Attention-based architectures: recurrent encoder-decoder

Generative and Deep Learning (GDL)

Davide Bacciu (davide.bacciu@unipi.it)

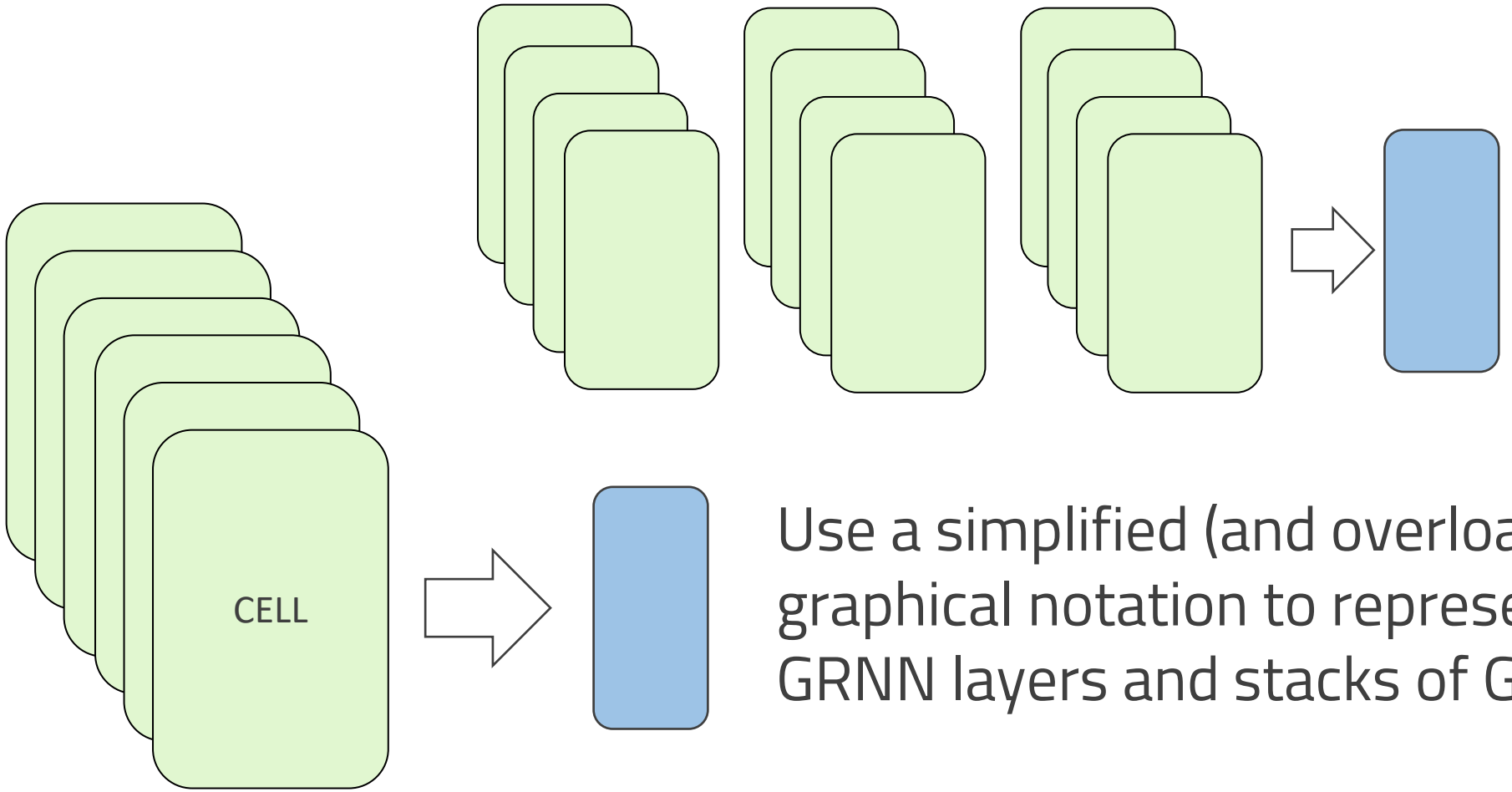
UNIVERSITÀ DI PISA

Objectives

- Advanced sequential tasks
  - Sequence-to-sequence
  - Dealing with compound output data
- Encoder-decoder recurrent architectures
- Cross-attention mechanism

Sequence-to-sequence

Graphical Notation for Compositionality



Use a simplified (and overloaded) graphical notation to represent GRNN layers and stacks of GRNN

Dealing with Compound Data

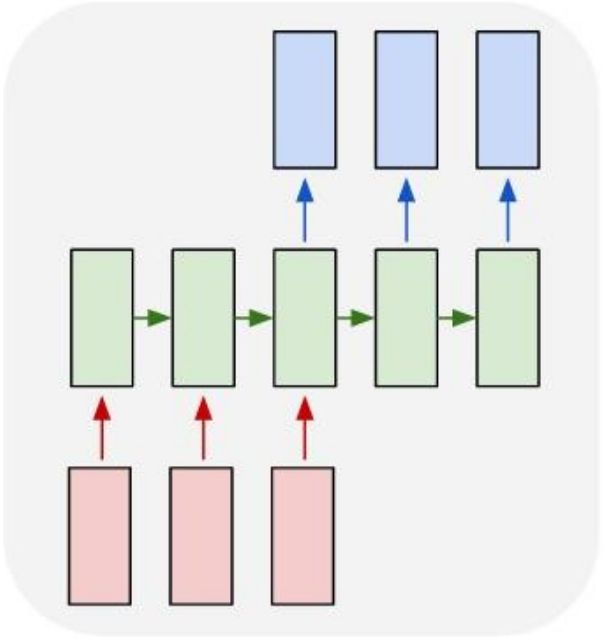
- GRNN are excellent to handle size/topology varying data in input
  - How can we handle size/topology varying outputs?
  - Sequence-to-sequence
- Structured data is compound information
  - Efficient processing needs the ability to focus on certain parts of such information
  - Cross-attention mechanism

Sequence Transduction

- Input and output are both sequences
- They may have different lengths
- Example: machine translation

How do we model the context here?

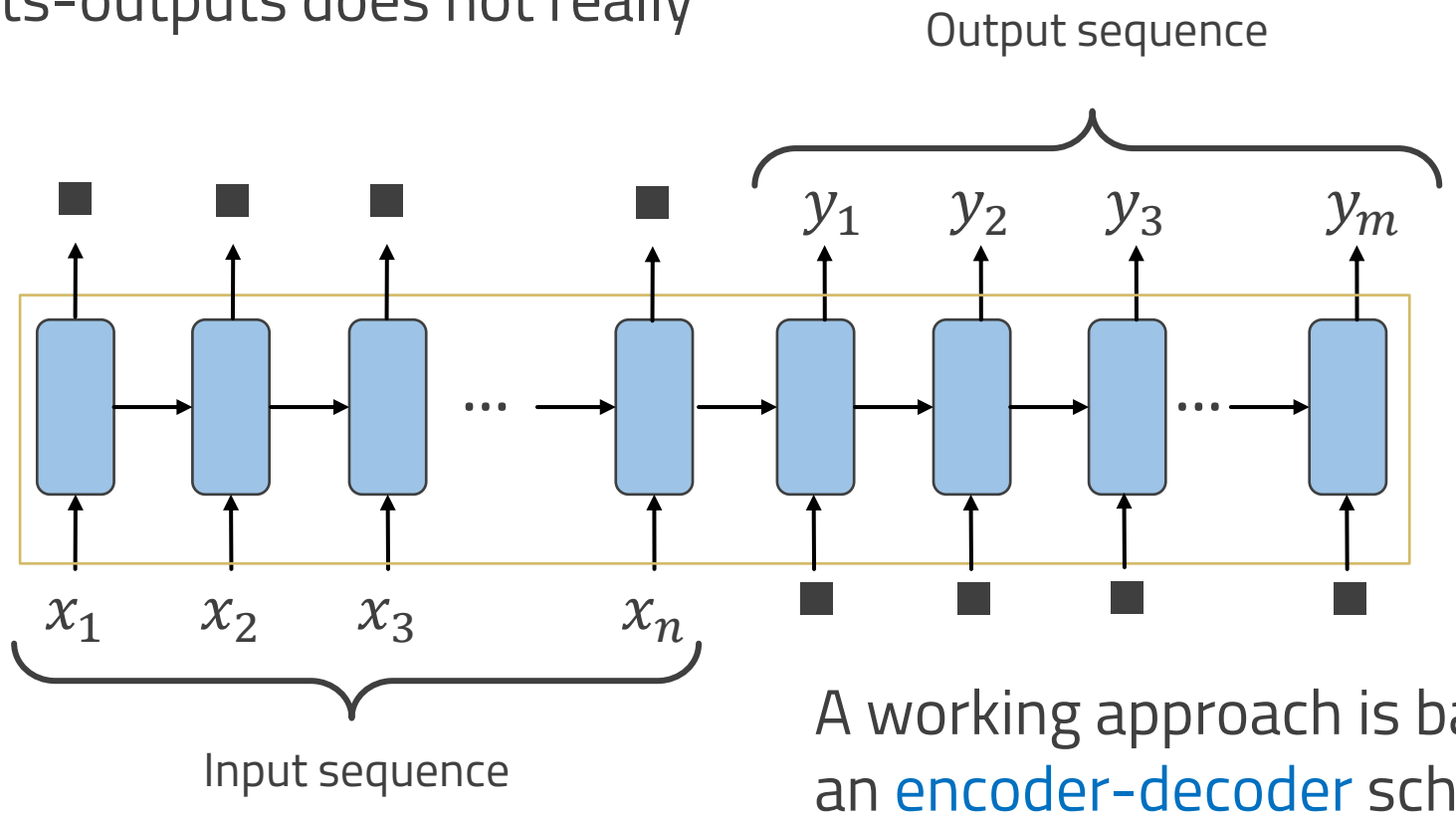
The cat is on the table → **Il gatto è sul tavolo**



