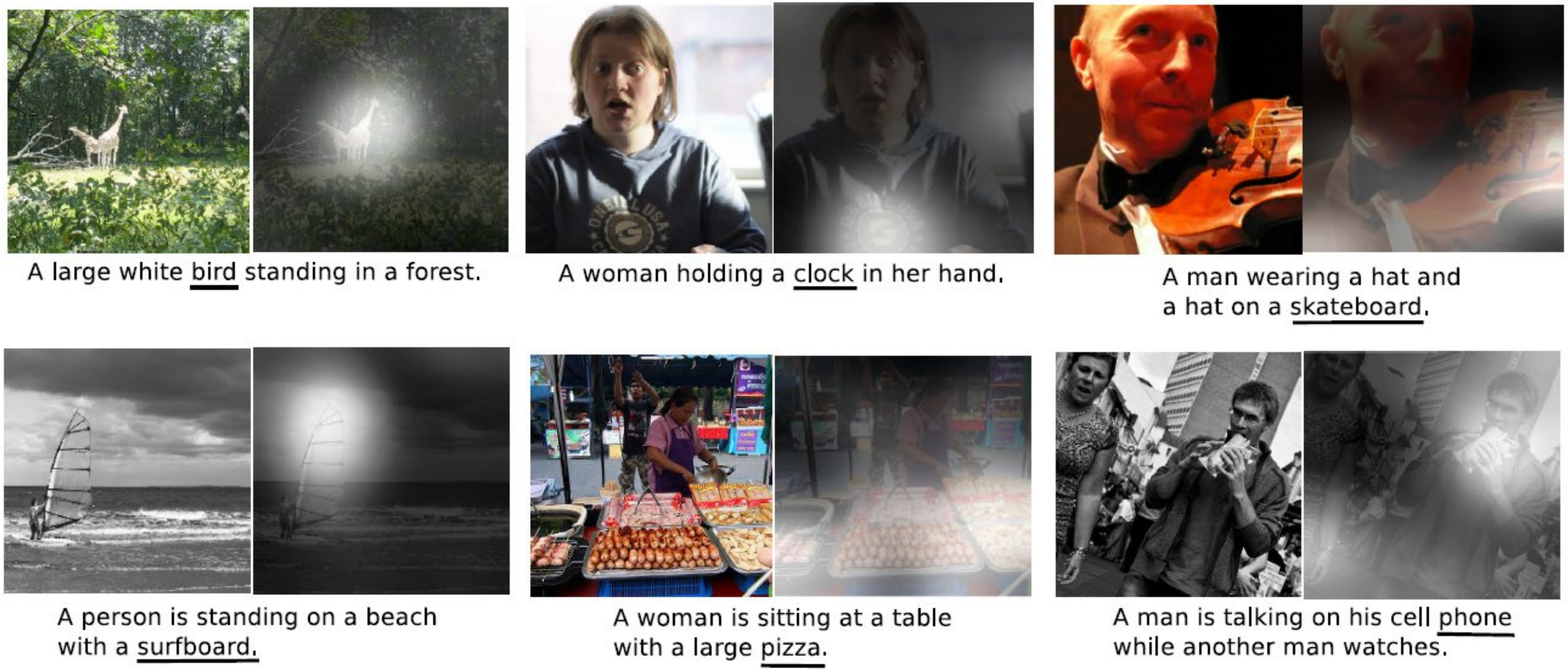
A group of people sitting on a boat in the water.

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

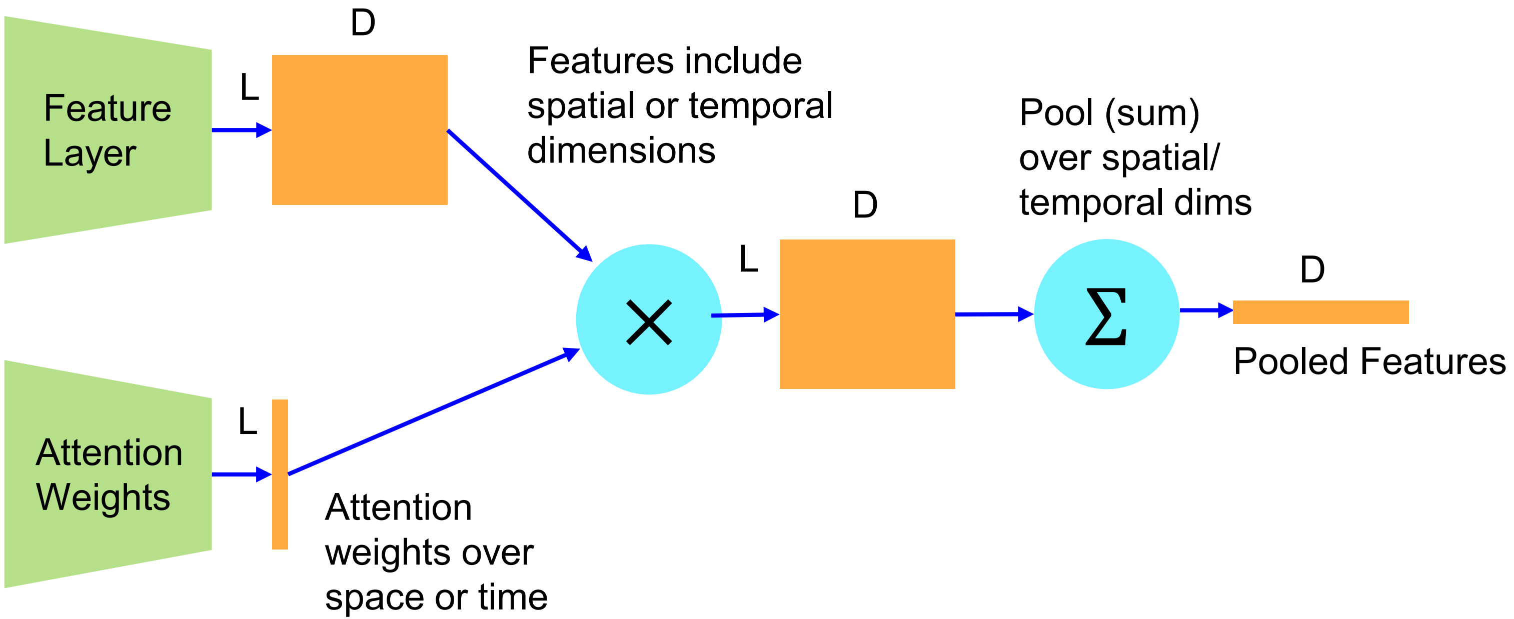
Soft Attention for Diagnosis

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.



