Attention-based architectures

INTELLIGENT SYSTEMS FOR PATTERN RECOGNITION (ISPR)

DAVIDE BACCIU – DIPARTIMENTO DI INFORMATICA - UNIVERSITA’ DI PISA

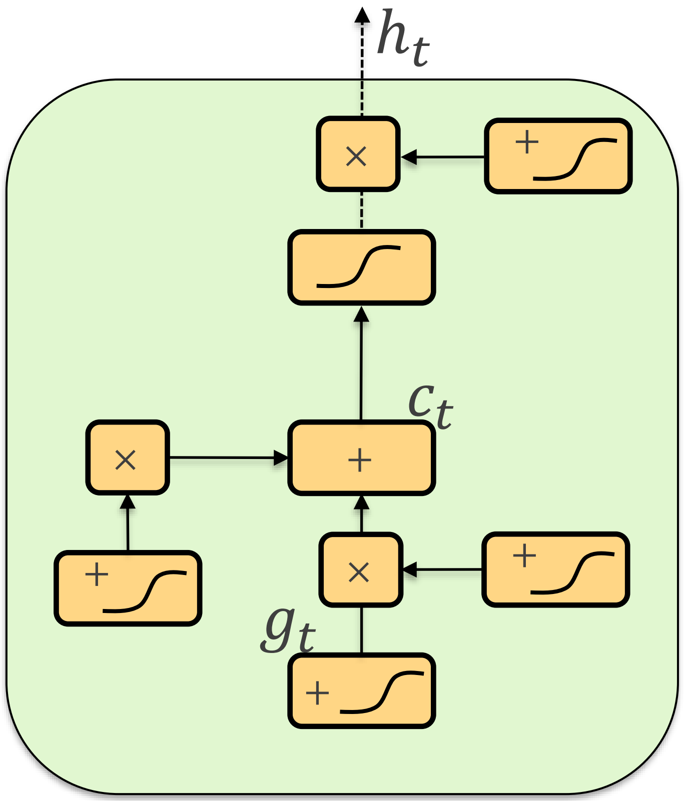
DAVIDE.BACCIU@UNIPI.IT

A 2 Lectures Outline

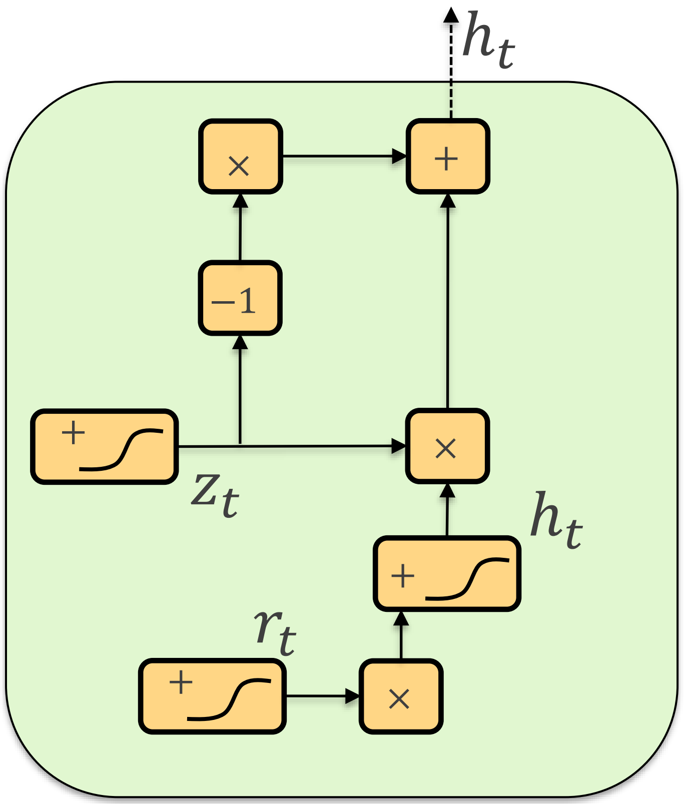
- L22 - Neural attention for structured/compound data
  - Sequence-to-sequence
  - Attention models
- L23 - Dealing with very long-term dependencies
  - Multiscale networks
  - Neural memories (more attention)
  - Differentiable memory read, write, indexing

Extra Lecture
Tomorrow 12/04/2024
– h16 – Aula E

Gated RNN Refresher



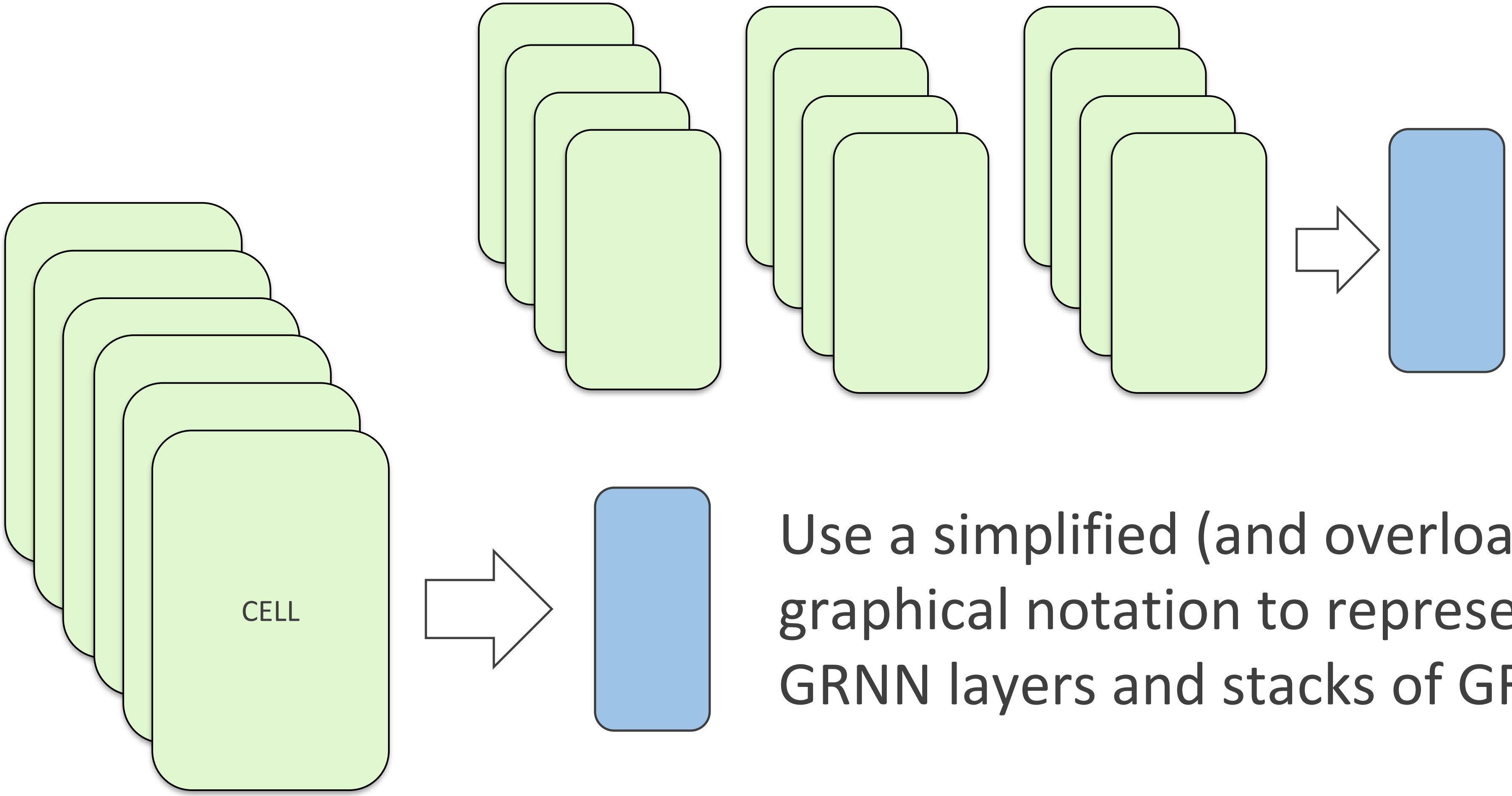
LSTM Cell



GRU Cell

UNIVERSITÀ DI PISA

Graphical Notation for Compositionality



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

Dealining 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
  - Attention mechanism

Sequence-to-sequence

Sequence Transduction

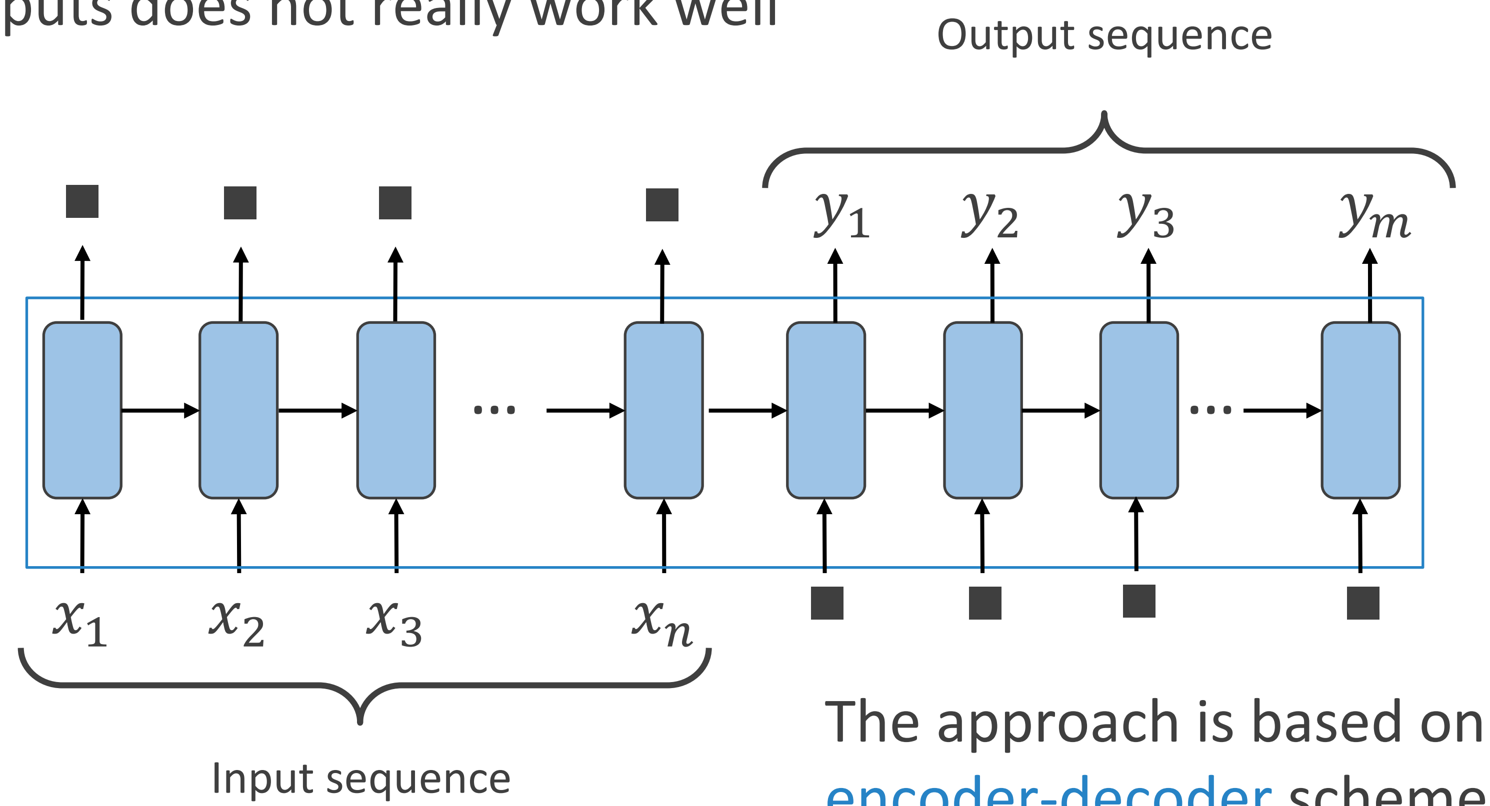
- Input and output are both sequences
- They may have different lengths
- Example: machine translation

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

How do we model the context here?