sequence to sequence

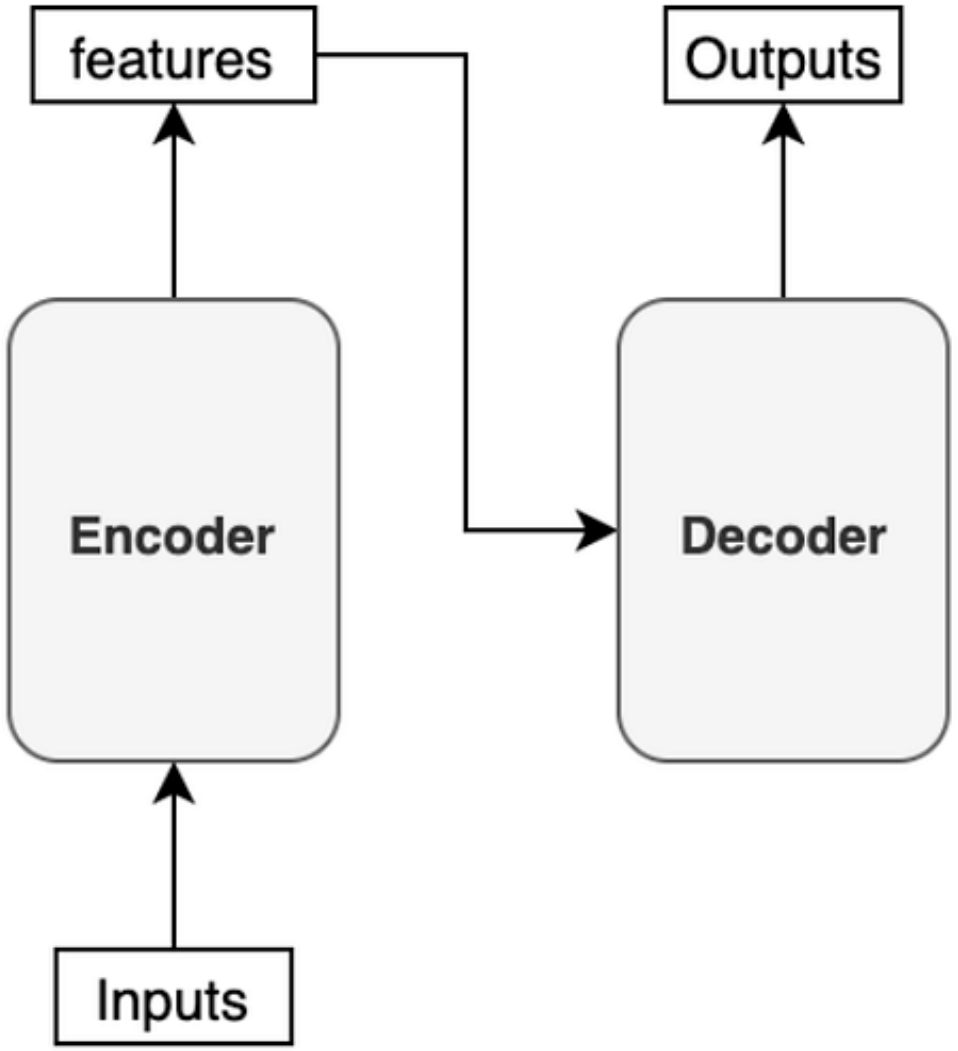
Learning to Output Variable Length Sequences

The idea of an unfolded RNN with blank inputs-outputs does not really work well



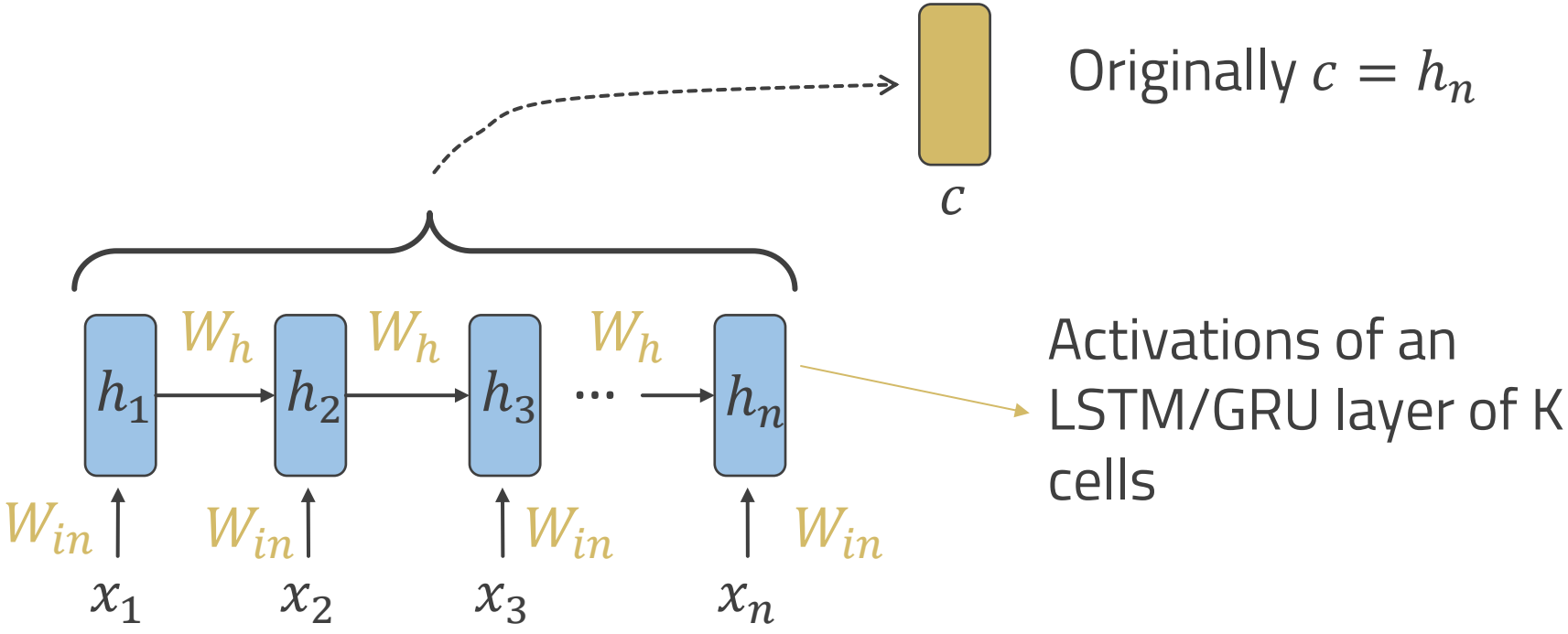
A working approach is based on an encoder-decoder scheme

Encoder-Decoder Architecture

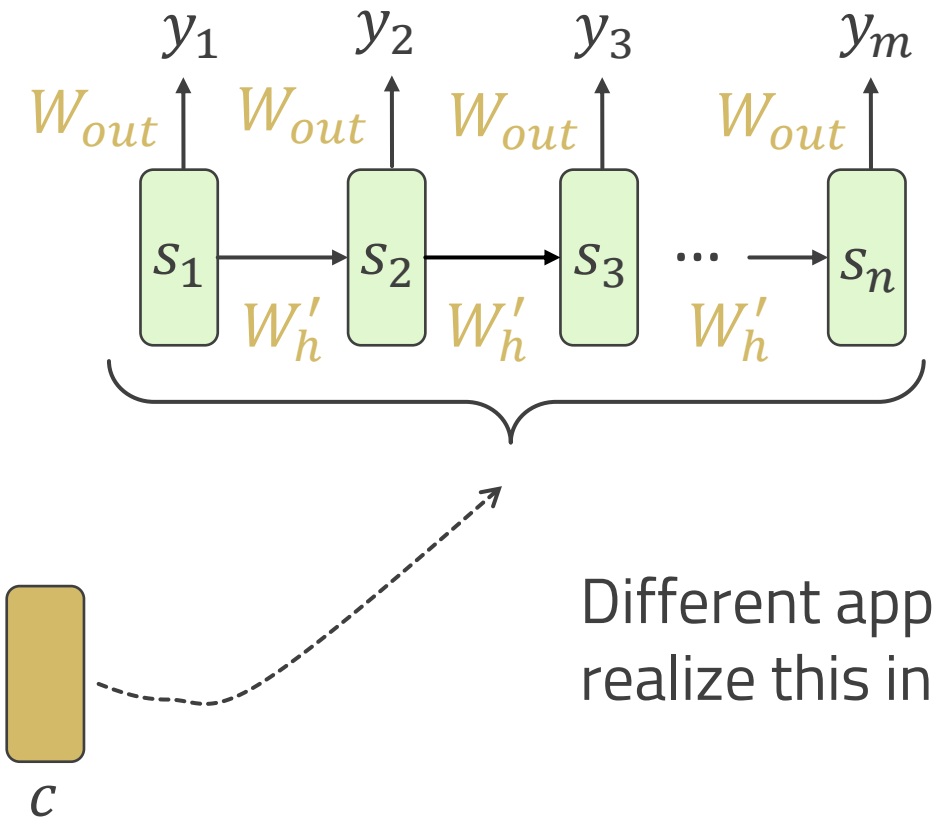


Encoder

Produce a compressed and fixed length representation c of all the input sequence $x_1, \dots, x_n$