Attention Mechanics

Typically, soft attention involves a feature layer, a weight predictor, and (optionally) pooling:



Gradients

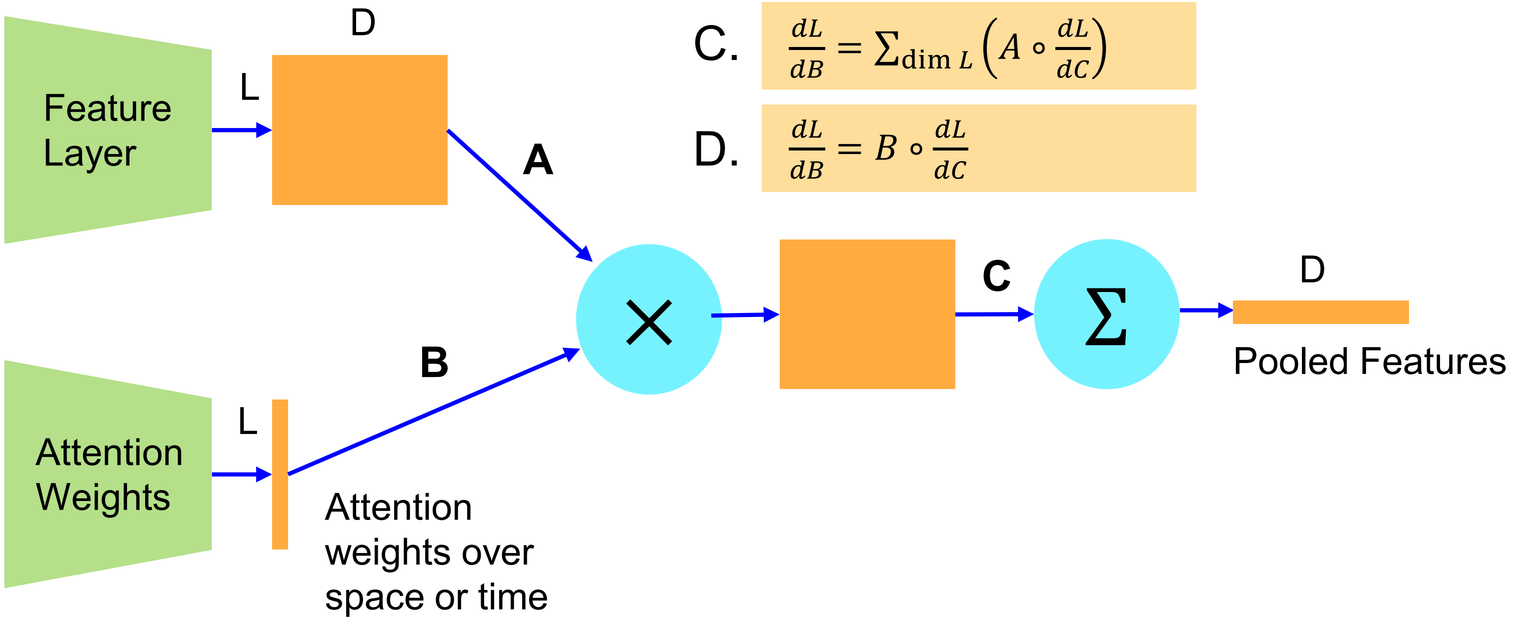
The attention gradient $\frac{dL}{dB}$ is given by

A. $\frac{dL}{dB} = A \circ \frac{dL}{dC}$

B. $\frac{dL}{dB} = \sum_{\dim D}\left(A \circ \frac{dL}{dC}\right)$

C. $\frac{dL}{dB} = \sum_{\dim L}\left(A \circ \frac{dL}{dC}\right)$

D. $\frac{dL}{dB} = B \circ \frac{dL}{dC}$

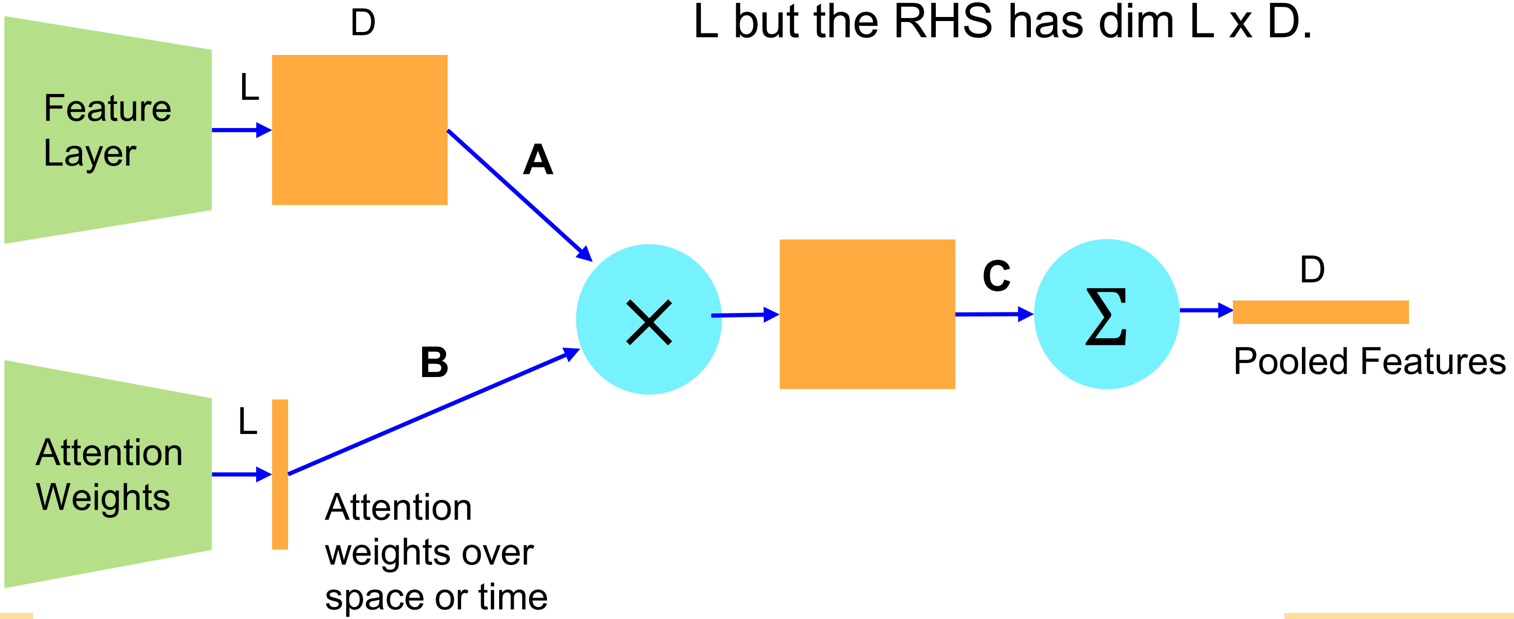


Oops!