Learning to Output Variable Length Sequences

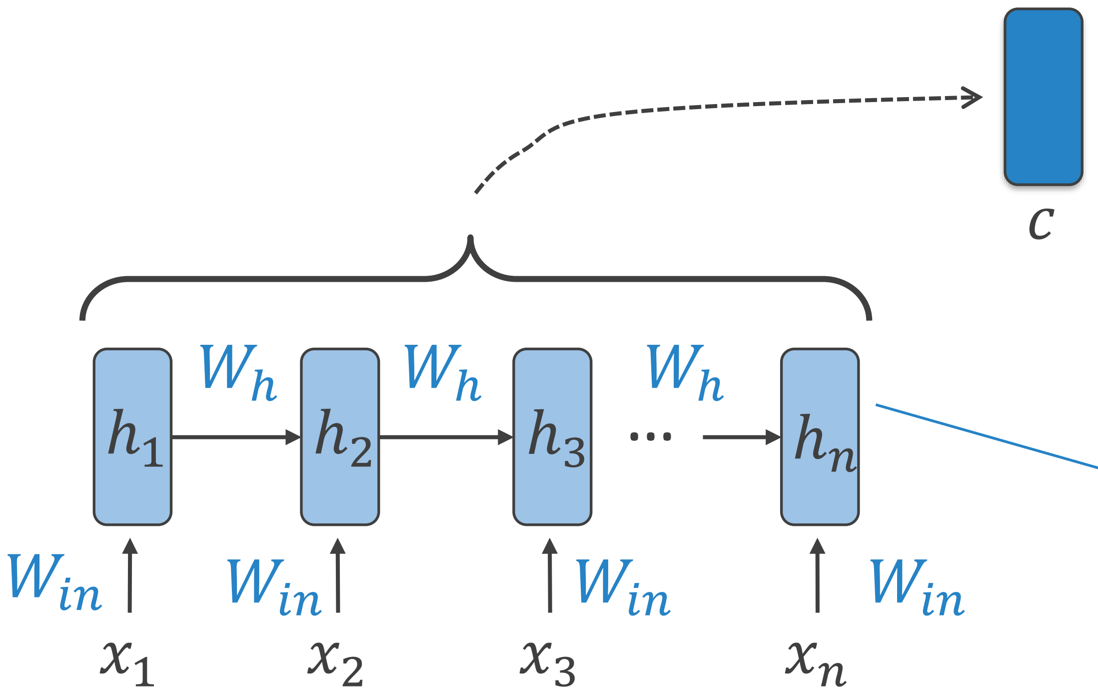
The idea of an unfolded RNN with blank inputs-outputs does not really work well



The approach is based on an encoder-decoder scheme

Encoder

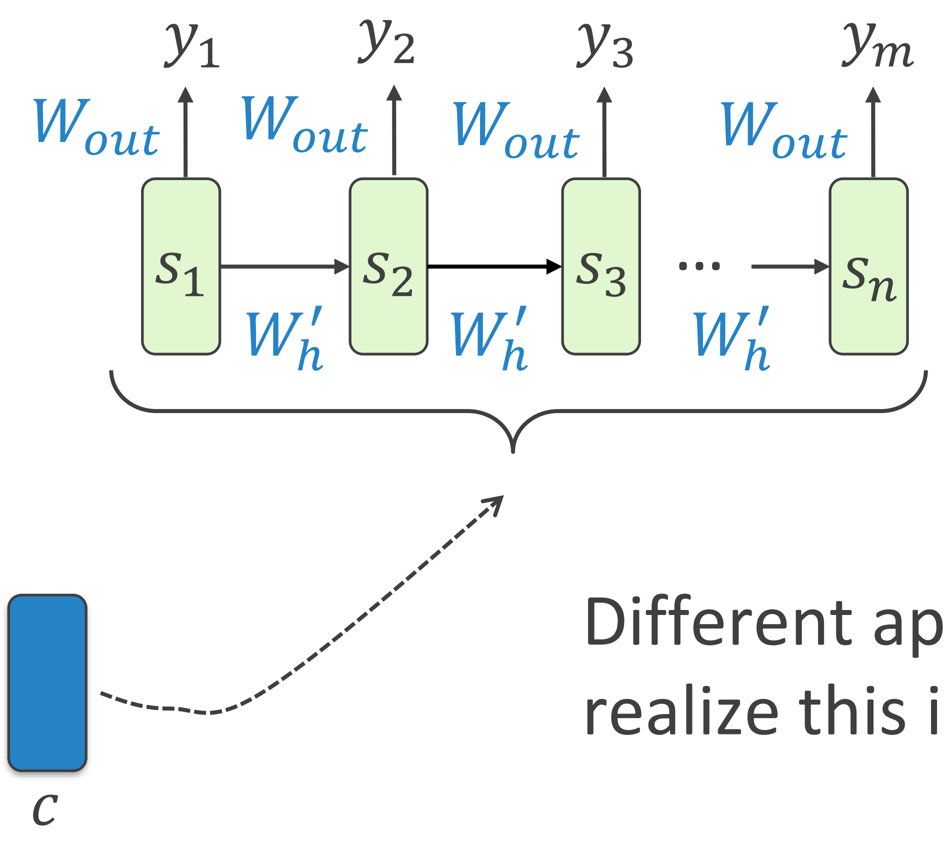
Produce a compressed and fixed length representation c of all the input sequence $x_1, \dots, x_n$



Originally $c = h_n$

Activations of an LSTM/GRU layer of K cells

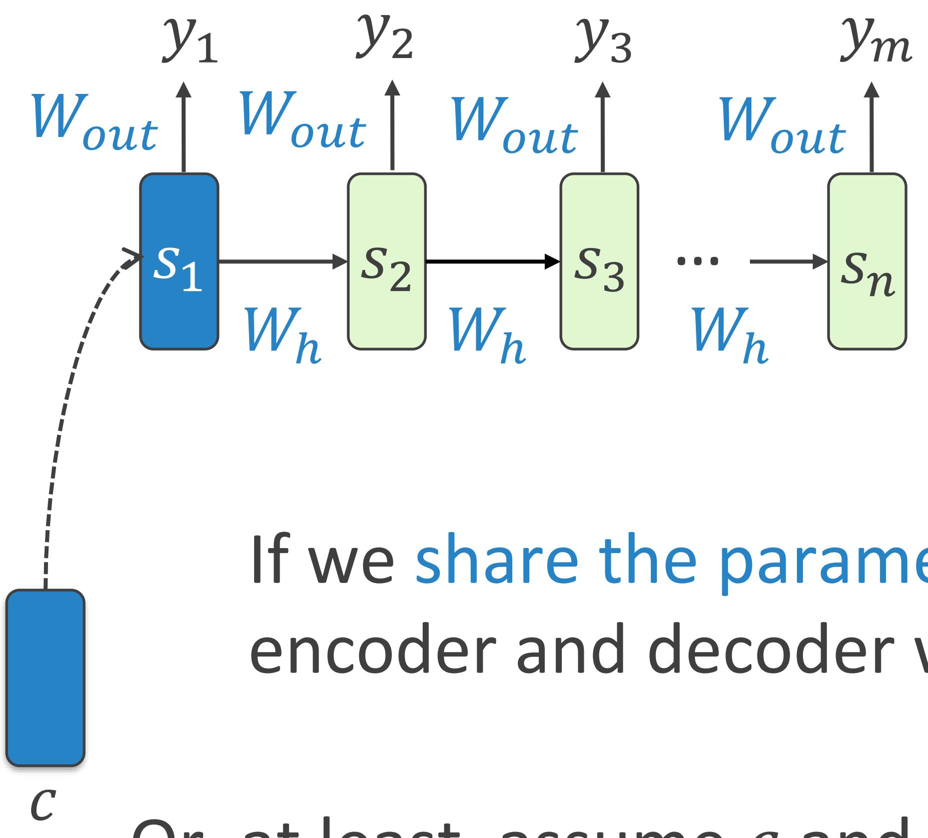
Decoder



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

Different approaches to realize this in practice

Decoder

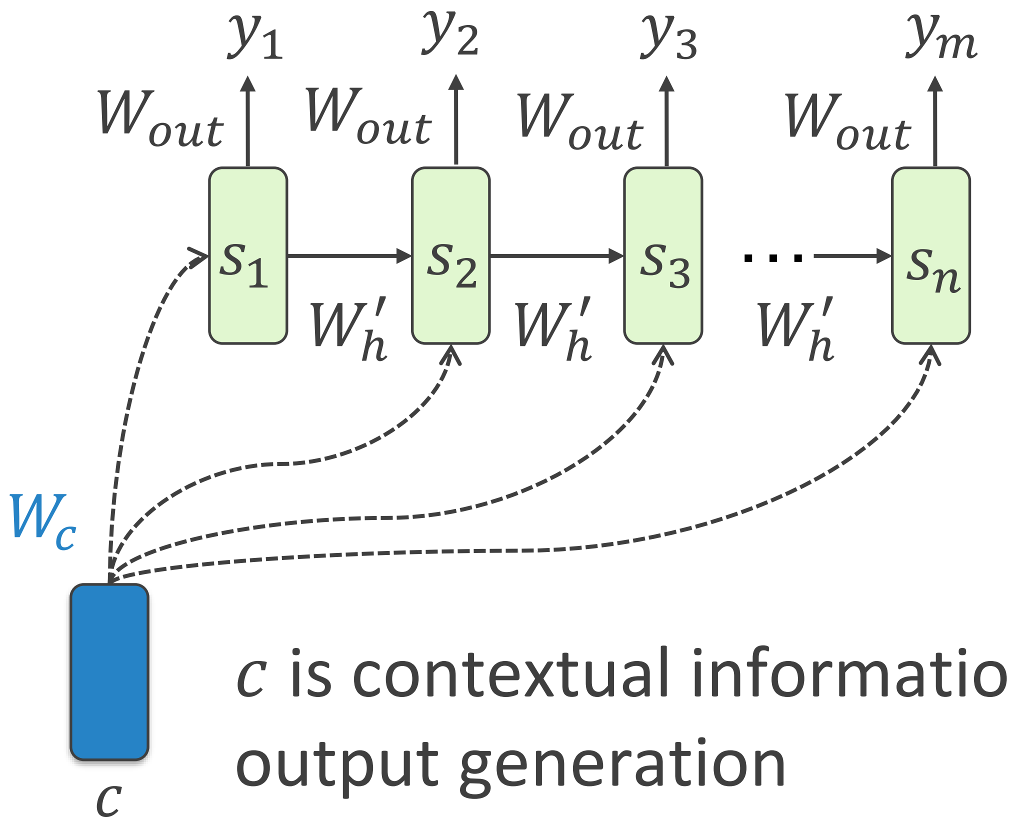


We risk to lose memory of c soon

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

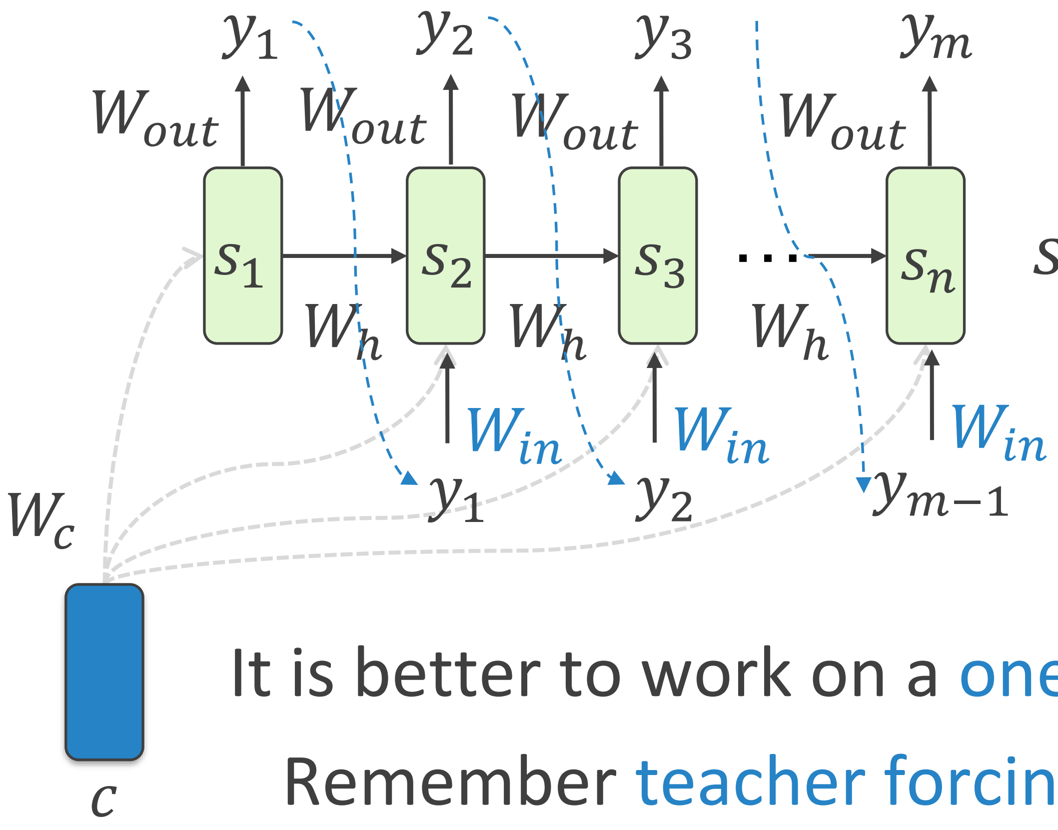
Or, at least, assume c and s_1 have compatible size