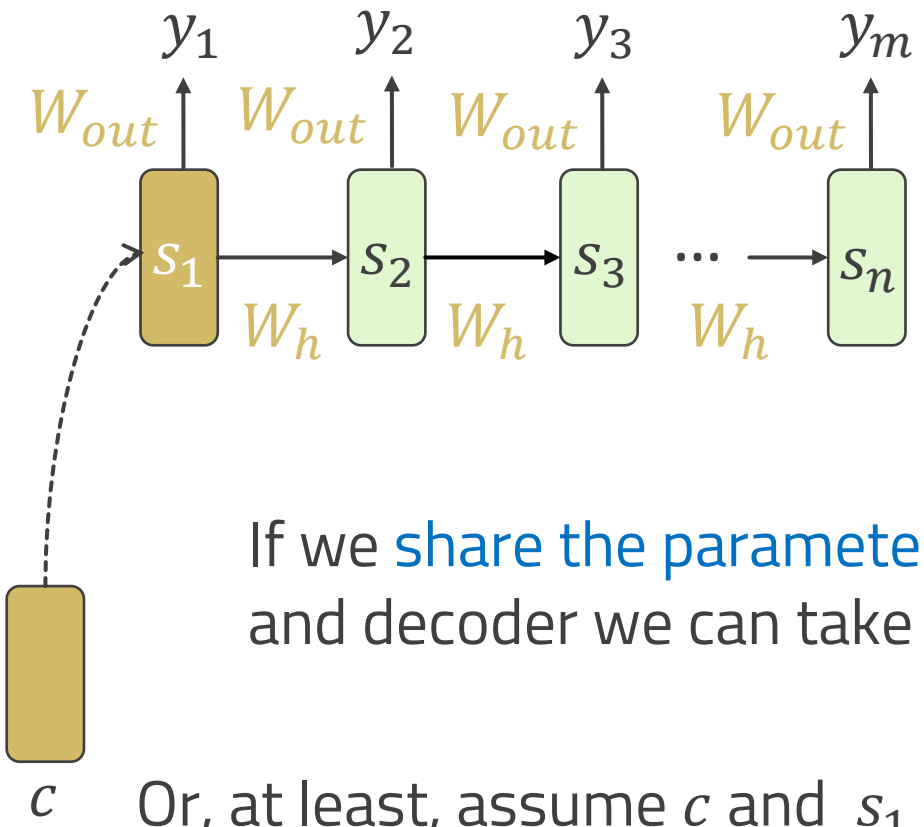
Decoder



A LSTM/GRU layer of K cells seeded by the context vector c

Different approaches to realize this in practice

Decoder

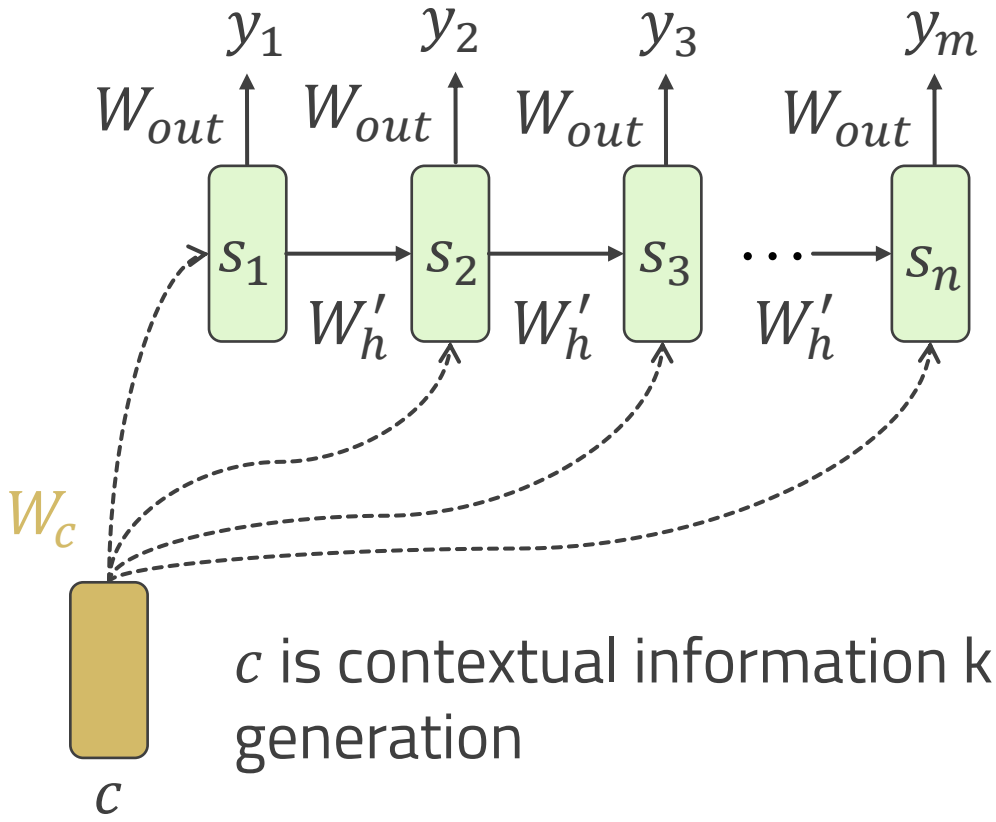


We risk to lose memory of c soon

If we share the parameters between encoder and decoder we can take $s_1 = c$

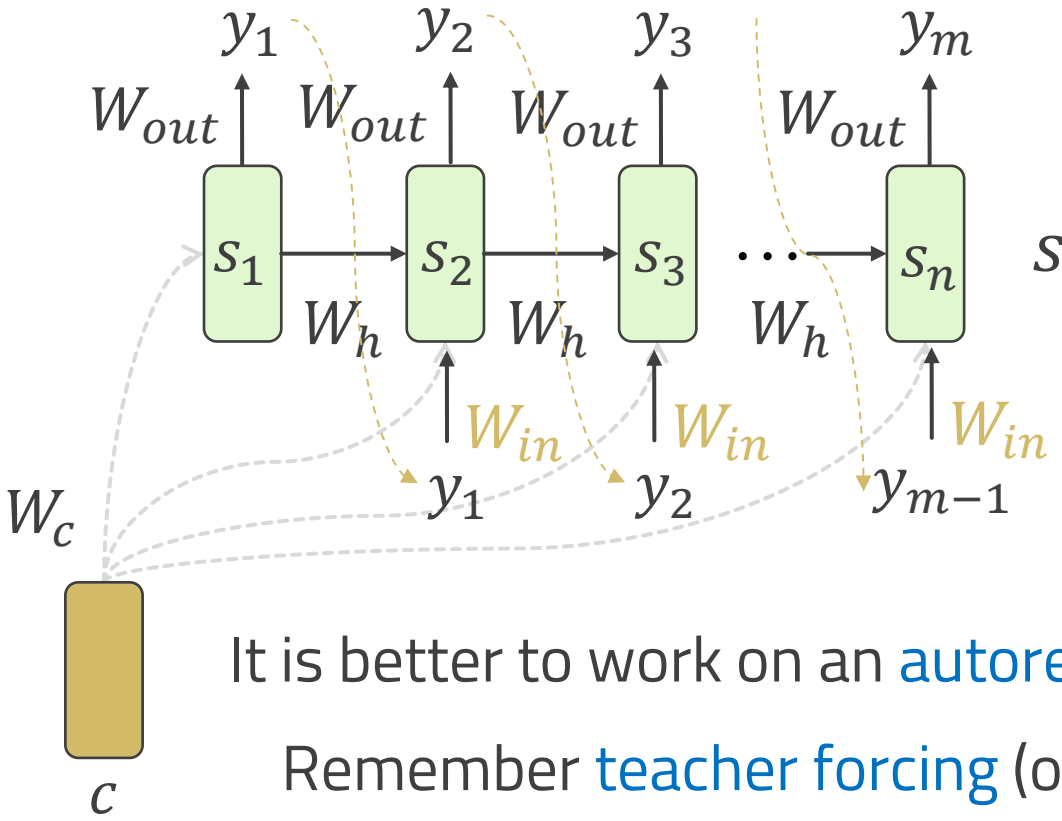
Or, at least, assume c and s_1 have compatible size

Decoder



c is contextual information kept throughout output generation

Decoder



$$s_i = f(c, s_{i-1}, y_{i-1})$$

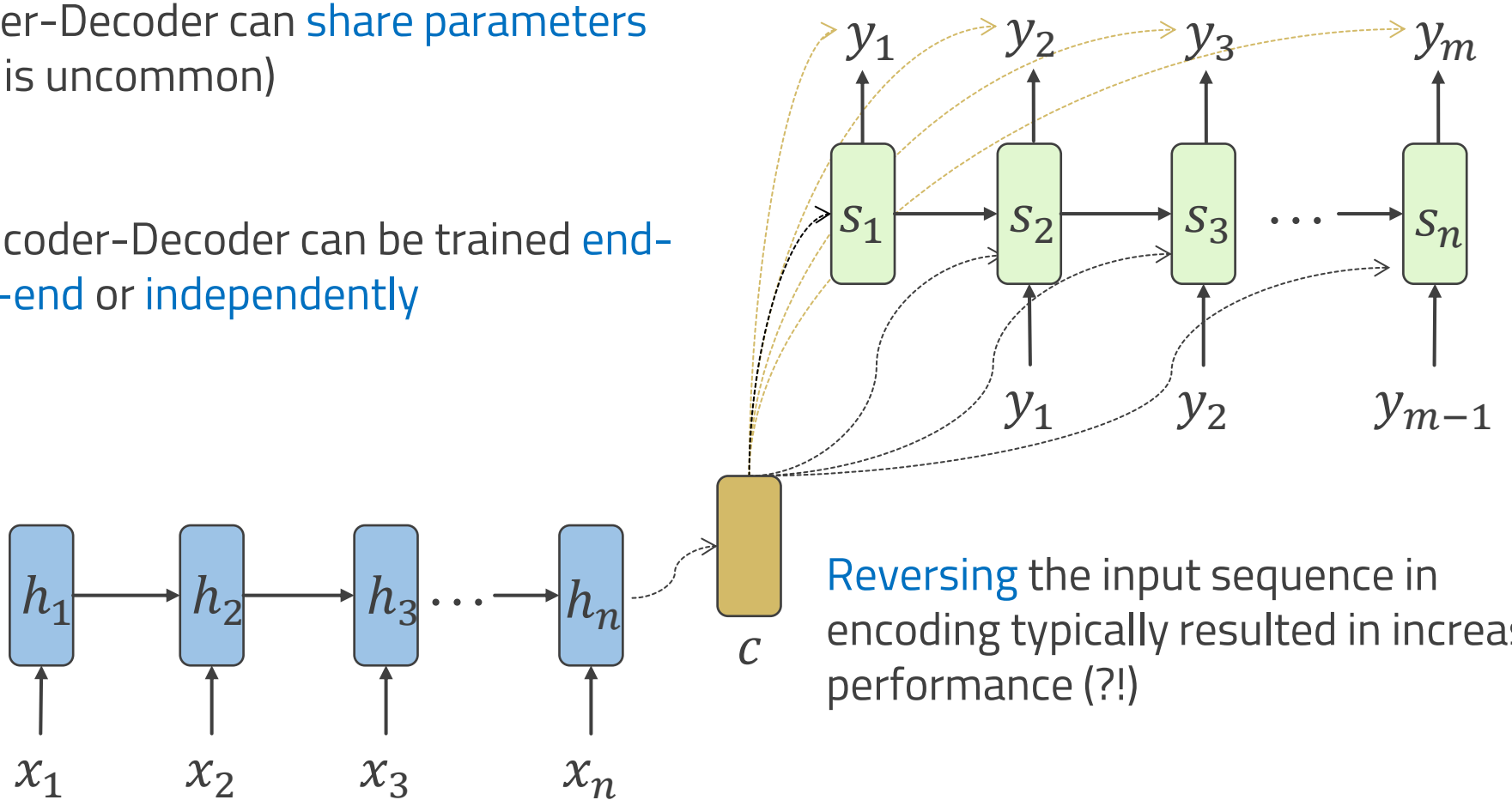
It is better to work on an autoregressive scheme