The attention gradient $\frac{dL}{dB}$ is given by

A. $\frac{dL}{dB} = A \circ \frac{dL}{dC}$

Dimensions mismatch: $\frac{dL}{dB}$ is dim L but the RHS has dim L x D.



Try Again

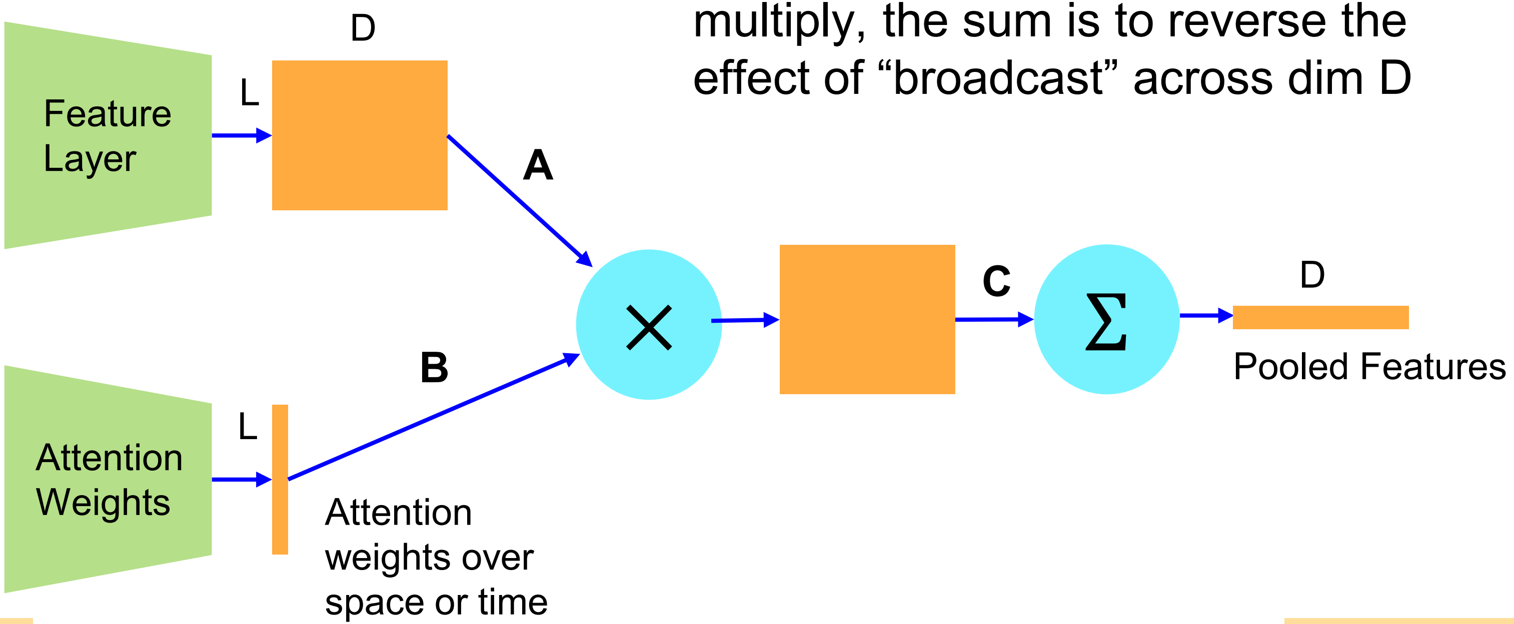
Continue

Correct!

The attention gradient $\frac{dL}{dB}$ is given by

B. $\frac{dL}{dB} = \sum_{\dim D} \left(A \circ \frac{dL}{dC} \right)$

$\left(A \circ \frac{dL}{dC} \right)$ is the backprop for the multiply, the sum is to reverse the effect of “broadcast” across dim D



Try Again

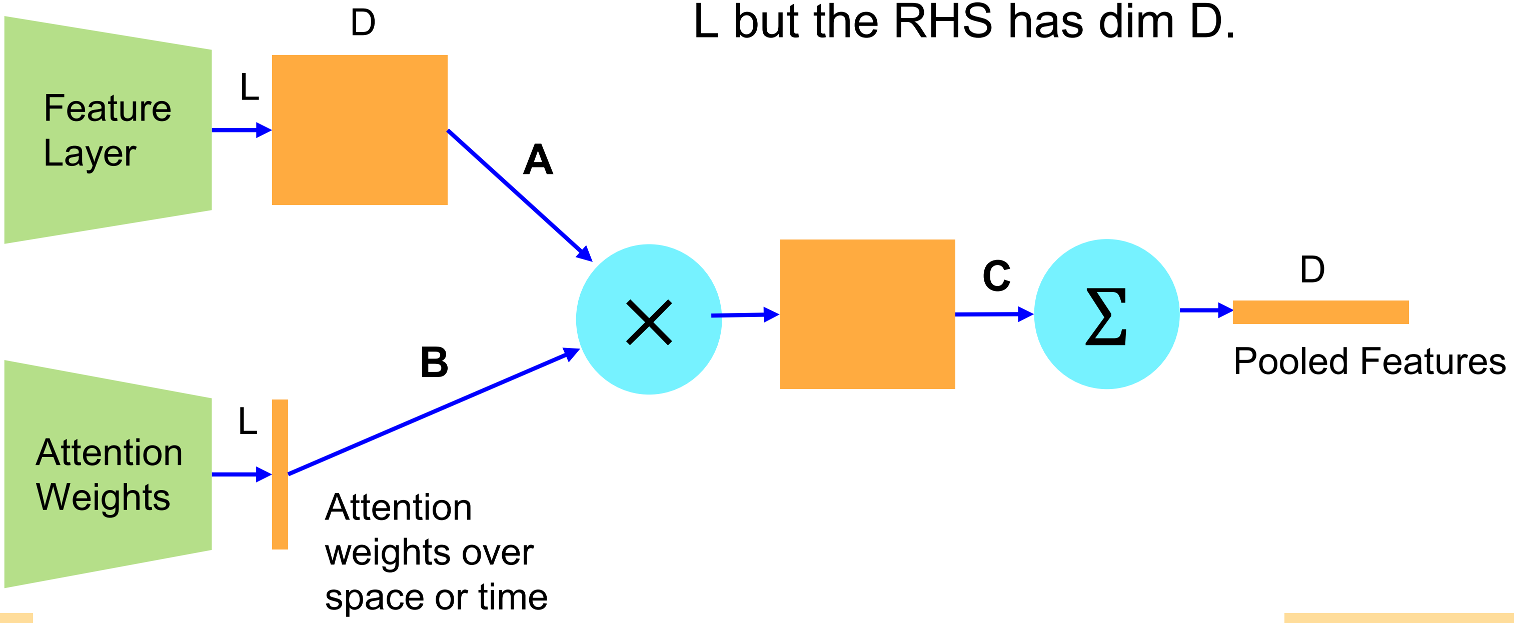
Continue

Oops!