Decoder



c is contextual information kept throughout output generation

Decoder



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

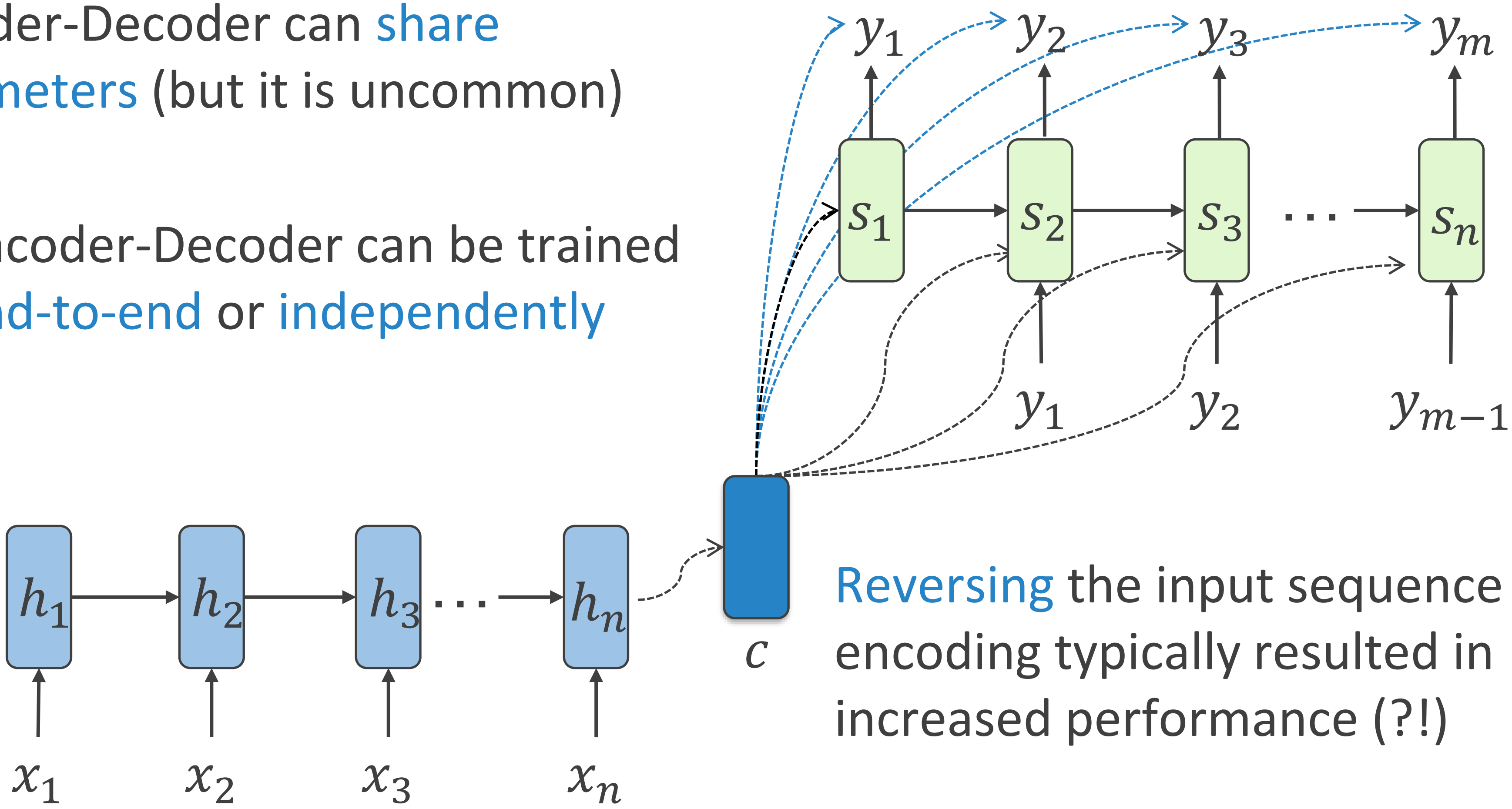
It is better to work on a one-step-ahead scheme

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 (?!)

Università di Pisa

A Motivating Example

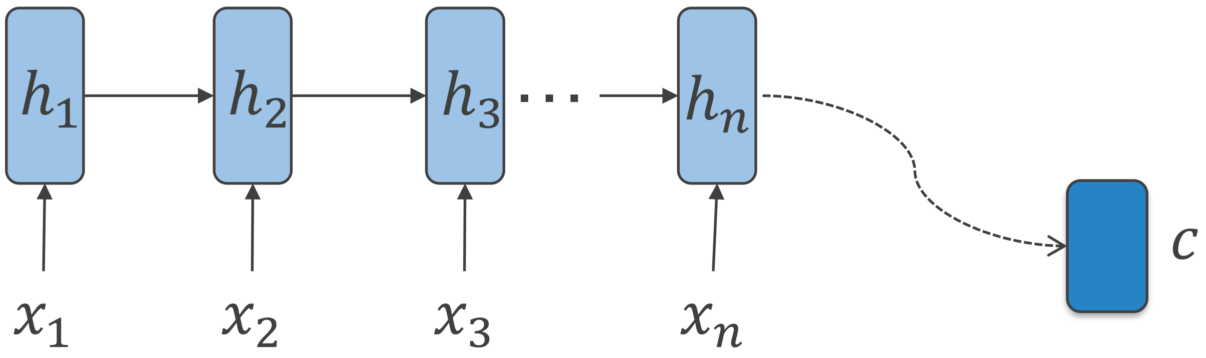
The cat is on the table

Il gatto è sul tavolo

Università di Pisa

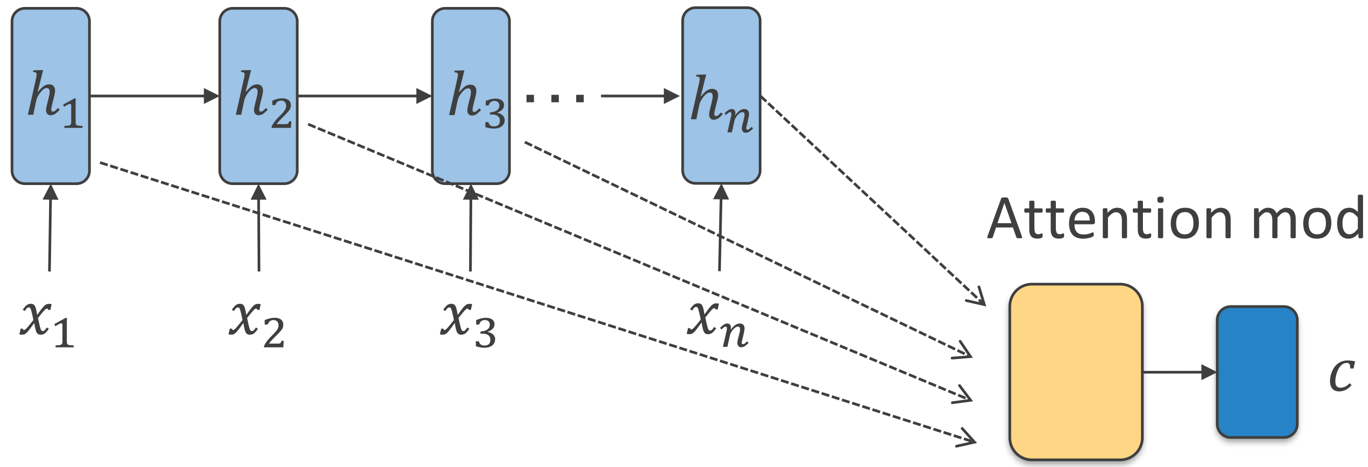
Attention

On the Need of Paying Attention



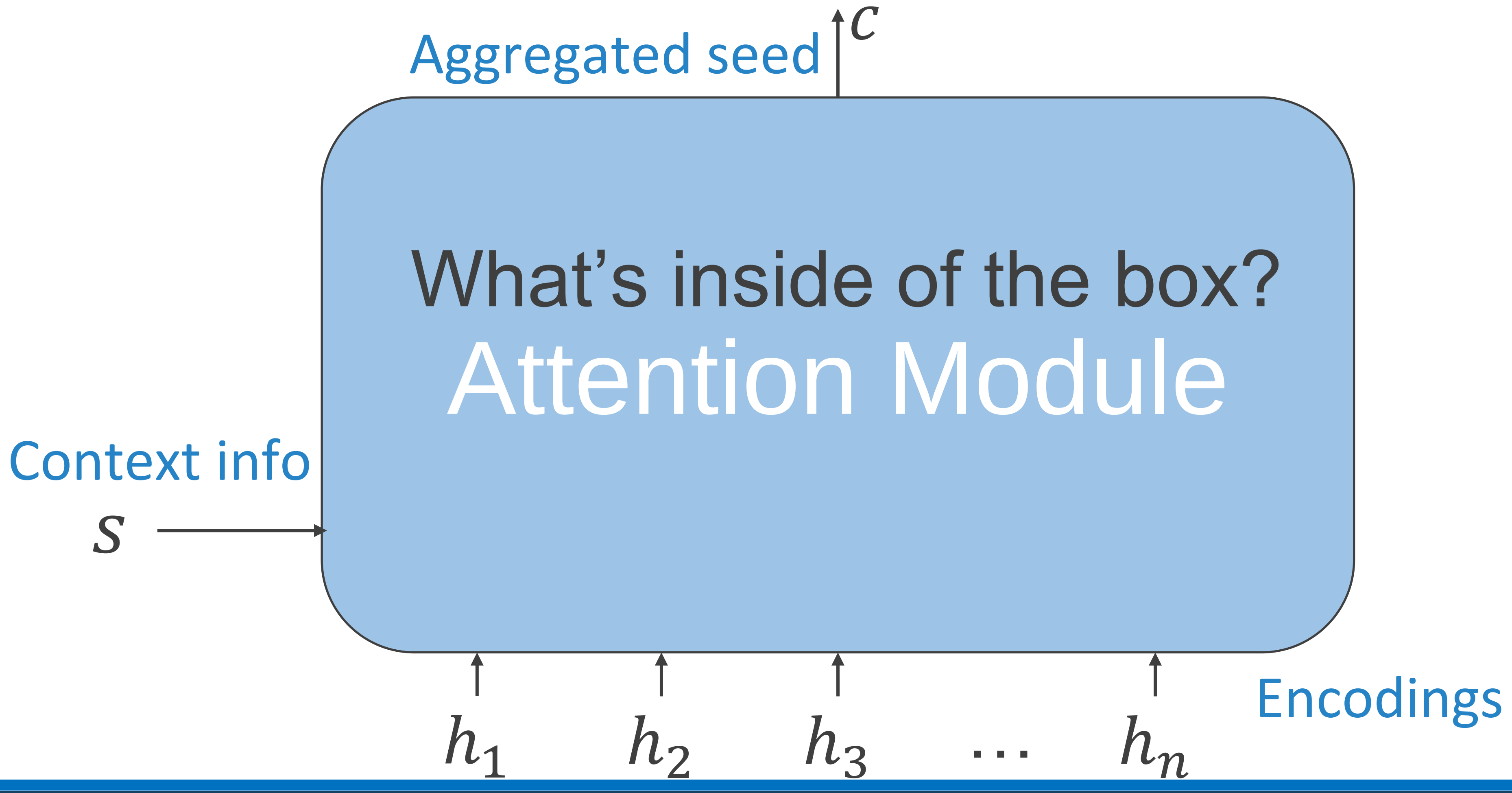
- Encoder-Decoder scheme assumes the hidden activation of the last input element summarizes sufficient information to generate the output
  - Bias toward most recent past
- Other parts of the input sequence might be very informative for the task
  - Possibly elements appearing very far from sequence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