Remember teacher forcing (only) at training time

Sequence-To-Sequence Learning

Encoder-Decoder can share parameters (but it is uncommon)

Encoder-Decoder can be trained end-to-end or independently



Reversing the input sequence in encoding typically resulted in increased performance (?!)

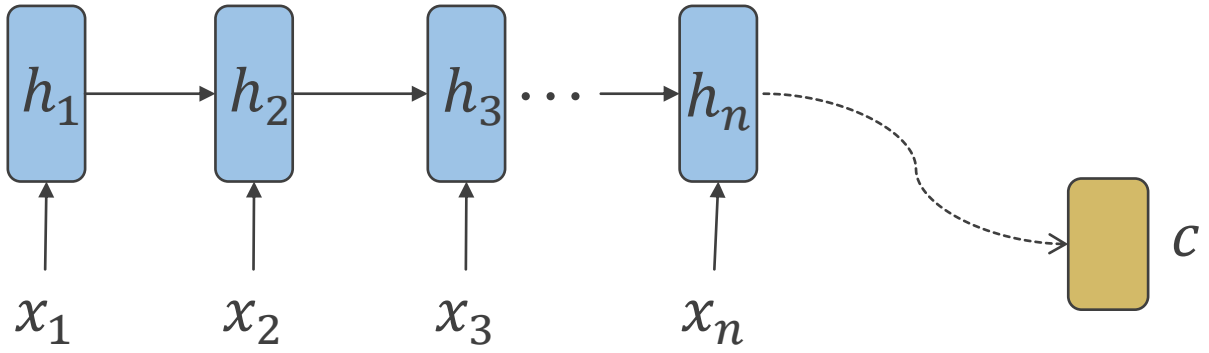
A Motivating Example

The cat is on the table

Il gatto è sul tavolo

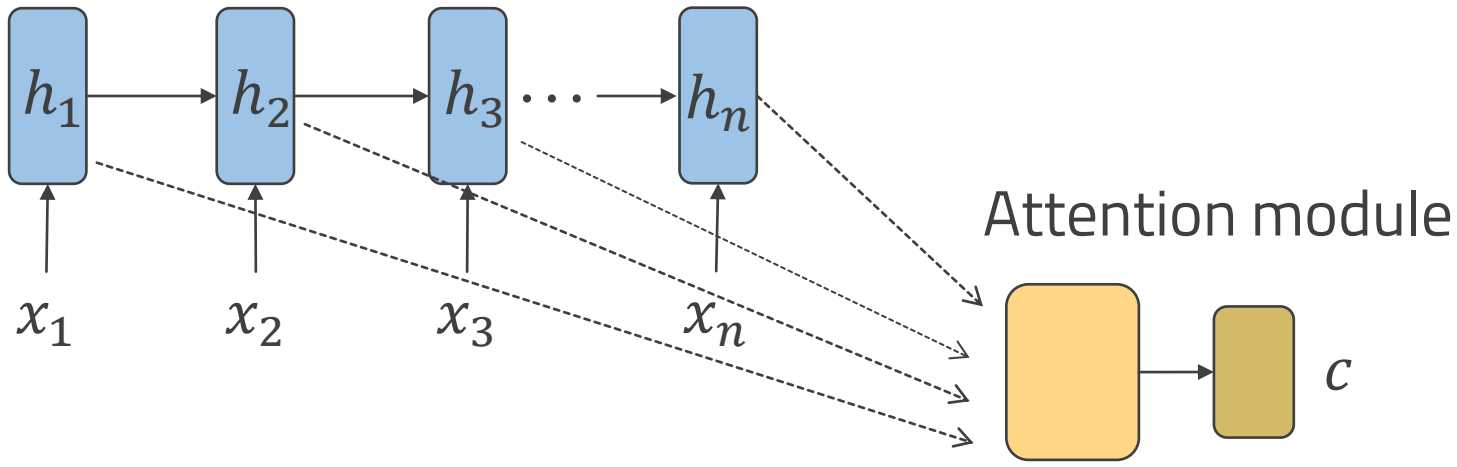
Soft-Attention

On the Need of Paying Attention



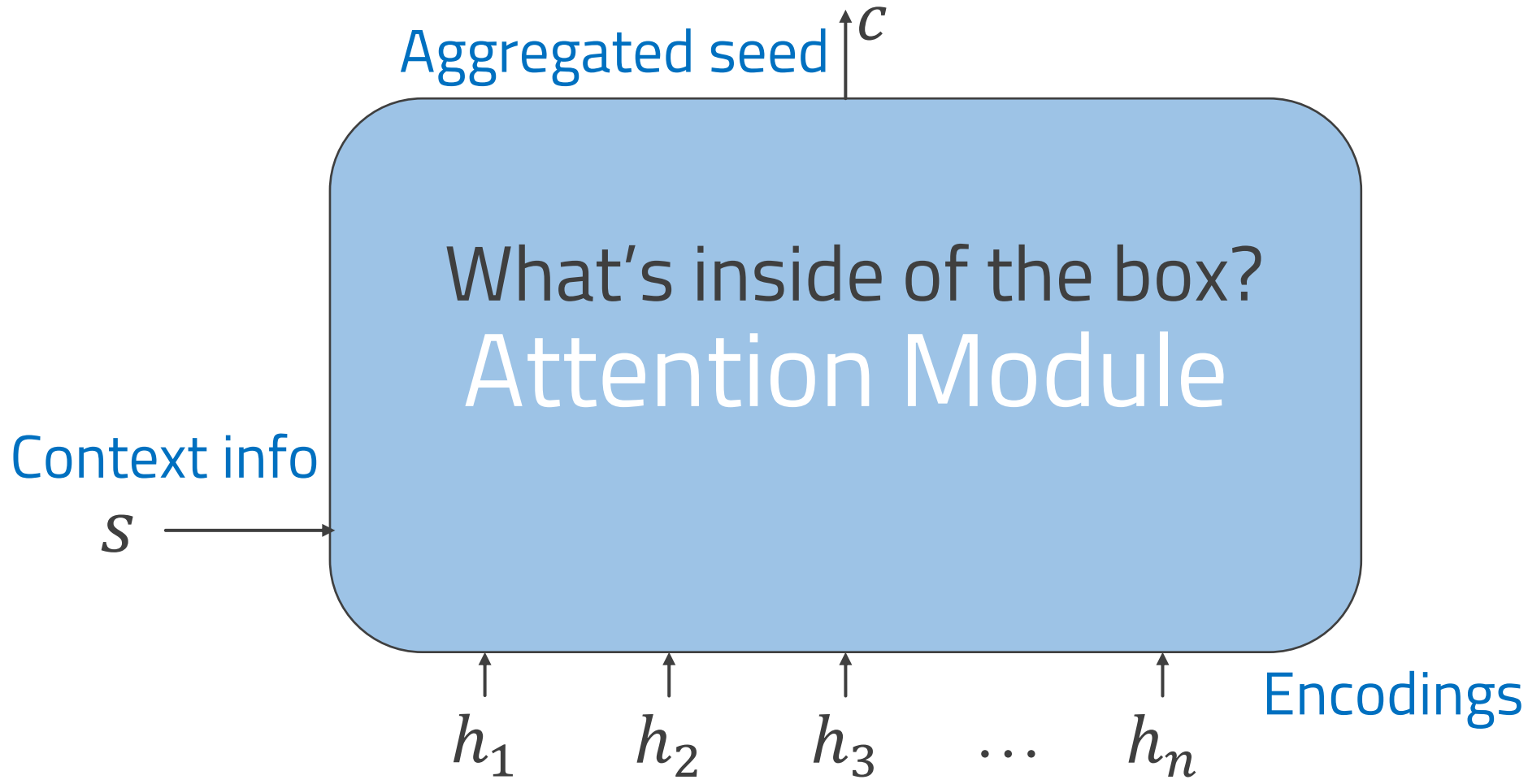
- Encoder-Decoder scheme assumes the hidden activation of the last input element summarizes sufficient information to generate the output
  - Bias toward most recent past
- Other parts of the input sequence might be very informative for the task
  - Possibly elements appearing very far from sequence end

On the Need of Paying Attention



Attention mechanisms select which part of the sequence to focus on to obtain a good c