The attention gradient $\frac{dL}{dB}$ is given by

C. $\frac{dL}{dB} = \sum_{\dim L} \left(A \circ \frac{dL}{dC} \right)$

Dimensions mismatch: $\frac{dL}{dB}$ is dim L but the RHS has dim D.



Try Again

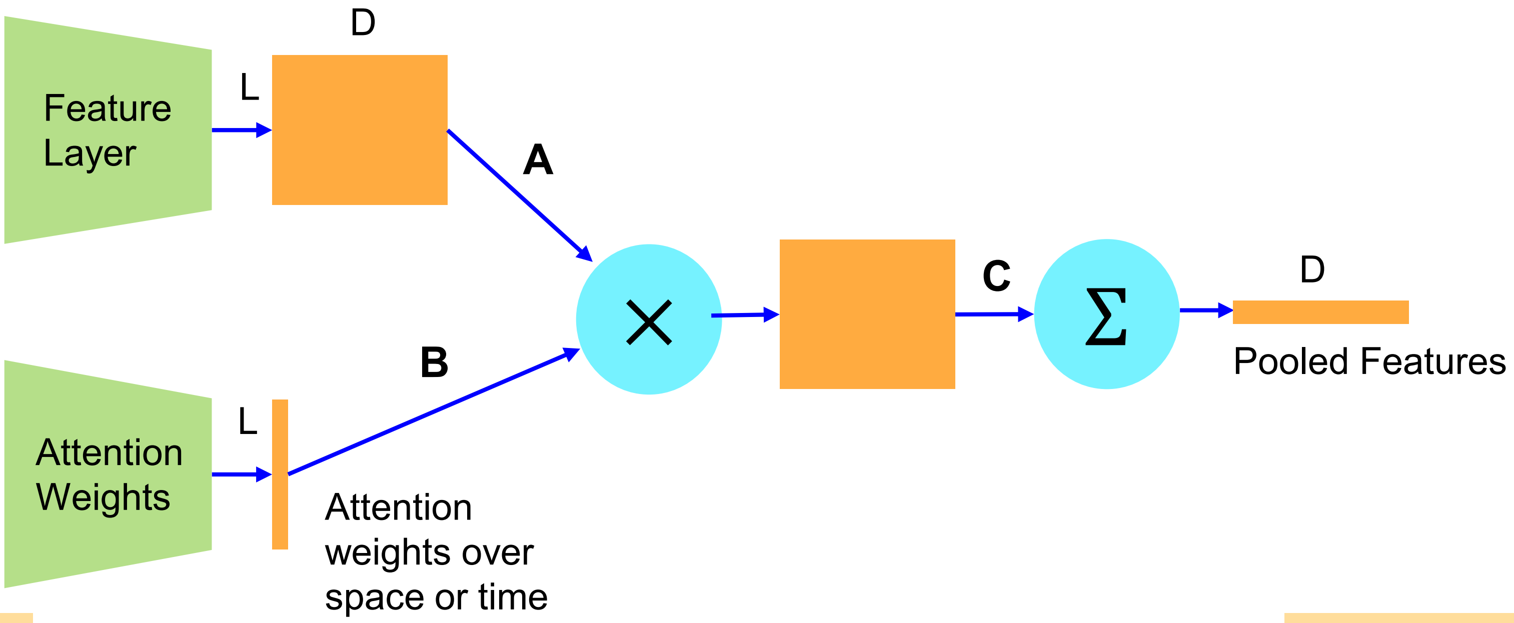
Continue

Oops!