On the Need of Paying Attention



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

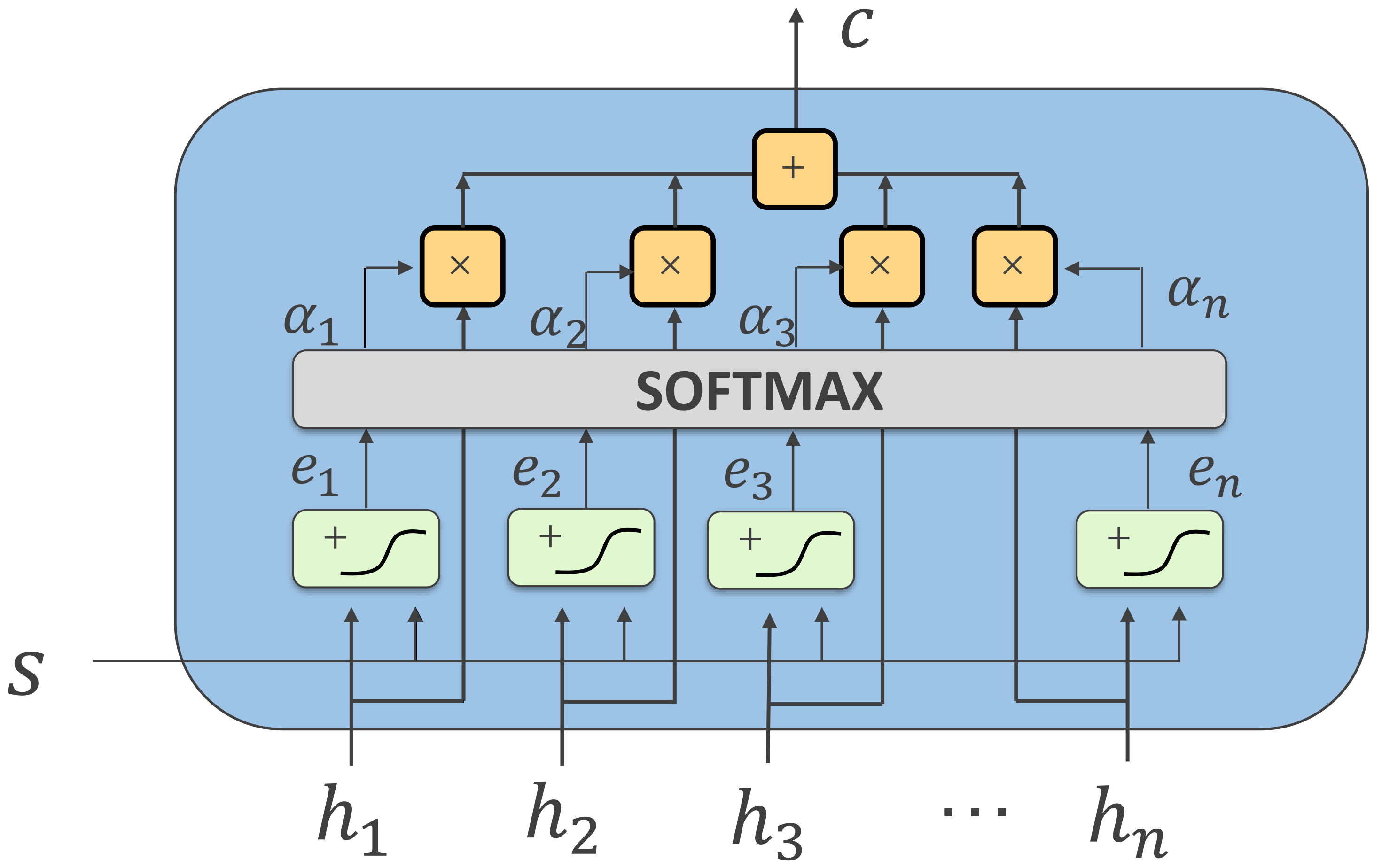
Attention Mechanisms – Blackbox View



What’s inside of the box?

The Revenge of the Gates!