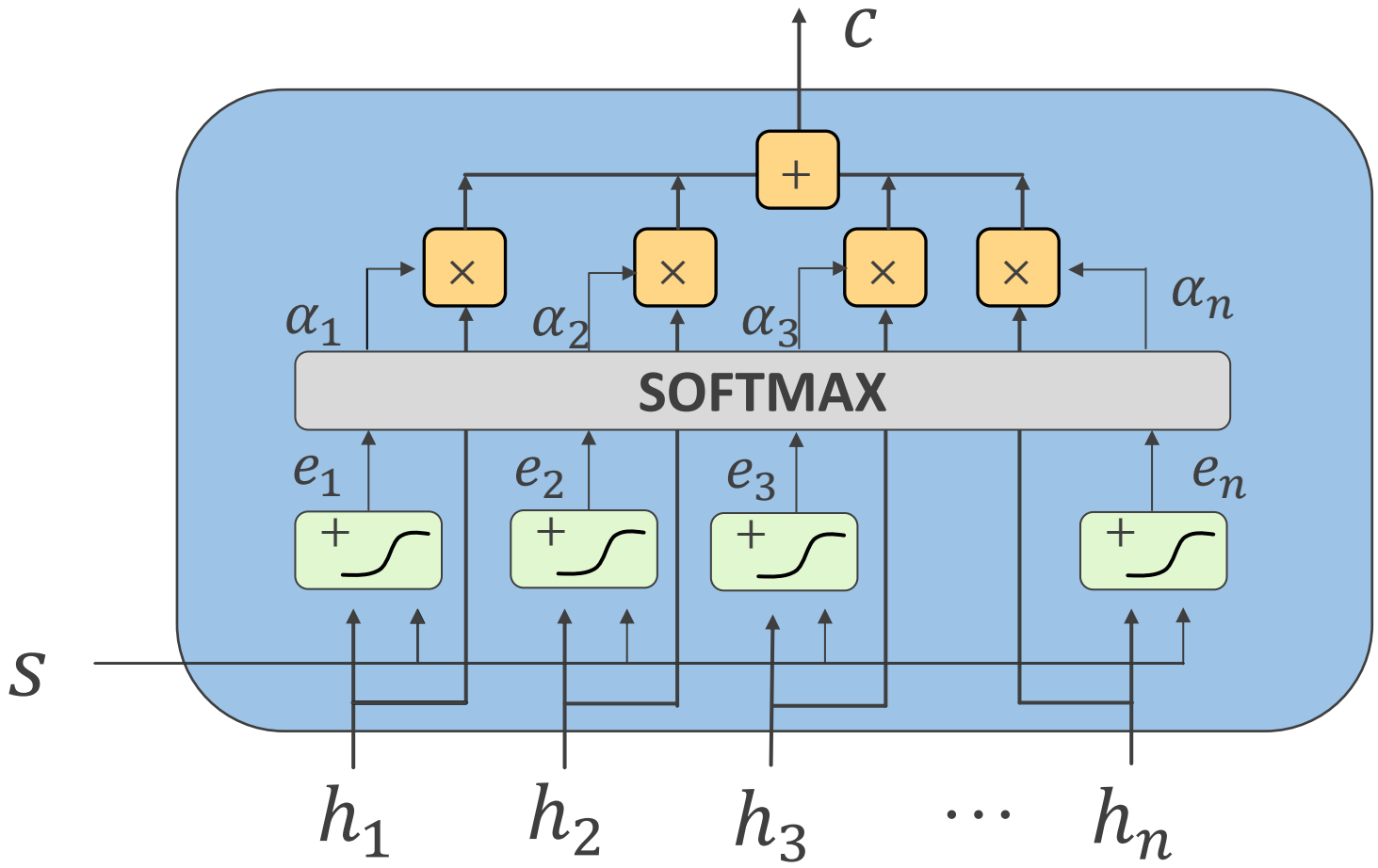
Attention Mechanisms – Blackbox View



What's inside of the box?

The Revenge of the Gates!