The attention gradient $\frac{dL}{dB}$ is given by

D. $\frac{dL}{dB} = B \circ \frac{dL}{dC}$

Just wrong…

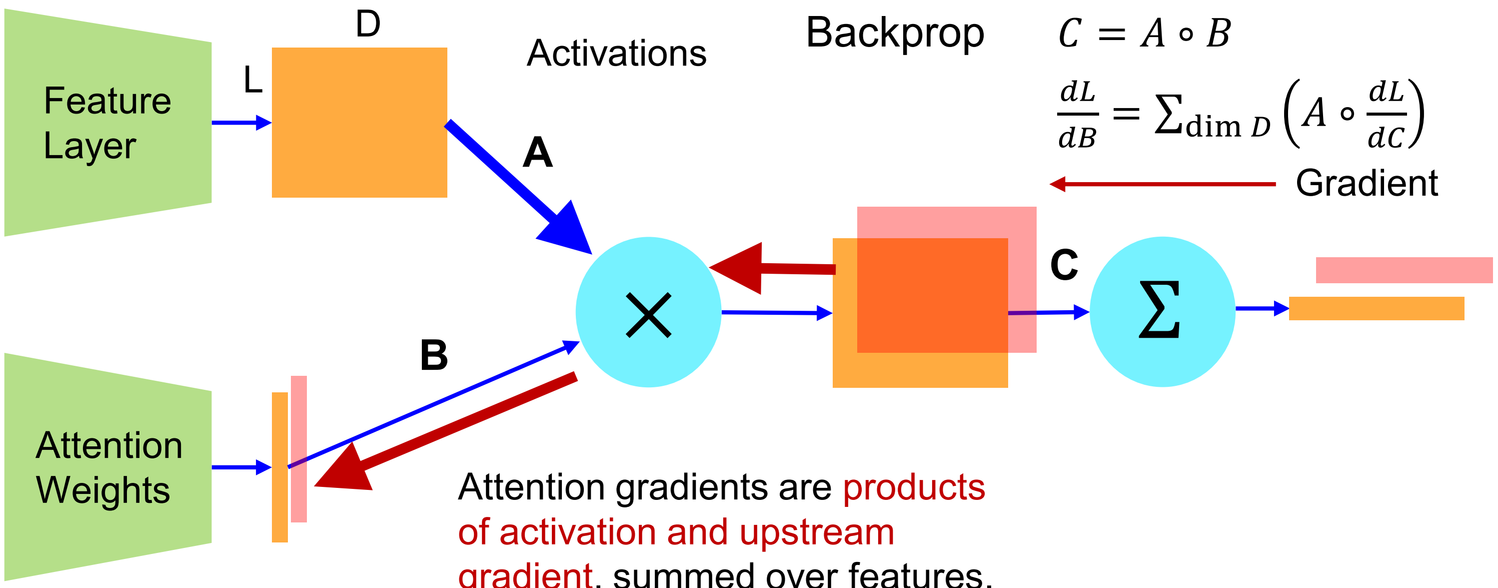


Try Again

Continue

Attention Mechanics: Salience

During training, the attention layers receives gradients which are the product of the upstream gradient and the feature layer activations (salience).



Attention gradients are products of activation and upstream gradient, summed over features.
This is a ***salience*** (importance) signal.

Attention and LSTMs

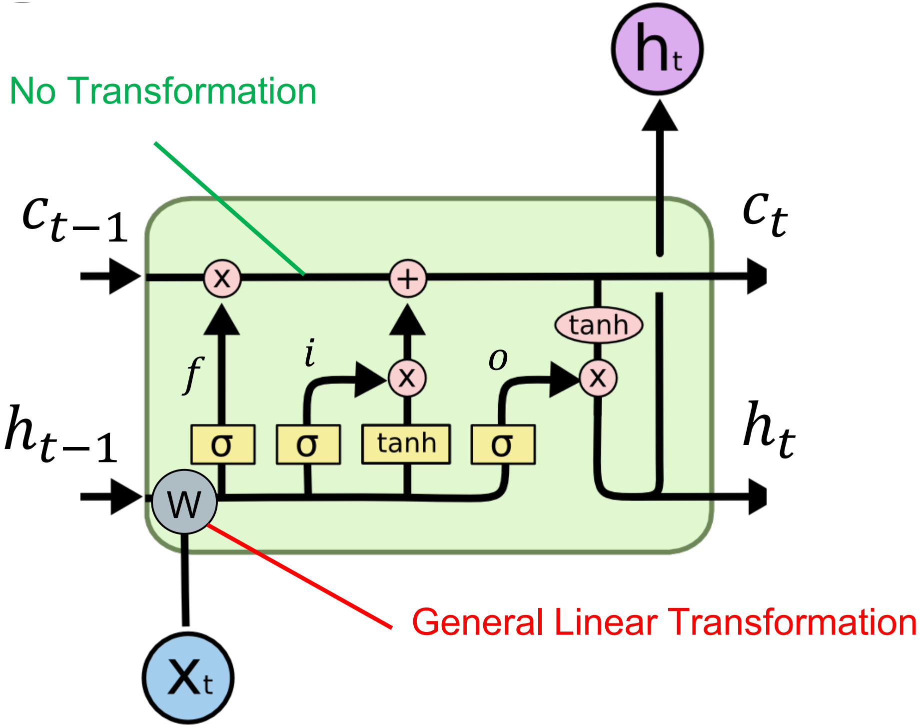
We saw something similar in LSTMs: i, f, o nodes learn to weight features.

They receive a salience gradient during training.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$$c_t^l = f \odot c_{t-1}^l + i \odot g$$

$$h_t^l = o \odot \tanh(c_t^l)$$



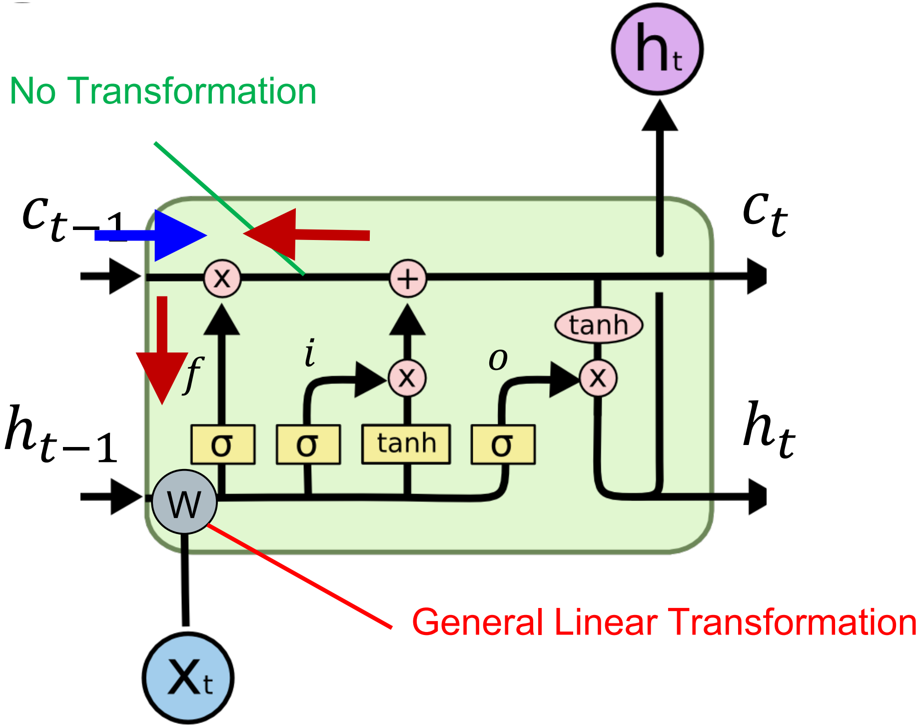
Attention and LSTMs

We saw something similar in LSTMs: i, f, o nodes learn to weight features.