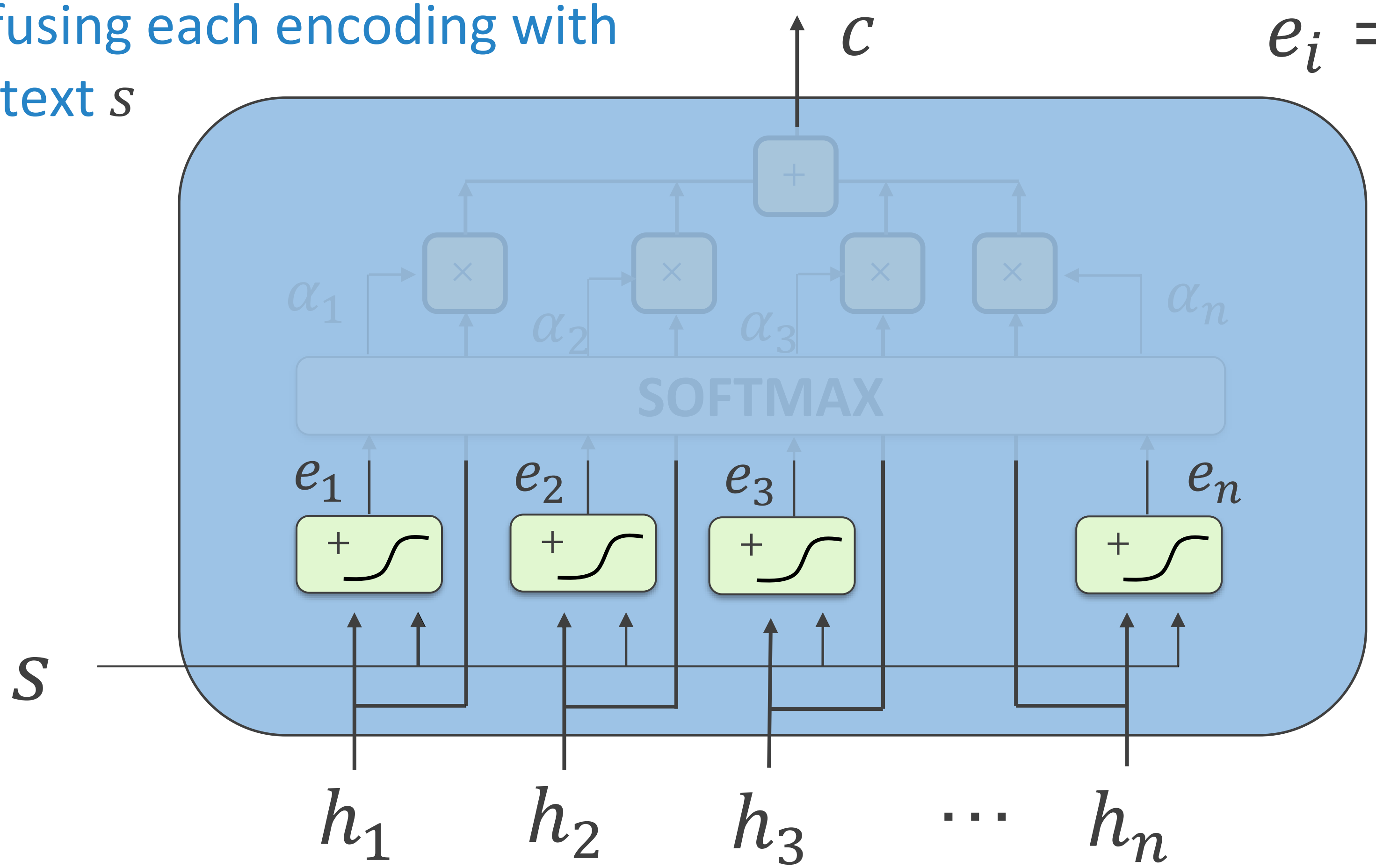
Opening the Box



Opening the Box – Relevance

Tanh layer fusing each encoding with current context s

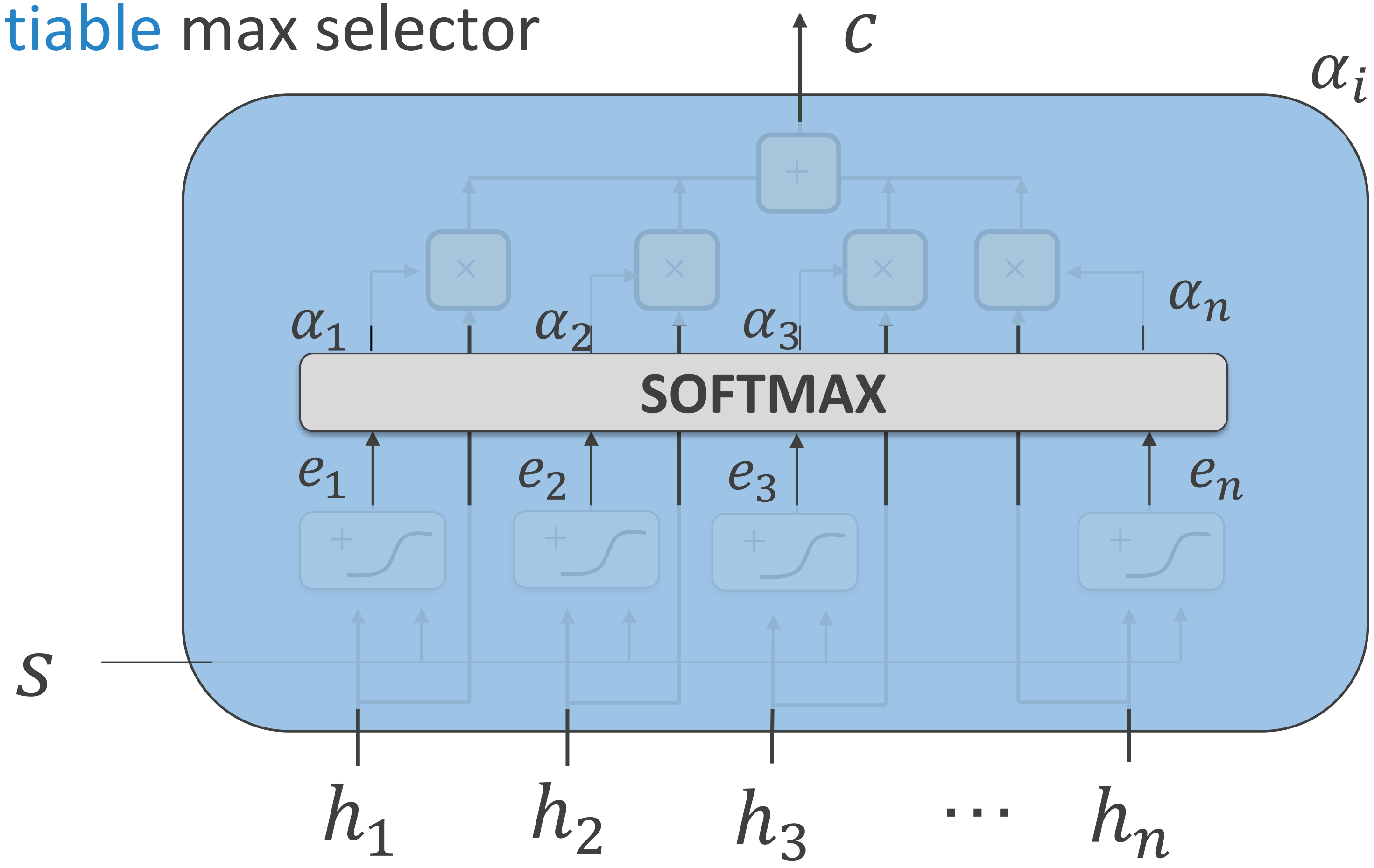
$$e_i = \mathrm{a}(s, h_i)$$



Opening the Box – Softmax

A differentiable max selector operator

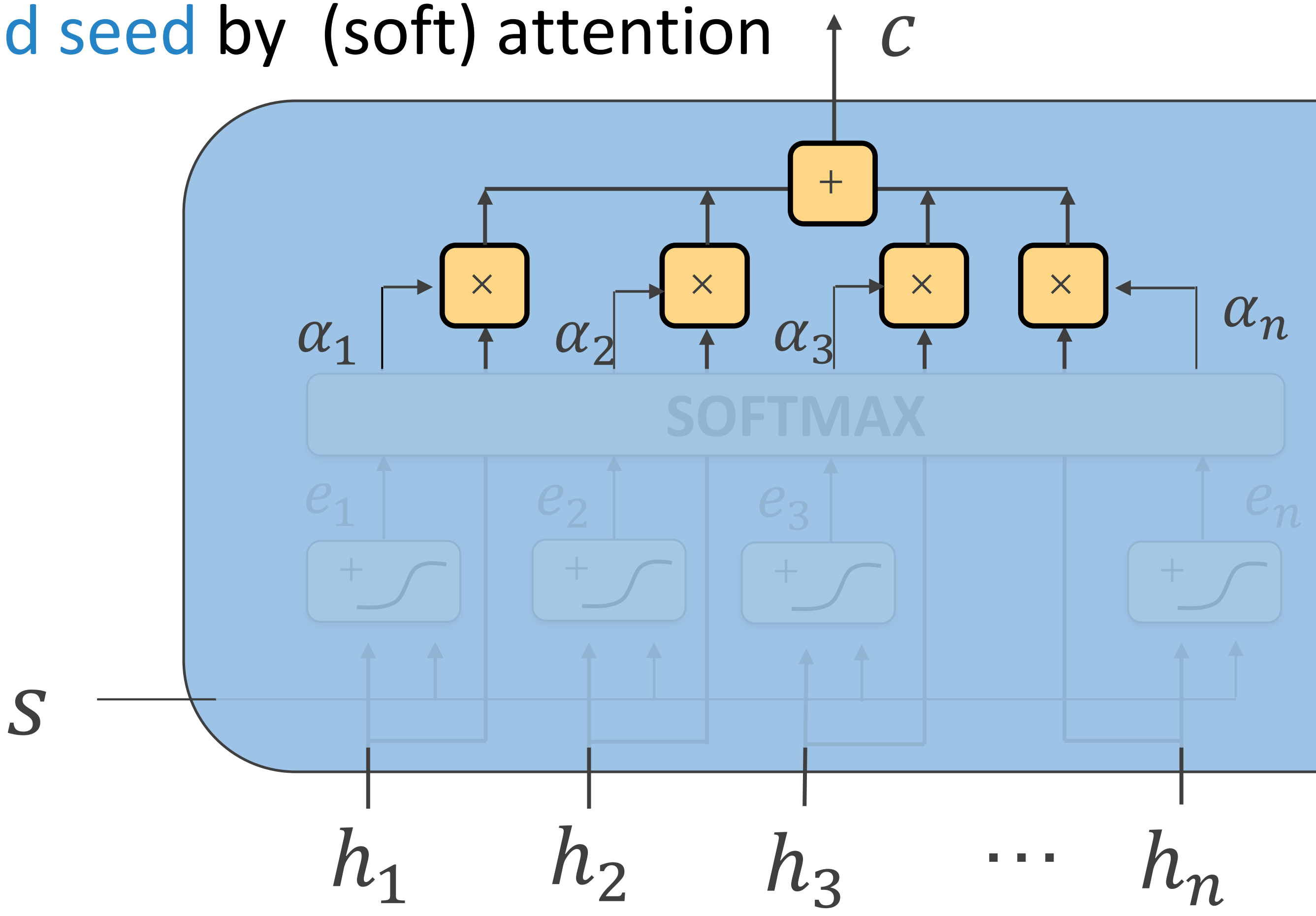
$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}$$



Opening the Box – Voting

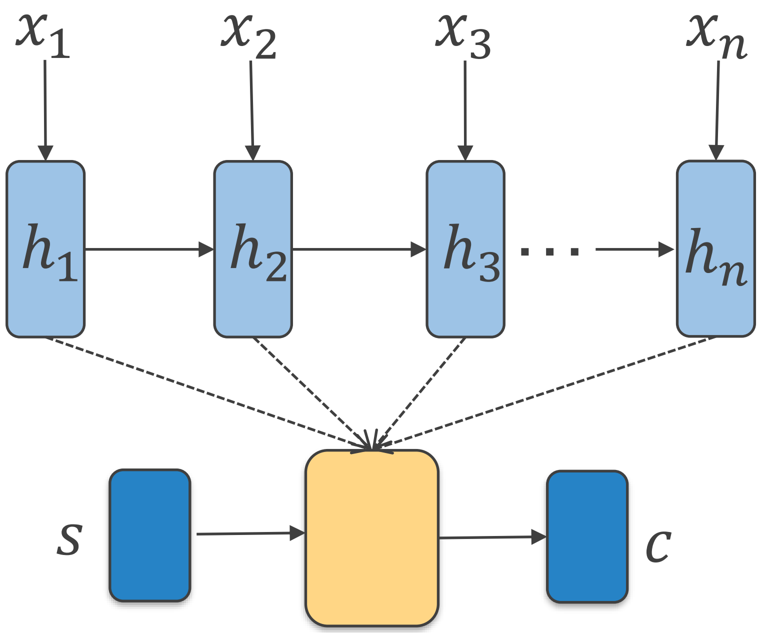
Aggregated seed by (soft) attention voting

$$c = \sum_i \alpha_i h_i$$



Attention - Equations

- Relevance: $e_i = \mathrm{a}(s, h_i)$
- Normalization: $\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}$
- Aggregation: $c = \sum_i \alpha_i h_i$