Opening the Box

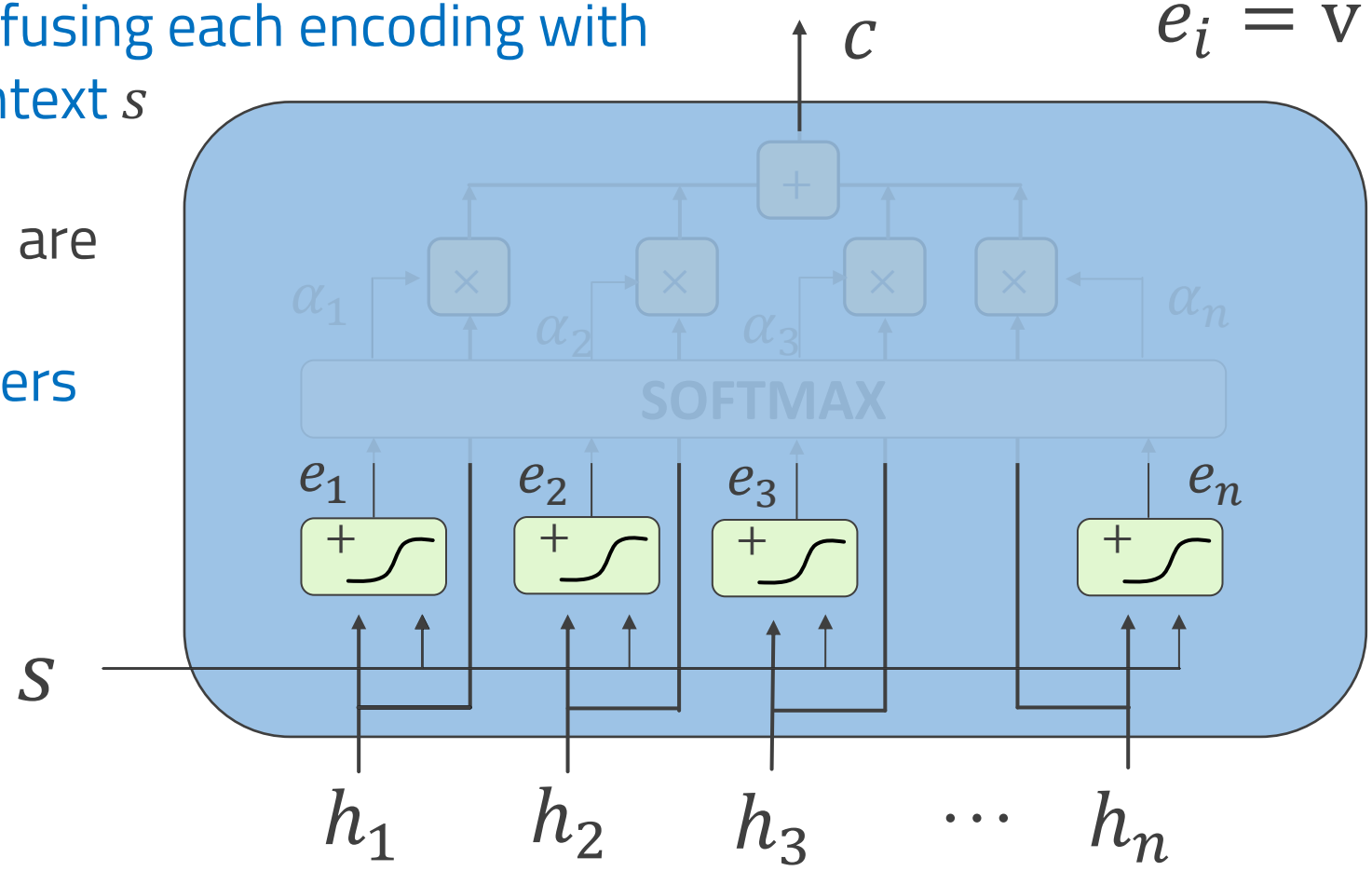


Opening the Box – Relevance/score

Tanh layer fusing each encoding with current context s

$$e_i = \mathrm{v}^{\mathrm{T}}\mathrm{net}_{\mathrm{W}}(s, h_i)$$

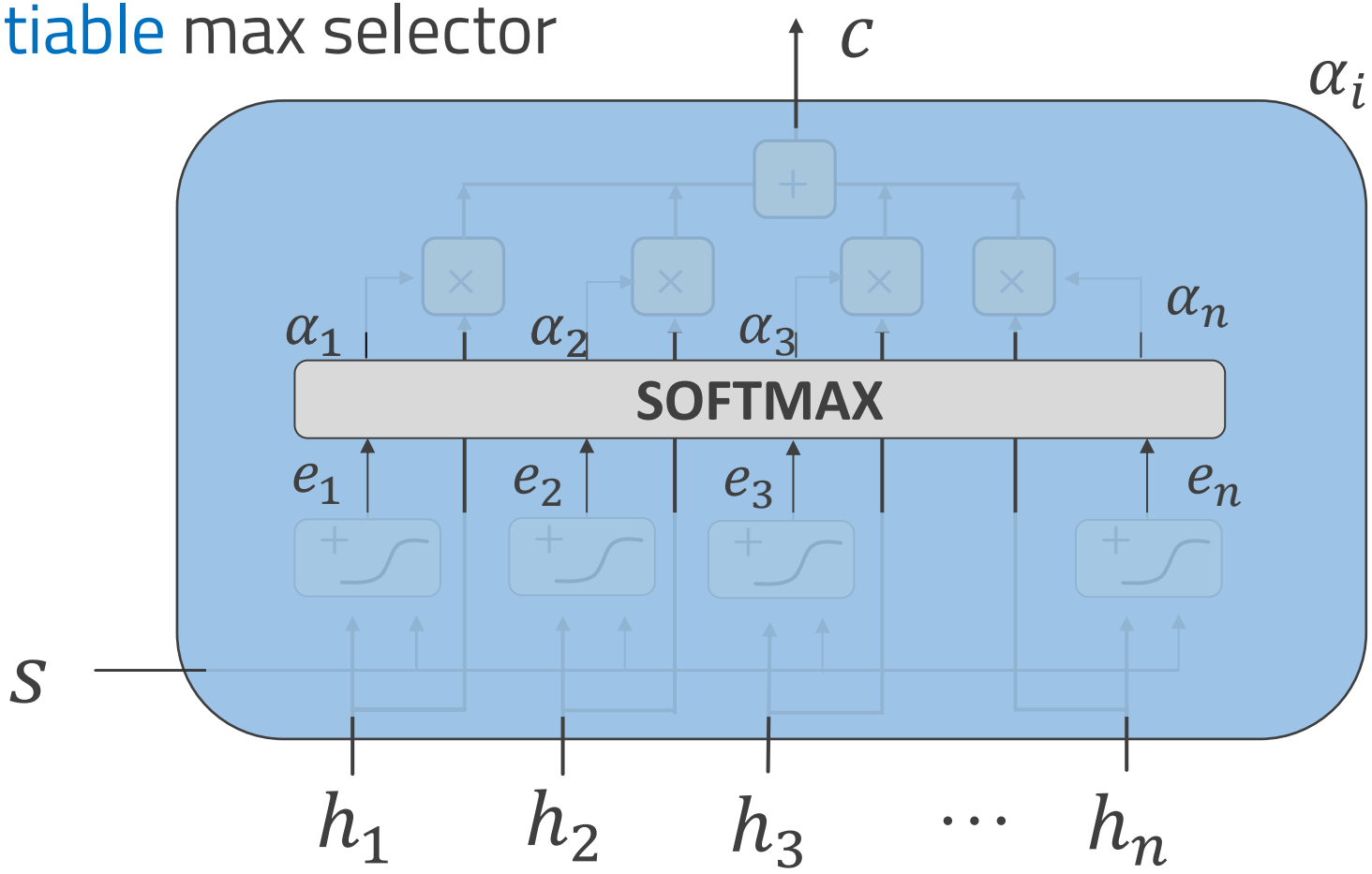
v and W are learned parameters



Opening the Box – Softmax

A differentiable max selector operator

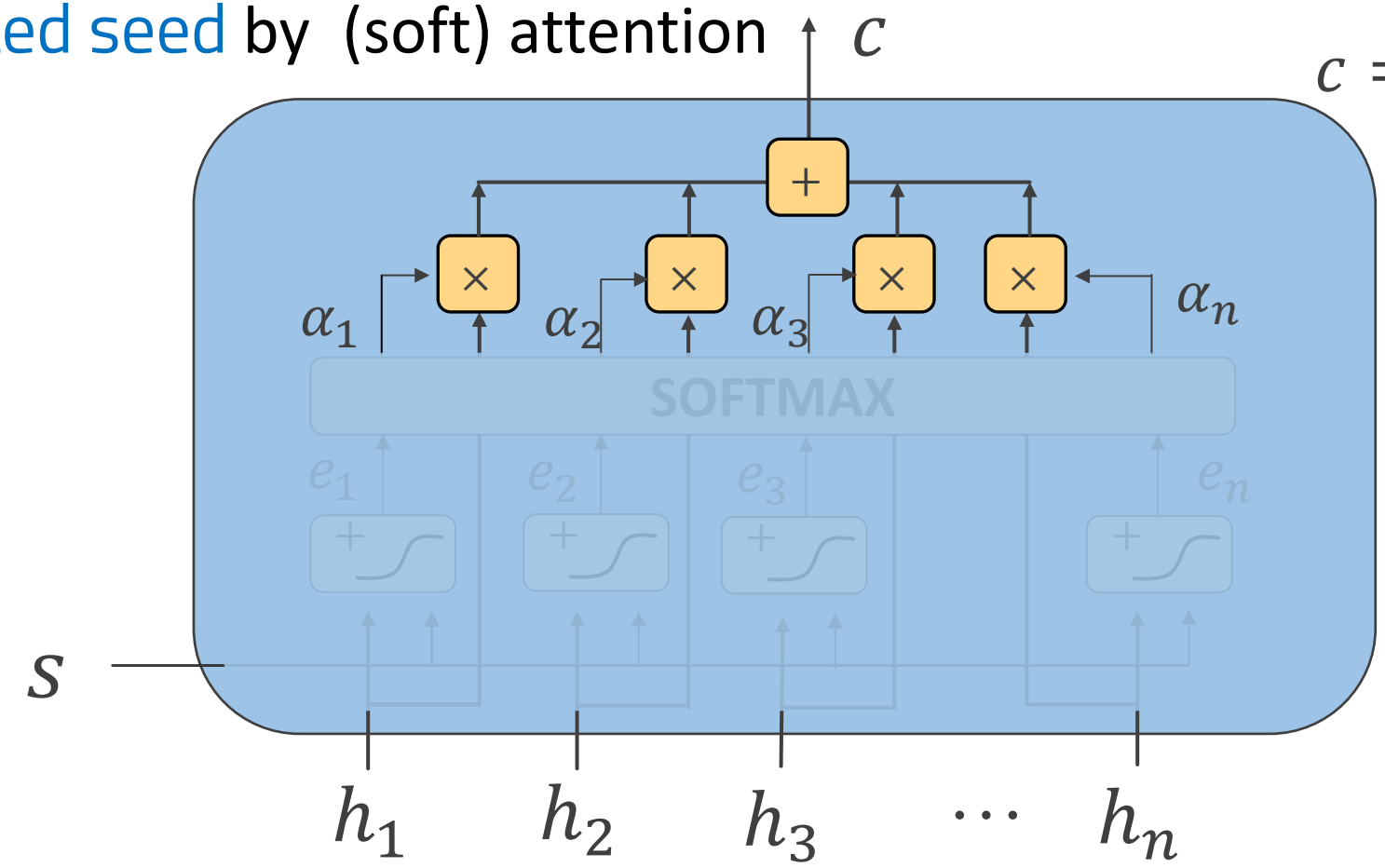
$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}$$



Opening the Box – Voting

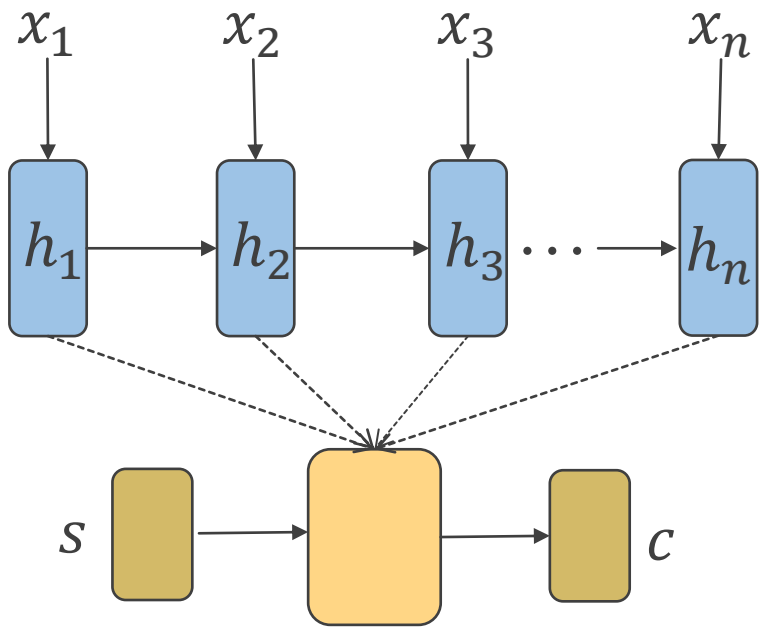
Aggregated seed by (soft) attention voting

$$c = \sum_i \alpha_i h_i$$



Soft-Attention - Equations

- Score: $e_i = \mathrm{v}^{\mathrm{t}}\mathrm{net}_{\mathrm{W}}(s, h_i)$
- Normalization: $\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}$
- Aggregation: $c = \sum_i \alpha_i h_i$