They receive a salience gradient during training.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$$c_t^l = f \odot c_{t-1}^l + i \odot g$$
$$h_t^l = o \odot \tanh(c_t^l)$$



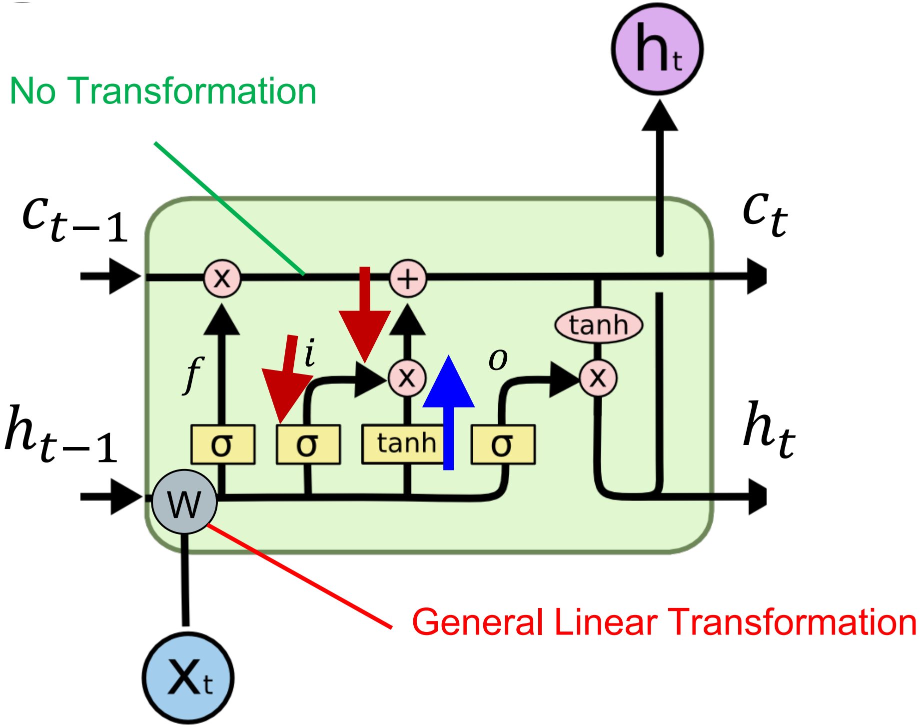
Attention and LSTMs

We saw something similar in LSTMs: i, f, o nodes learn to weight features.

They receive a salience gradient during training.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$$c_t^l = f \odot c_{t-1}^l + i \odot g$$
$$h_t^l = o \odot \tanh(c_t^l)$$



Attention and LSTMs

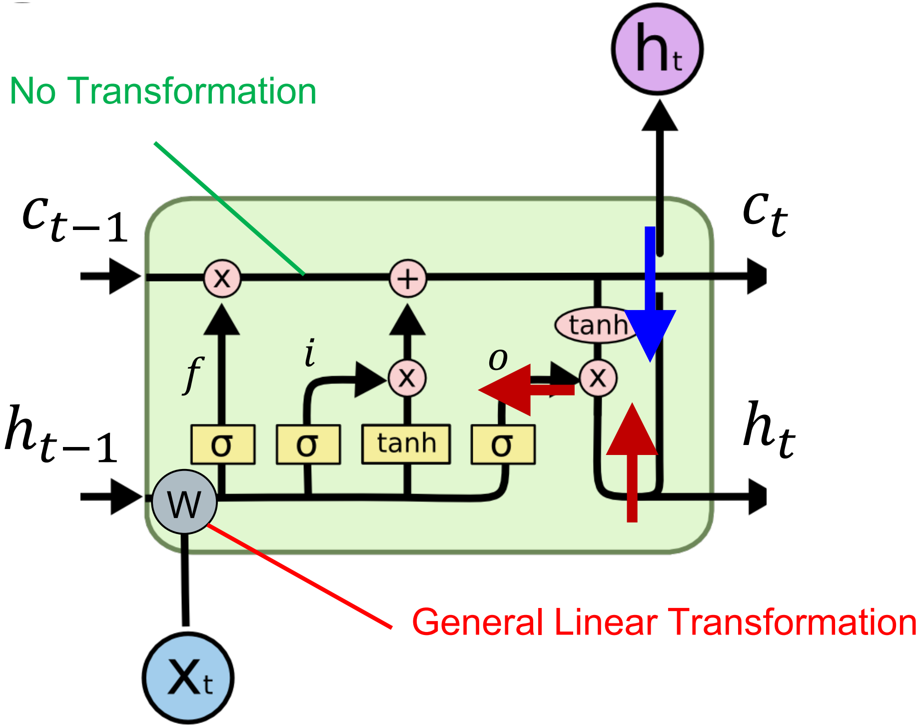
We saw something similar in LSTMs: i, f, o nodes learn to weight features.

They receive a salience gradient during training.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$$c_t^l = f \odot c_{t-1}^l + i \odot g$$

$$h_t^l = o \odot \tanh(c_t^l)$$



Attention as Explanation

Deep Network behavior is generally inscrutable.

Deep Networks do not model data like classical ML models.

Activations don’t have obvious meaning (mostly).

Attention maps are explanations of net behavior because they identify the influential parts of the input stream.

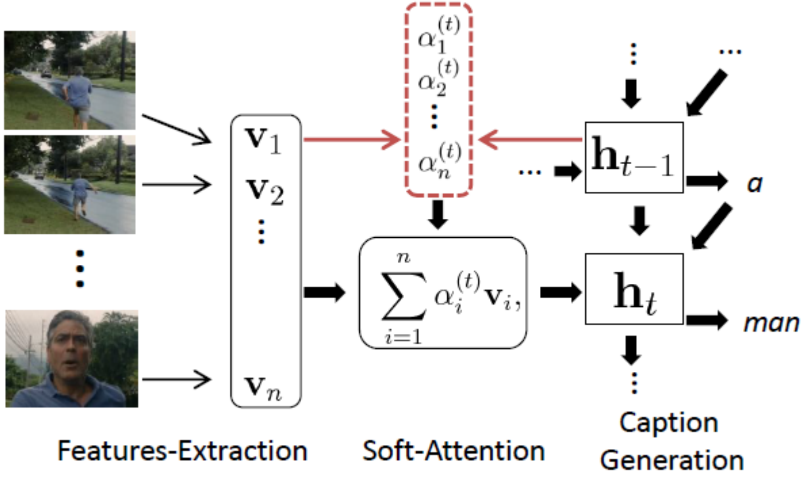
Soft Attention for Video

"Describing Videos by Exploiting Temporal Structure," Li Yao et al, arXiv 2015.