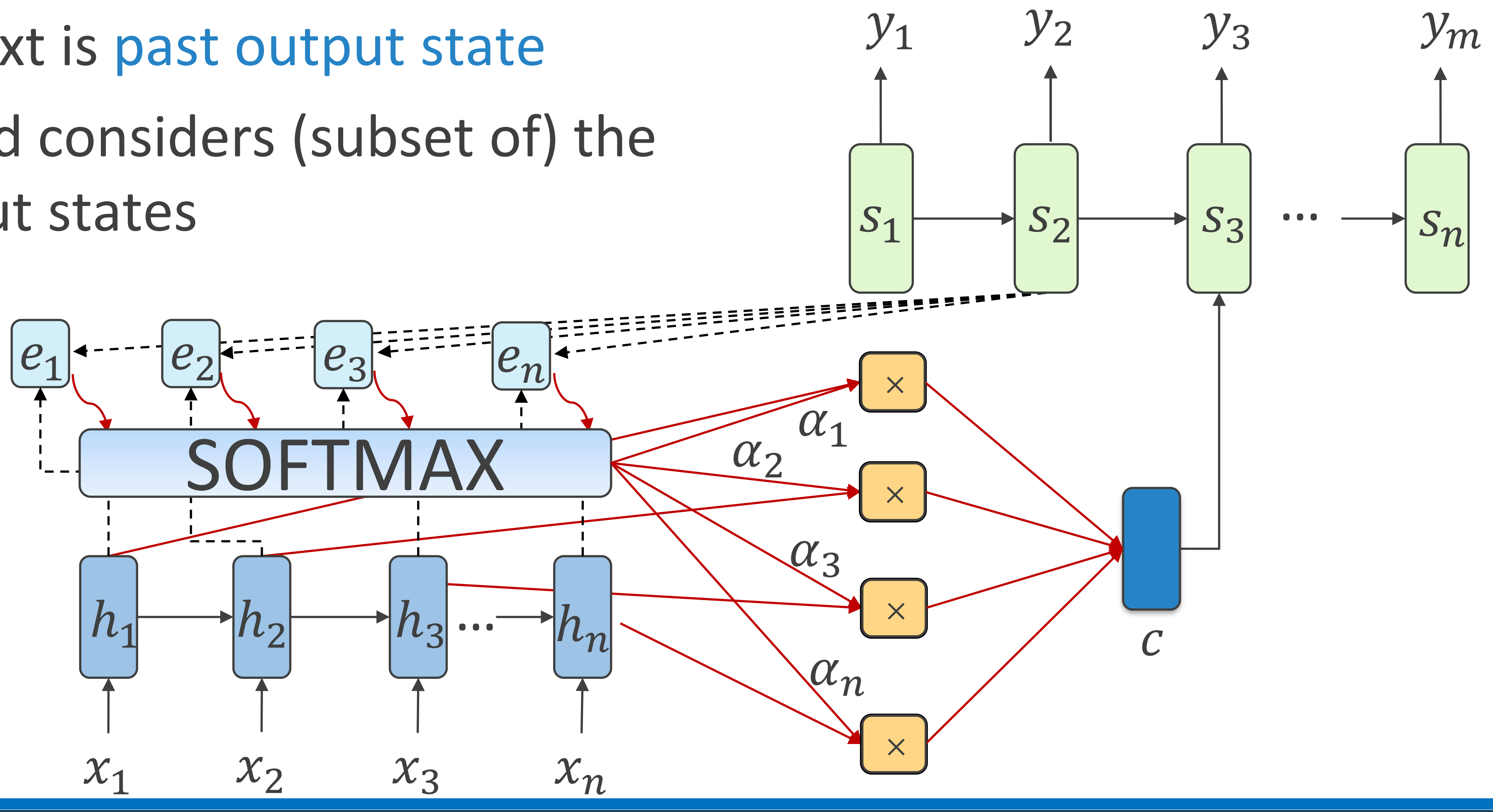


Attention module

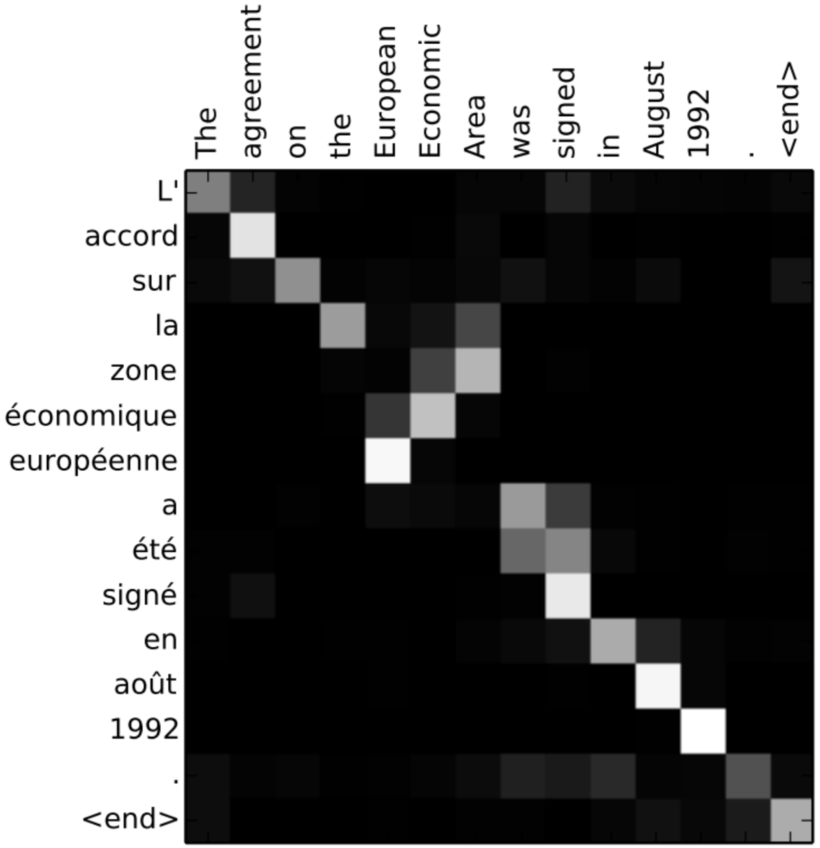
Attention in Seq2Seq

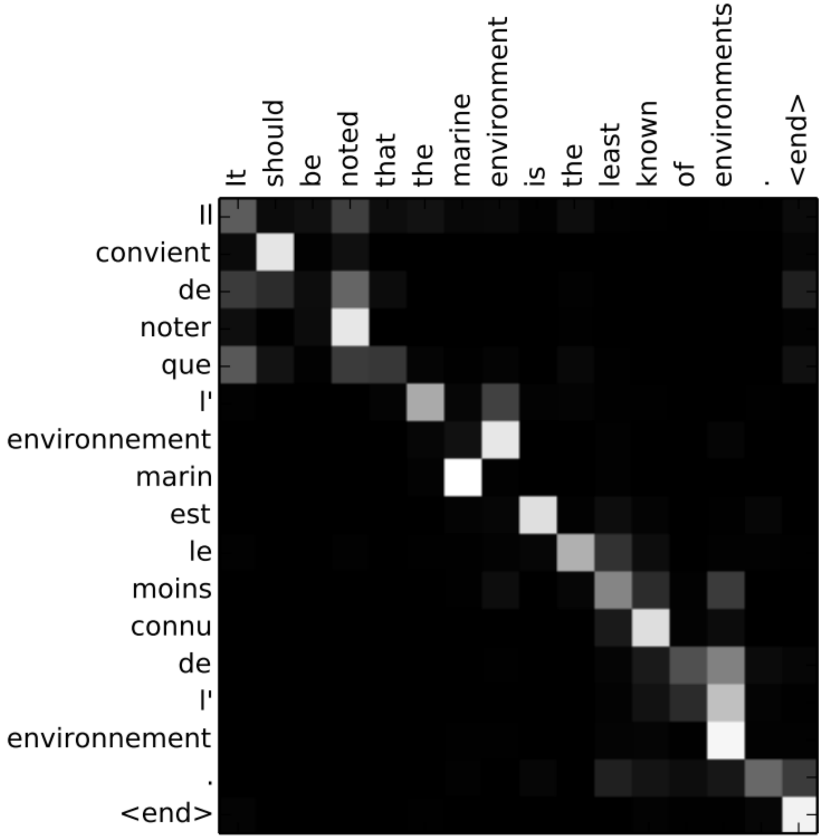
Context is past output state

- Seed considers (subset of) the input states



Learning to Translate with Attention

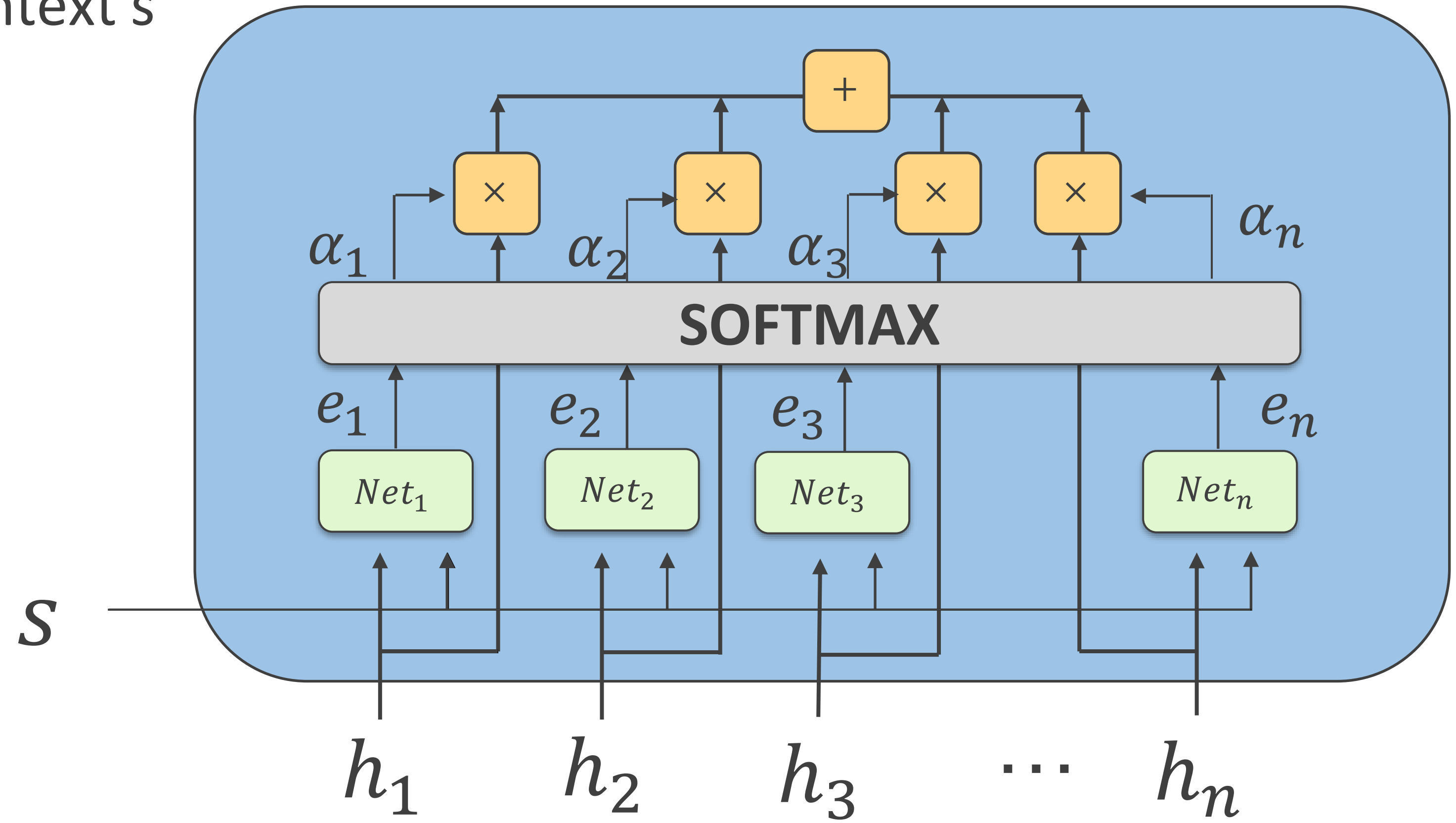




Bahdanau et al, Show, Neural machine translation by jointly learning to align and translate, ICLR 2015

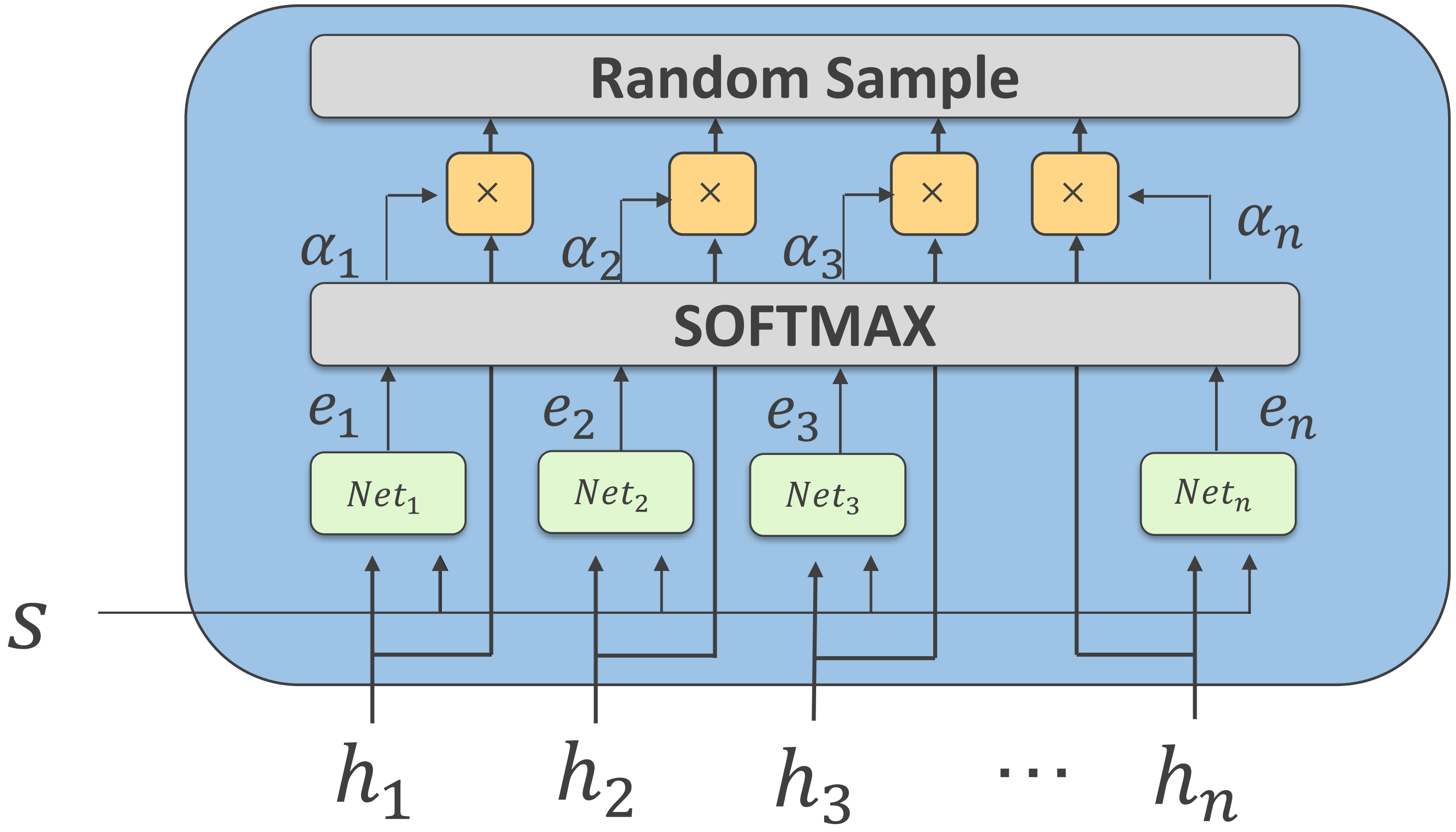
Advanced Attention – Generalize Relevance

This component determines how much each h is correlated/associated with current context s



Advanced Attention – Hard Attention

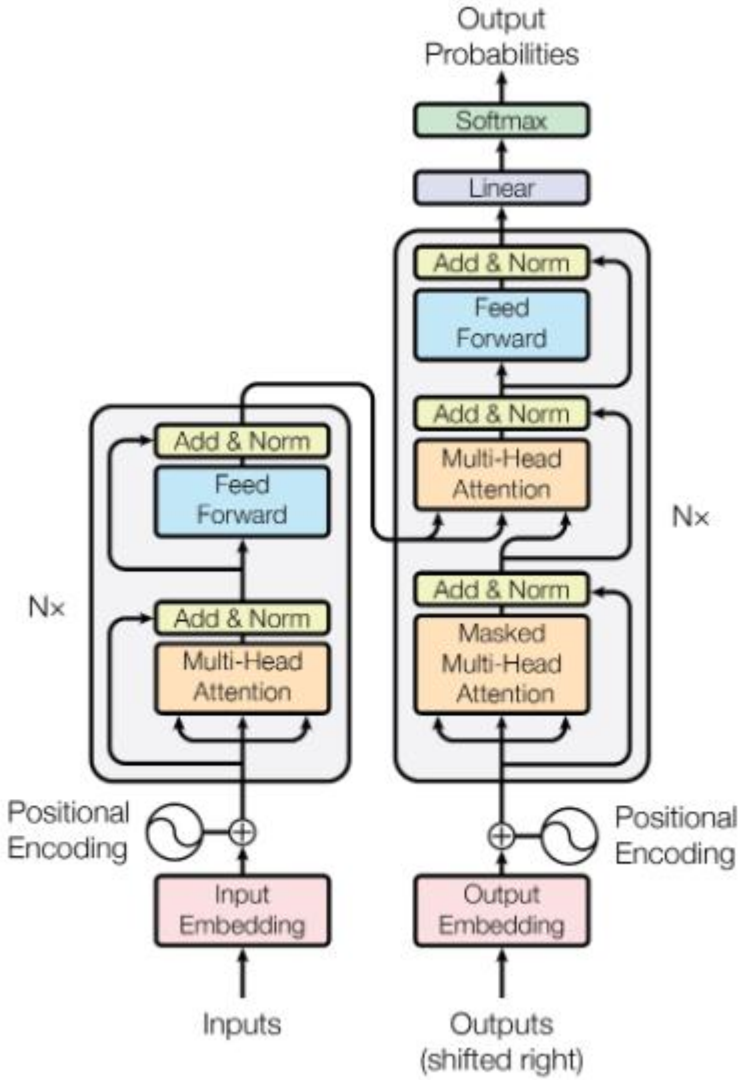
Sample a single encoding using probability α_i