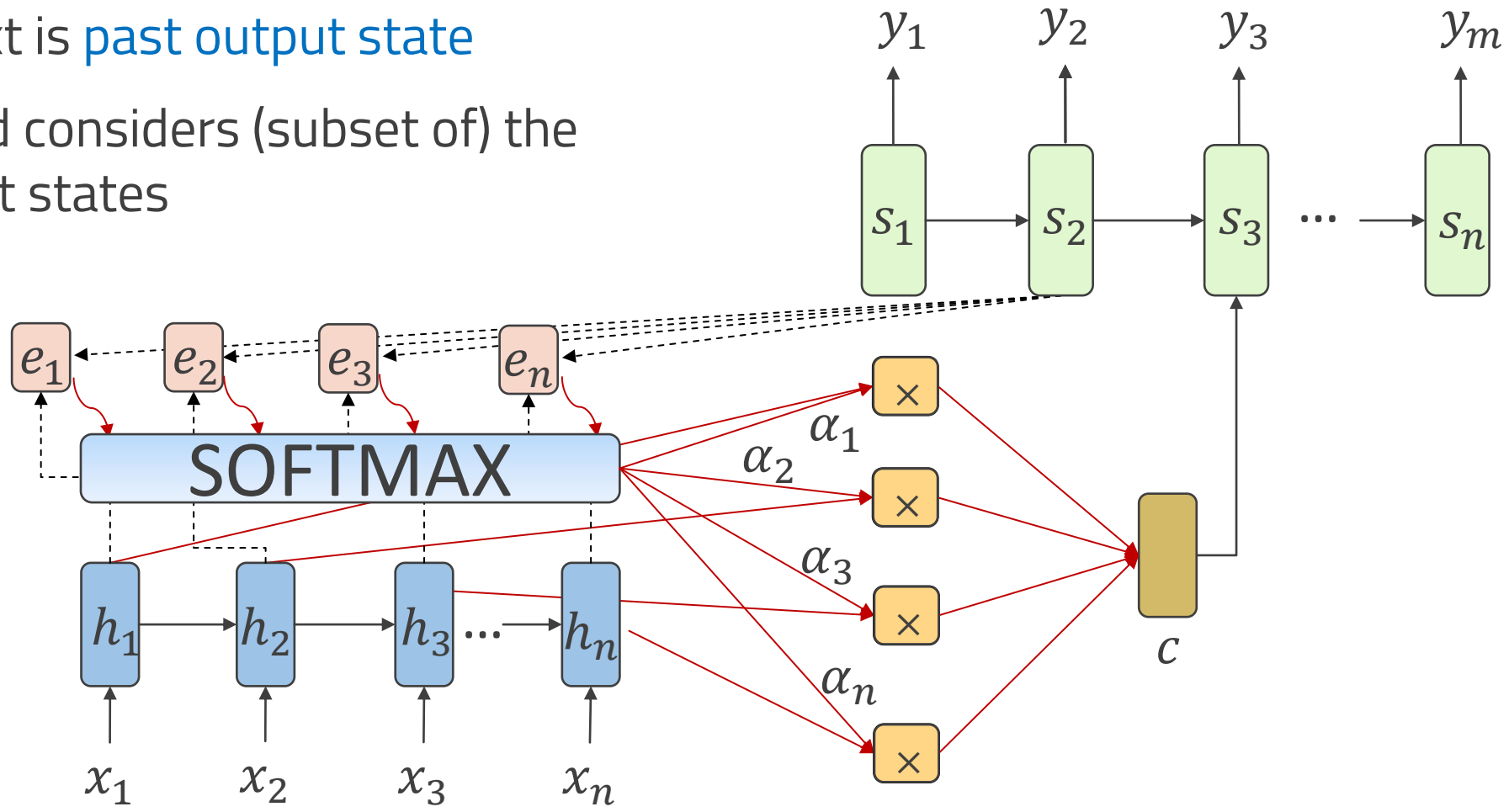


Attention module

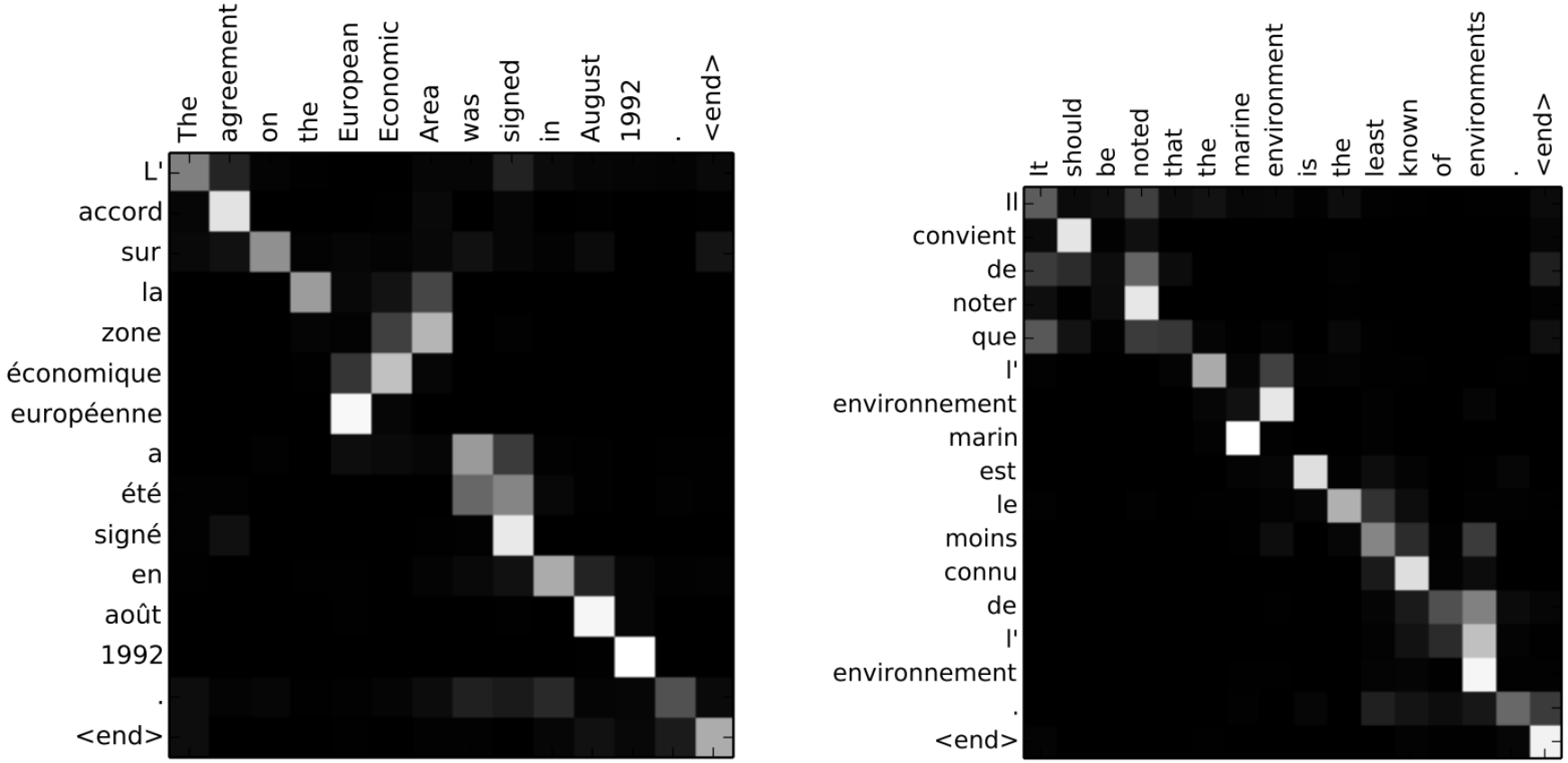
Cross-Attention in Seq2Seq

Context is past output state

Seed considers (subset of) the input states



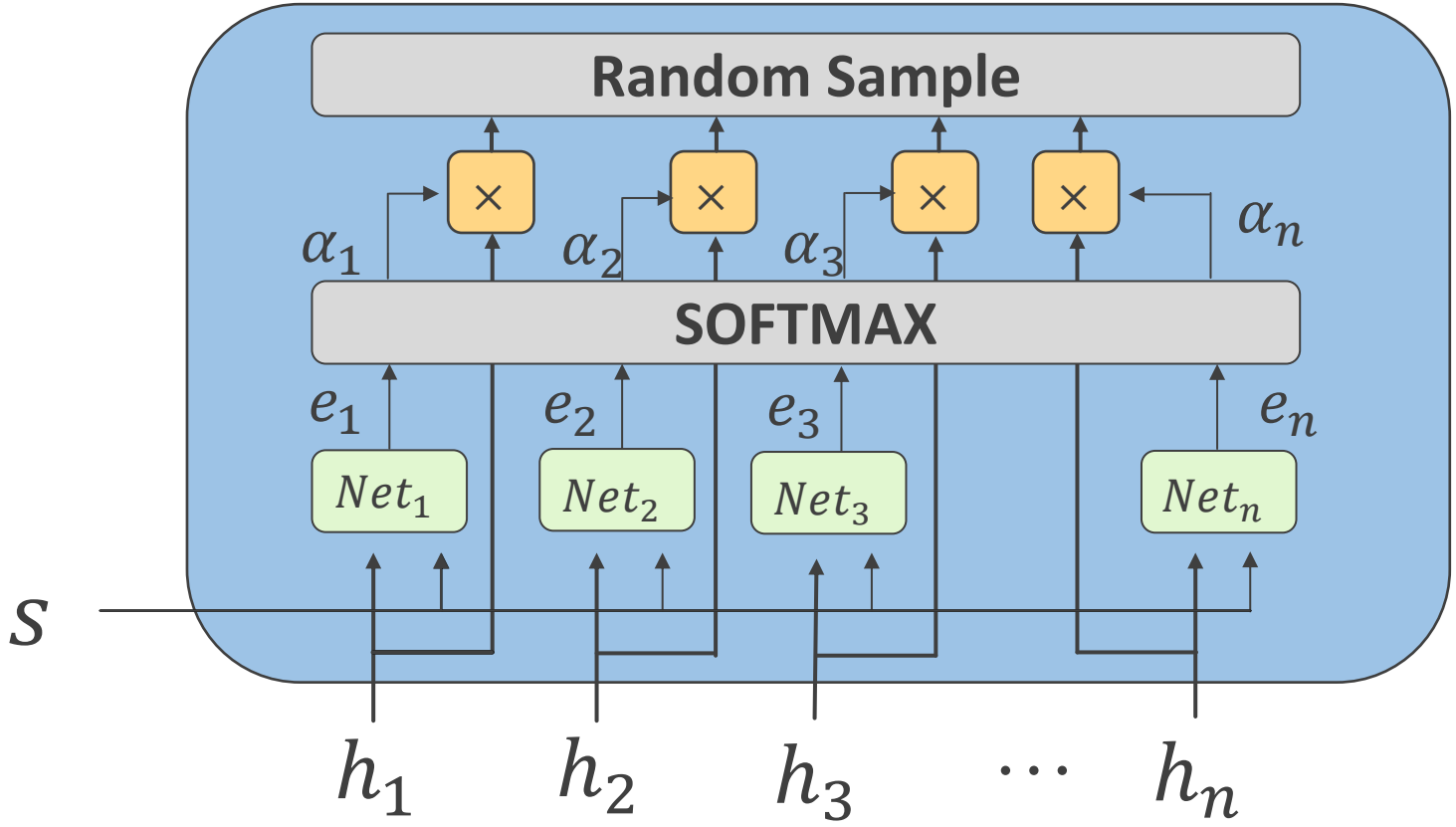
Learning to Translate with Attention



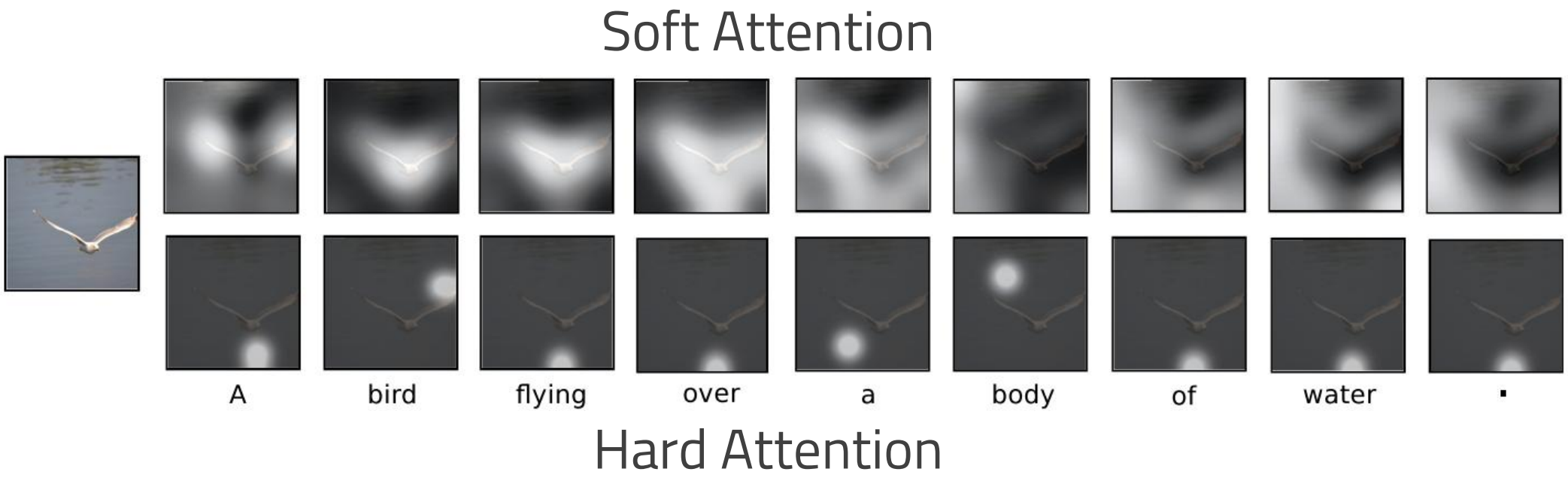
Bahdanau et al, Show, Neural machine translation by jointly learning to align and translate, ICLR 2015

Hard-Attention

Sample a single encoding using probability α_i



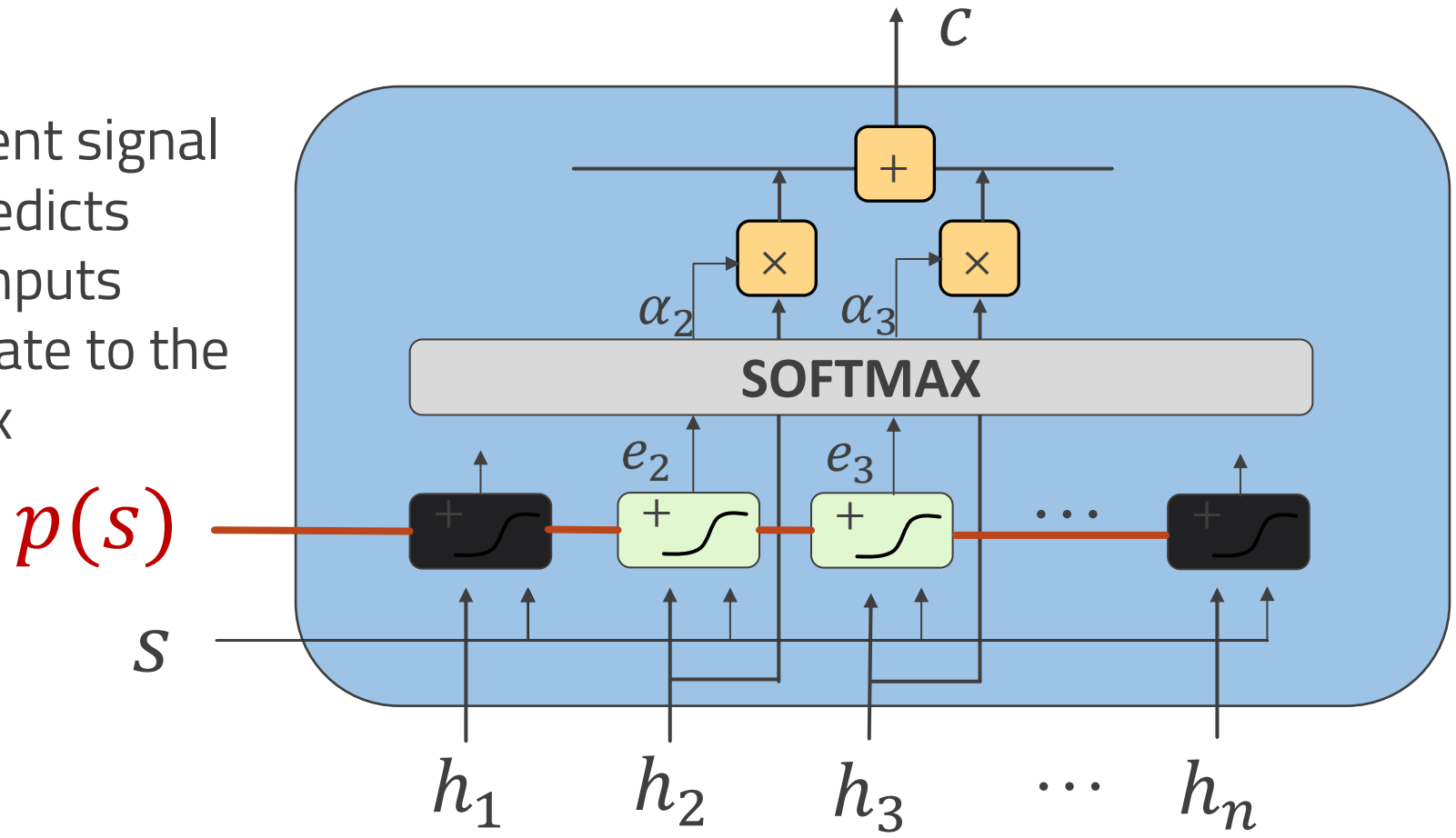
Attention-Based Captioning – Focus Shifting



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

Local Attention

Alignment signal $p(s)$ predicts which inputs participate to the softmax



Making hard-attention differentiable

Wrap-up

Generalizing Attention

Different forms depending on how we score association of inputs with current context s

Name	Alignment score function
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_{t-1}; \boldsymbol{h}_i])$
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_i^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.

Also referred to as concat attention