Soft Attention for Video

The attention model:

$\alpha_i^{(t)}$ is the attention weight to frame i at time t

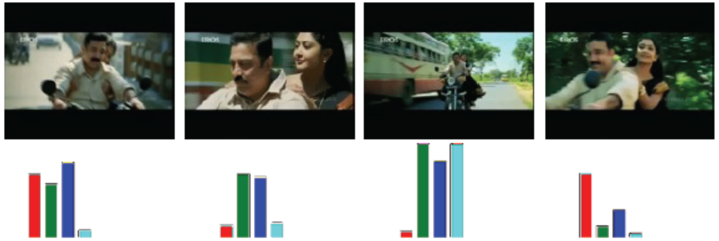


n is the number of video frames

Number of times steps is The caption length

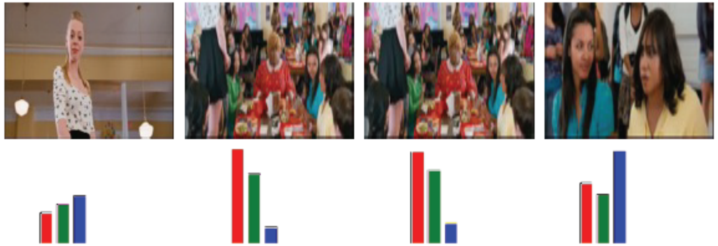
“Describing Videos by Exploiting Temporal Structure,” Li Yao et al, arXiv 2015.

Examples



+Local+Global: A man and a woman are talking on the road

Ref: A man and a woman ride a motorcycle



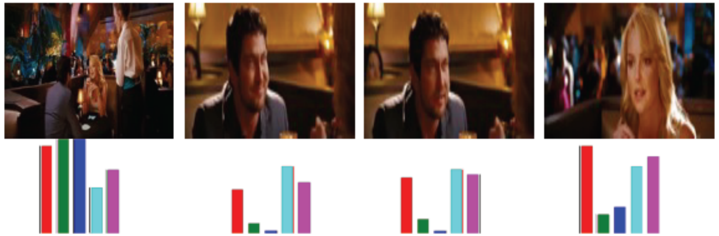
+Local+Global: the girl grins at him

Ref: SOMEONE and SOMEONE swap a look

***+Local+Global*:** Someone is frying a fish in a pot

+Local: Someone is frying something
+Global: The person is cooking
Basic: A man cooking its kitchen

Ref: A woman is frying food



***+Local+Global*:** as SOMEONE sits on the table, SOMEONE shifts his gaze to SOMEONE

***+Local*:** with a smile SOMEONE arrives
***+Global*:** SOMEONE sits at a table
***Basic*:** now, SOMEONE grins

***Ref*:** SOMEONE gaze at SOMEONE

Soft Attention for Video

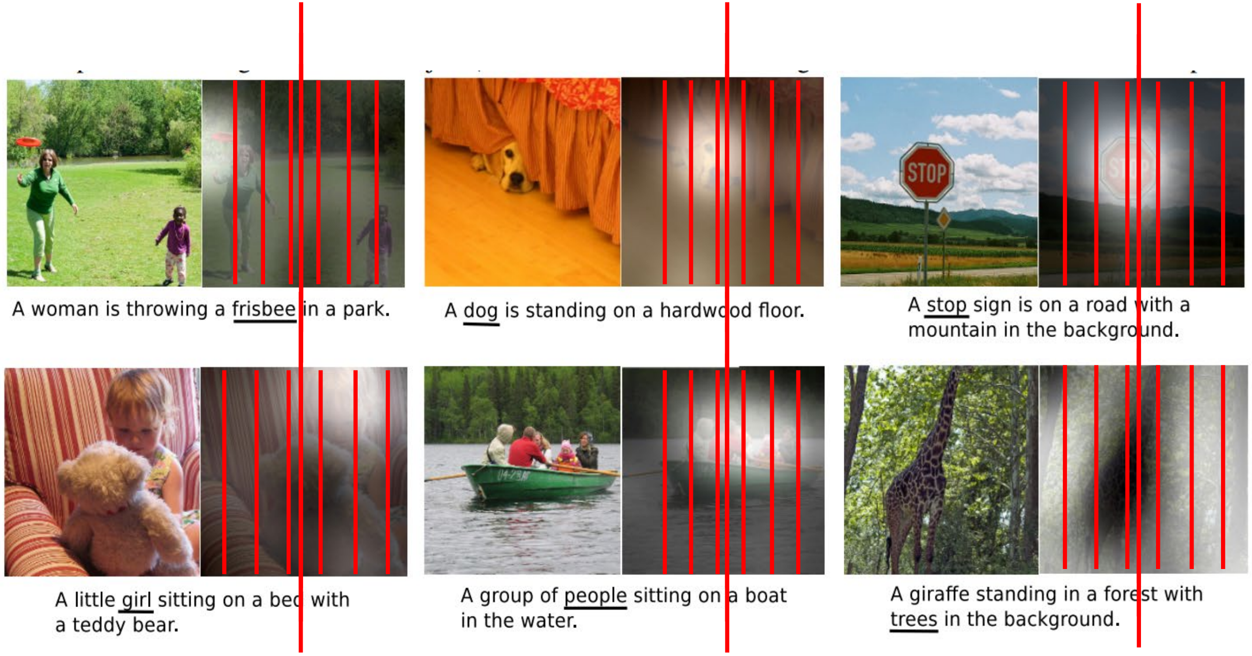
Table 1. Performance of different variants of the model on the Youtube2Text and DVS datasets.

	Youtube2Text				DVS			
Model	BLEU	METEOR	CIDEr	Perplexity	BLEU	METEOR	CIDEr	Perplexity
Enc-Dec (Basic)	0.3869	0.2868	0.4478	33.09	0.003	0.044	0.044	88.28
+ Local (3-D CNN)	0.3875	0.2832	0.5087	33.42	0.004	0.051	0.050	84.41
+ Global (Temporal Attention)	0.4028	0.2900	0.4801	27.89	0.003	0.040	0.047	66.63
+ Local + Global	**0.4192**	**0.2960**	**0.5167**	**27.55**	**0.007**	**0.057**	**0.061**	**65.44**
Venugopalan *et al.* [41]	0.3119	0.2687	-	-	-	-	-	-
+ Extra Data (Flickr30k, COCO)	0.3329	0.2907	-	-	-	-	-	-
Thomason *et al.* [37]	0.1368	0.2390	-	-	-	-	-	-

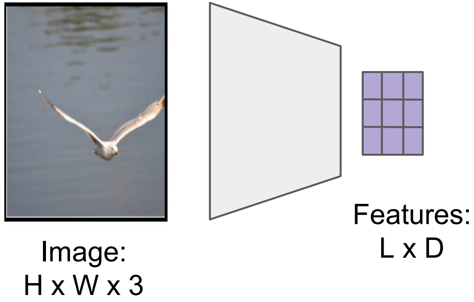
"Describing Videos by Exploiting Temporal Structure," Li Yao et al, arXiv 2015.