Attention Is All You Need, https://arxiv.org/pdf/1706.03762.pdf

Transformers

- First pure attention-based model
- Self-attention
- No recurrence
- Encoder-decoder architecture

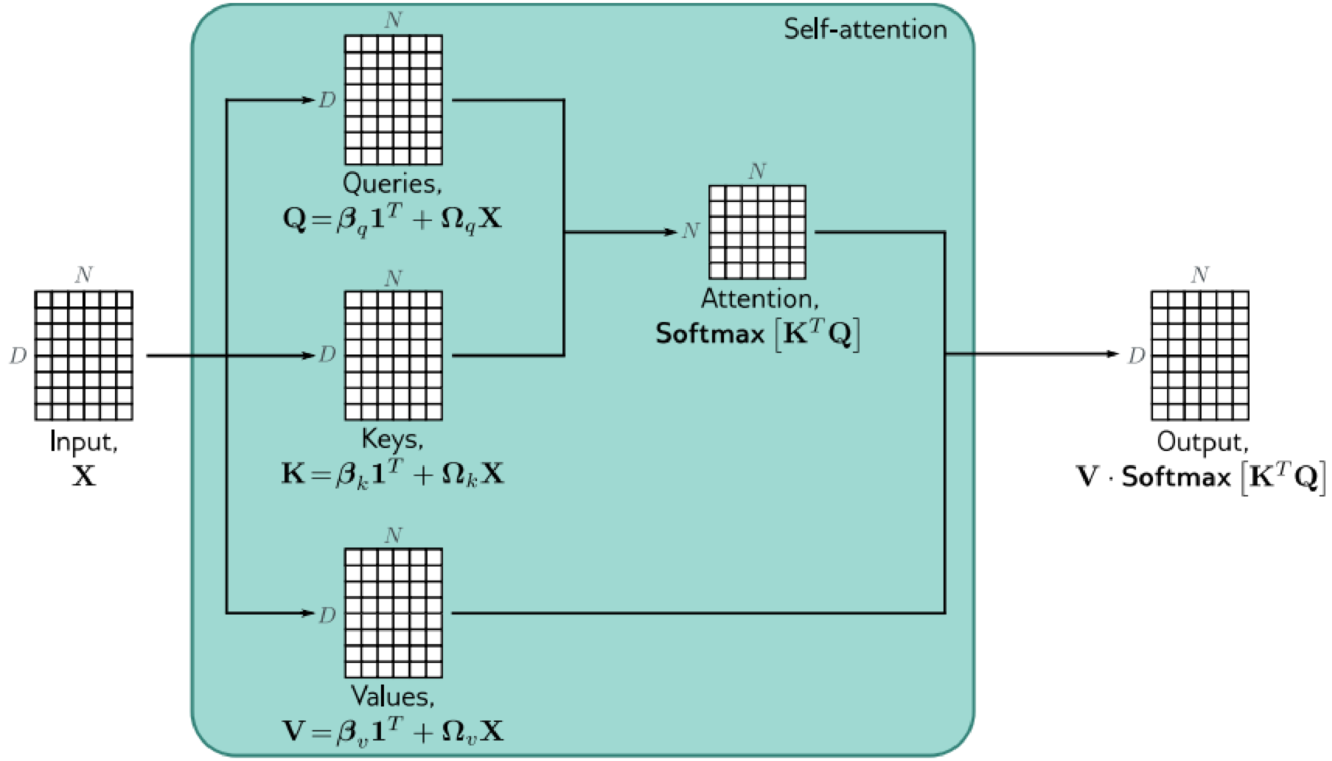


Self Attention

Each element of an input sequence X_i projects into 3 vectors: **query**, **key** and **value**

Scaled self-attention

$$\sum_j softmax_j\left(\frac{Q_i \cdot \boldsymbol{K}^T}{\sqrt{d_k}}\right)V_j$$



Self Attention – K,V,Q Generation

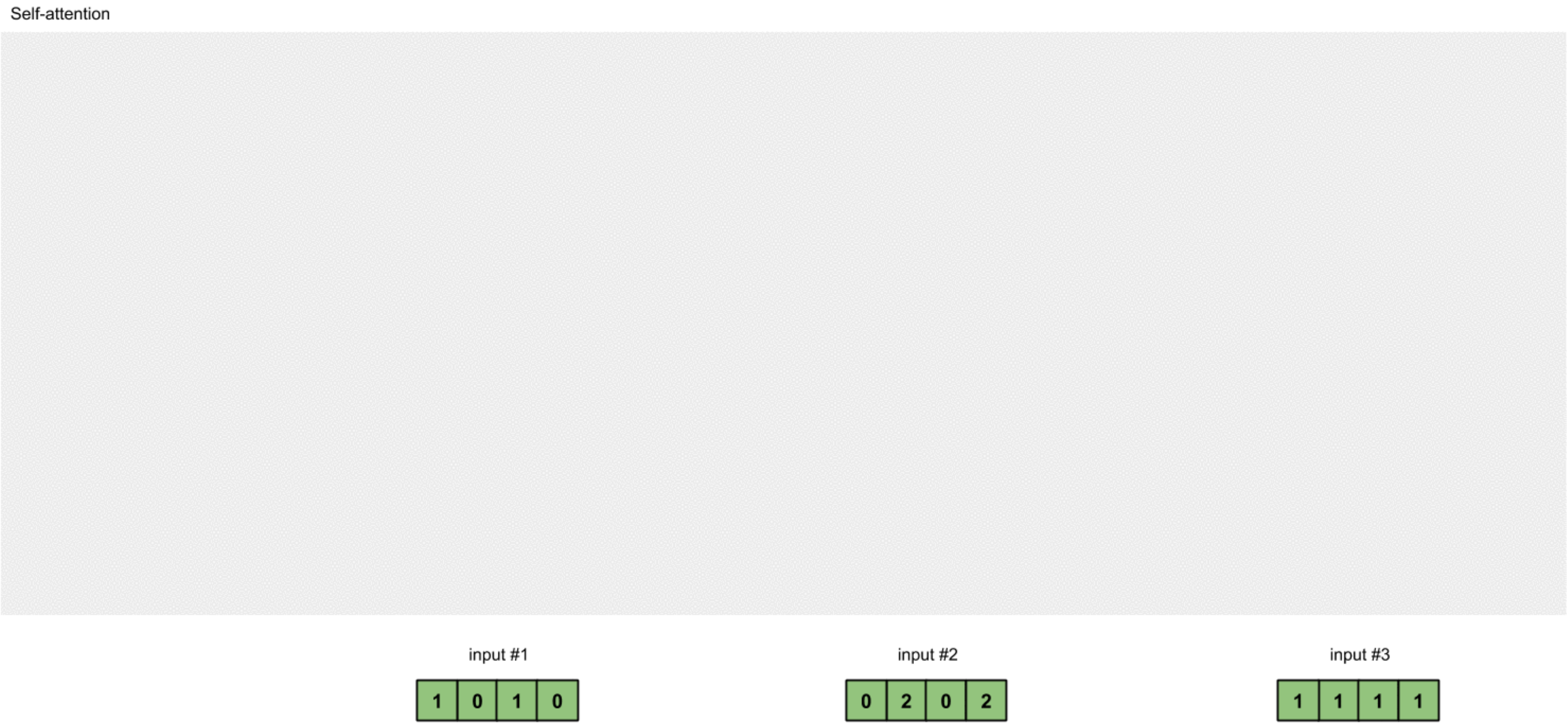
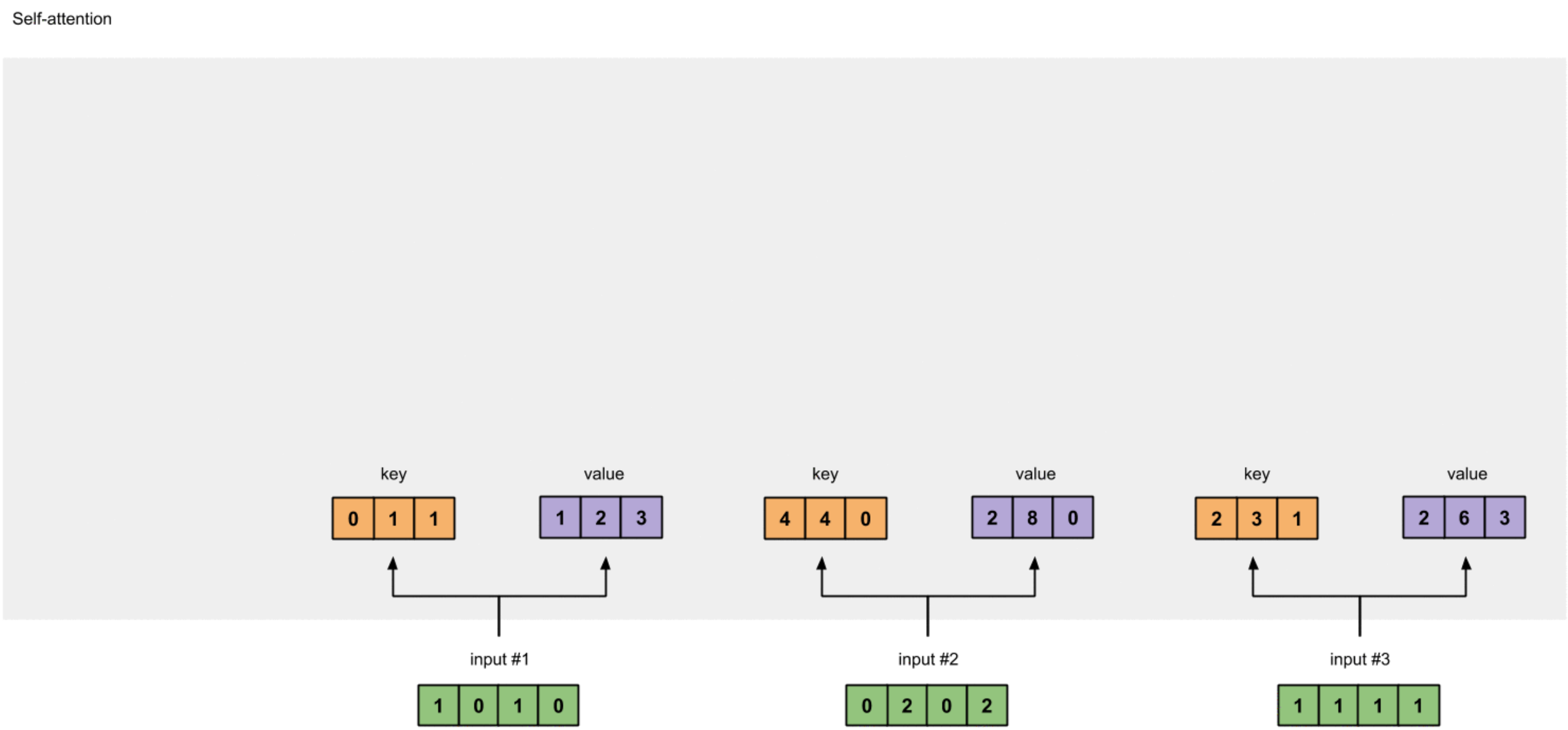