We will see more about these in the next lecture

Attention-Based Captioning - Interpretation

Learns to correlate textual and visual concepts

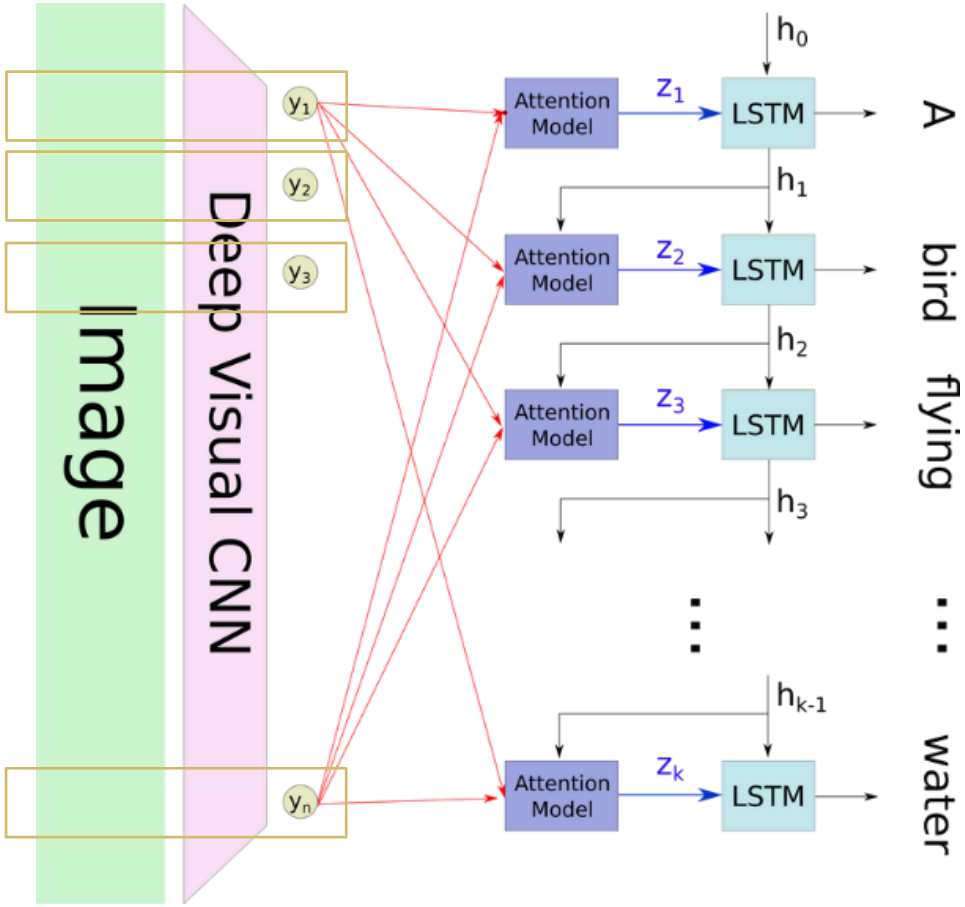
Helps understanding why the model fails

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

Heterogenous Encoder-Decoder Models

Encodings associated to *n* image regions

From convolutional layers rather than from fully connected



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

Take Home Messages

- Attention.. Attention.. and, again, attention
  - Soft attention is nice because makes everything fully differentiable
  - Hard attention is stochastic hence cannot Backprop
  - Empirical evidences of them being sensitive to different things
- Encoder-Decoder scheme
  - A general architecture to compose heterogeneous models and data
  - Decoding allows sampling complex predictions from an encoding conditioned distribution
- Setting the ground for a well-known architecture
  - The Transformer

Next Lecture

A particular form of attention which changed the world

- Self-attention
- Transformer
- Inductive bias and gradient propagation
- Self-supervised training