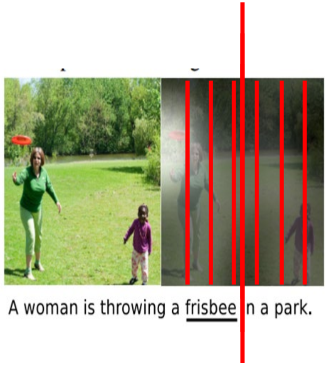
Soft Attention for Captioning

Attention constrained to fixed grid!



Attending to arbitrary regions?

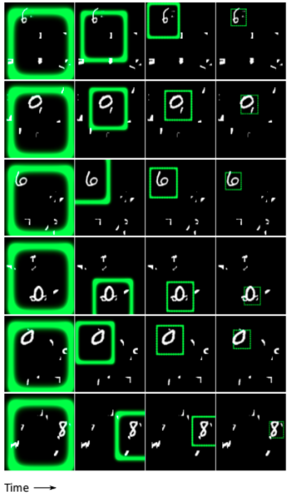




Attention mechanism from Show, Attend, and Tell only lets us softly attend to fixed grid positions … can we do better?

Attending to Arbitrary Regions: DRAW

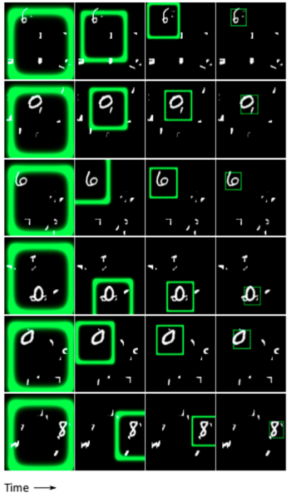
Classify images by attending to arbitrary regions of the *input*



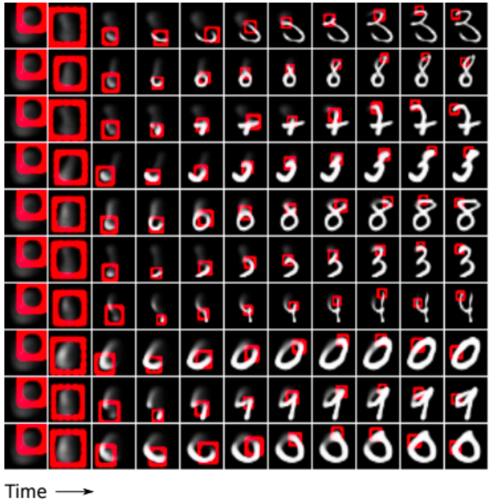
Gregor et al, “DRAW: A Recurrent Neural Network For Image Generation”, ICML 2015

Attending to Arbitrary Regions: DRAW

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



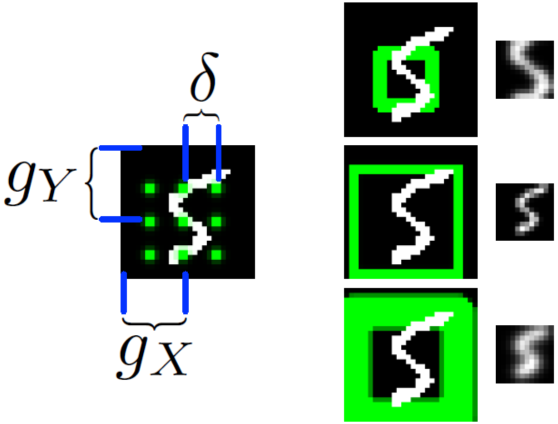
Generate images by attending to arbitrary regions of the *output*



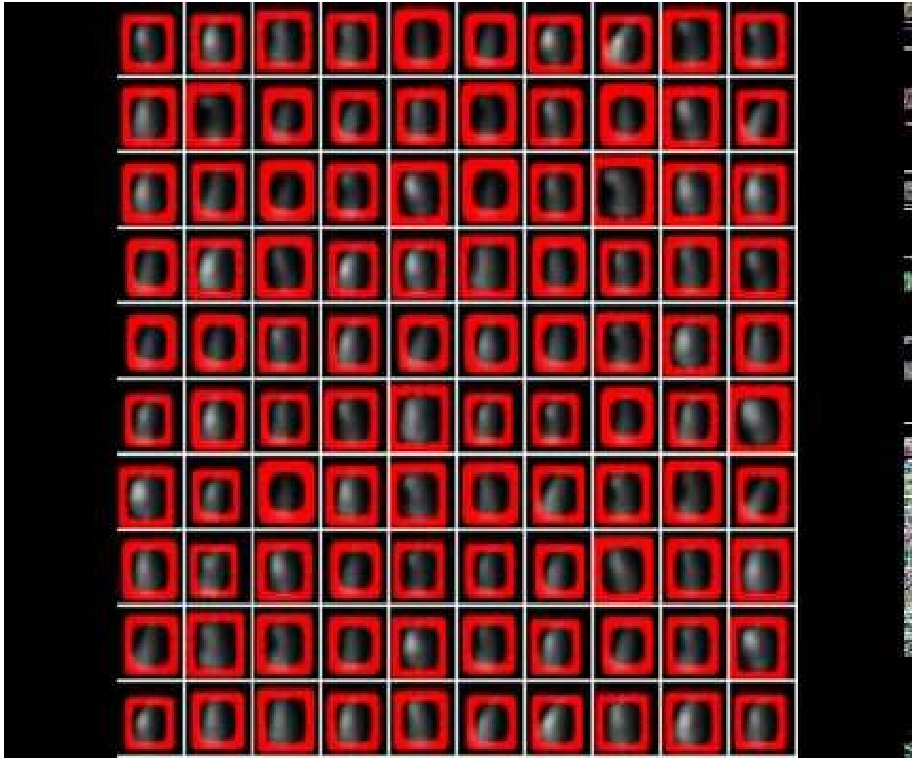
Gregor et al, “DRAW: A Recurrent Neural Network For Image Generation”, ICML 2015

Attending to Arbitrary Regions: DRAW

Attention is a parametric distribution: both location and scale can vary:



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



Attention Takeaways

Performance:

Attention models can ***improve accuracy*** and ***reduce computation*** at the same time.

Salience:

Attention models learn to predict salience, i.e. to emphasize relevant input data across space or time.

Attention Takeaways

Explainability:

Attention models encode explanations.

Both locus and trajectory help understand what’s going on.

Hard vs. Soft:

Soft models are easier to train, hard models require reinforcement learning.