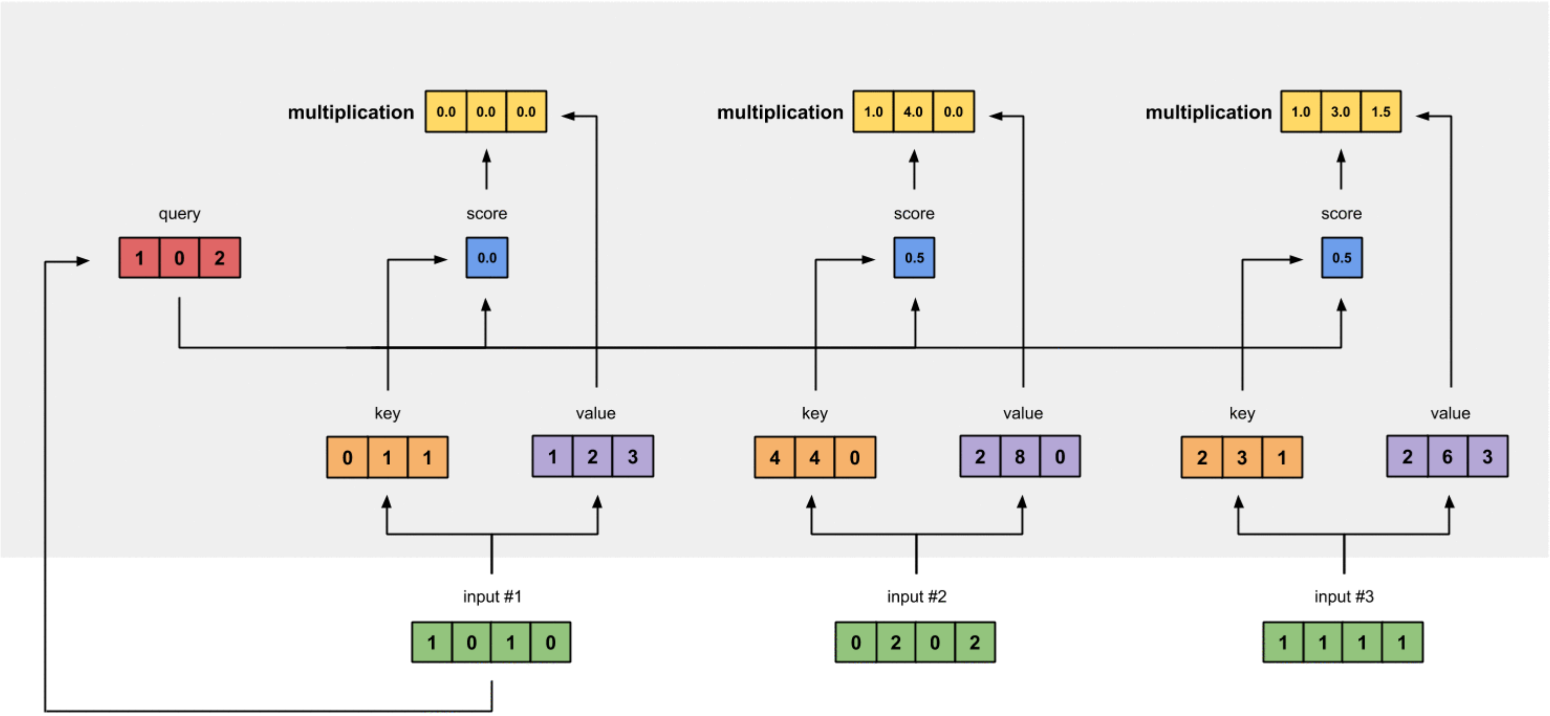
Figure credit to this [article]

Self Attention – Compute Attention Score

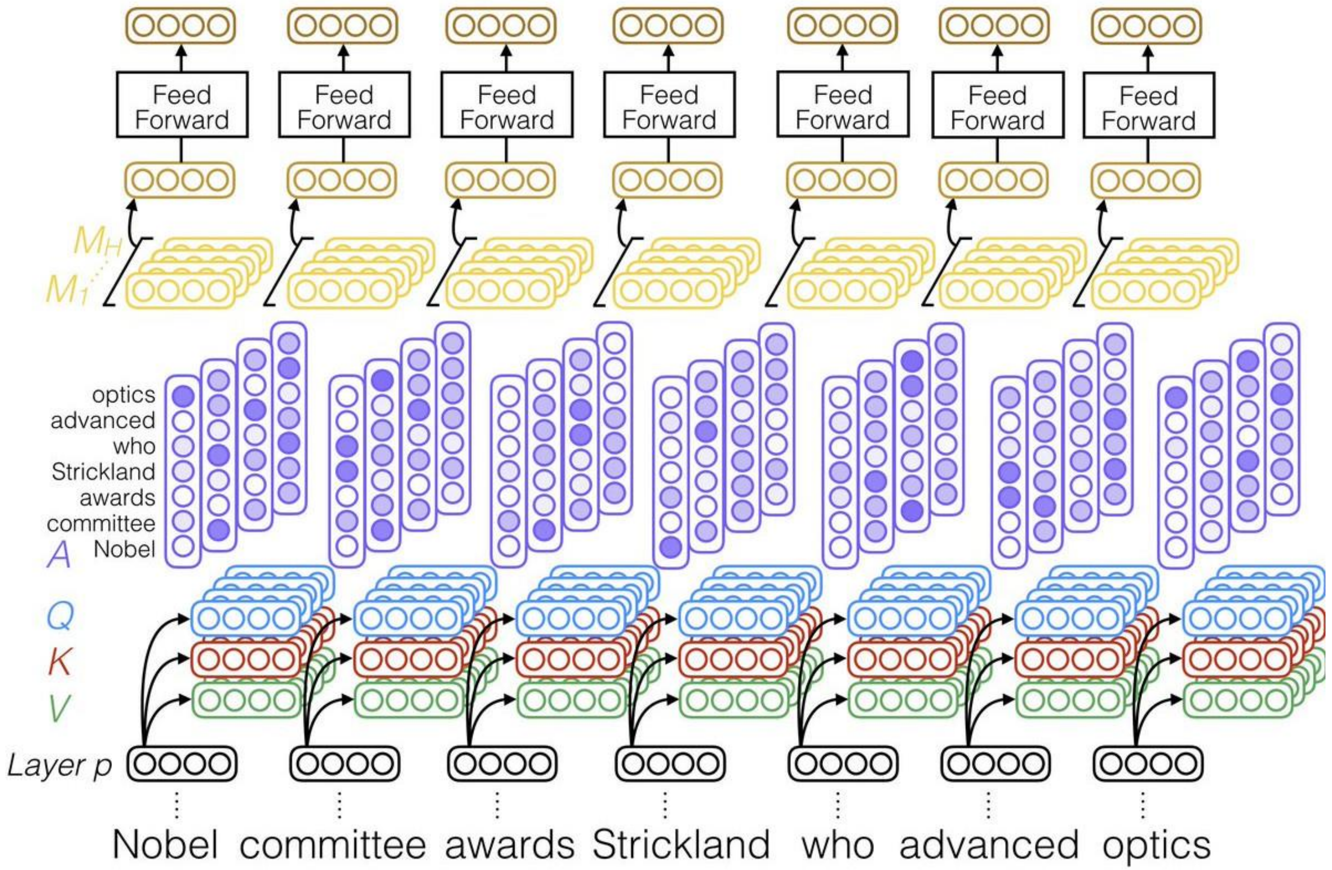


Self Attention – Produce Output

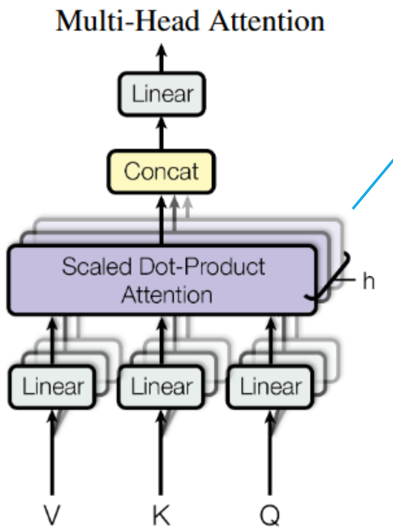
Self-attention



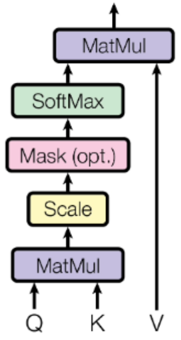
Self Attention – MultiHead



Strubell et al, Linguistically-Informed Self-Attention for Semantic Role Labeling, EMNLP 2018



concat



Is self-attention a good mechanism to model temporal dependencies?

What happens if I randomly shuffle some tokens?

UNIVERSITÀ DI PISA

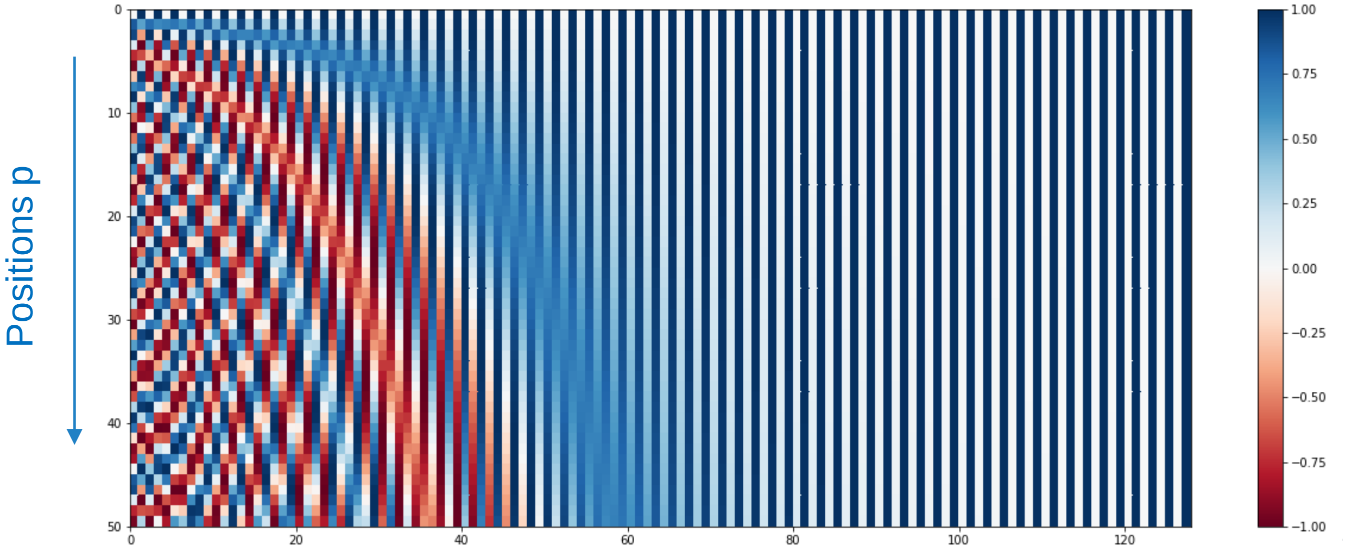
Attention Is All You Need,
https://arxiv.org/pdf/1706.03762.pdf

(Absolute) Positional Encoding

- Self-attention is order-independent
- But in sequences we need ordering information
- word embedding + positional embedding

$$PE(p, 2i) = \sin(p/\overbrace{10000^{2i/d}}^{\text{frequency}})$$
$$PE(p, 2i+1) = cos(p/10000^{2i/d})$$

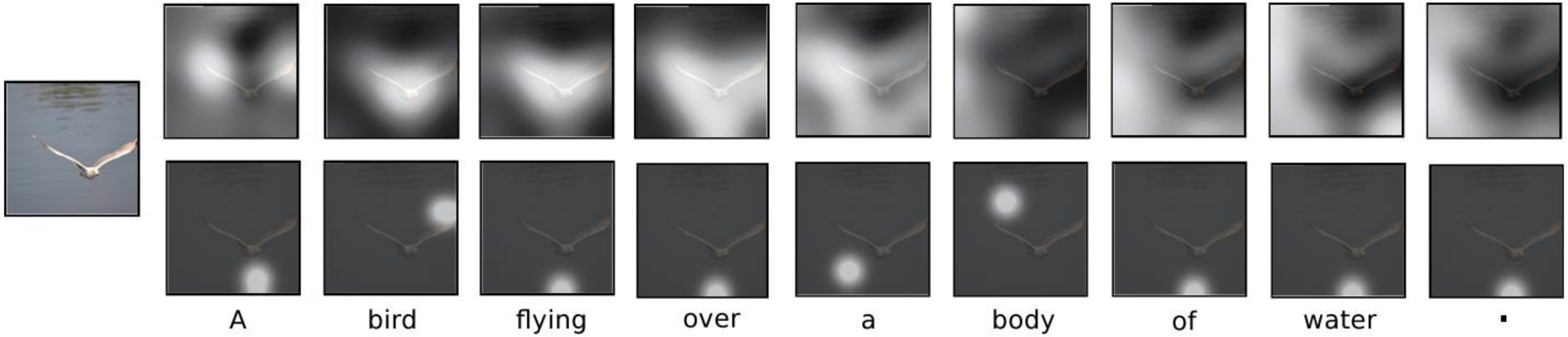
embedding dimensions $i \in \{1, d\}$



Attention in Vision

Attention-Based Captioning – Focus Shifting

Soft Attention



Hard Attention

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

Attention-Based Captioning - Generation

Learns to correlate textual and visual concepts

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.

Helps understanding why the model fails

A large white bird standing in a forest.

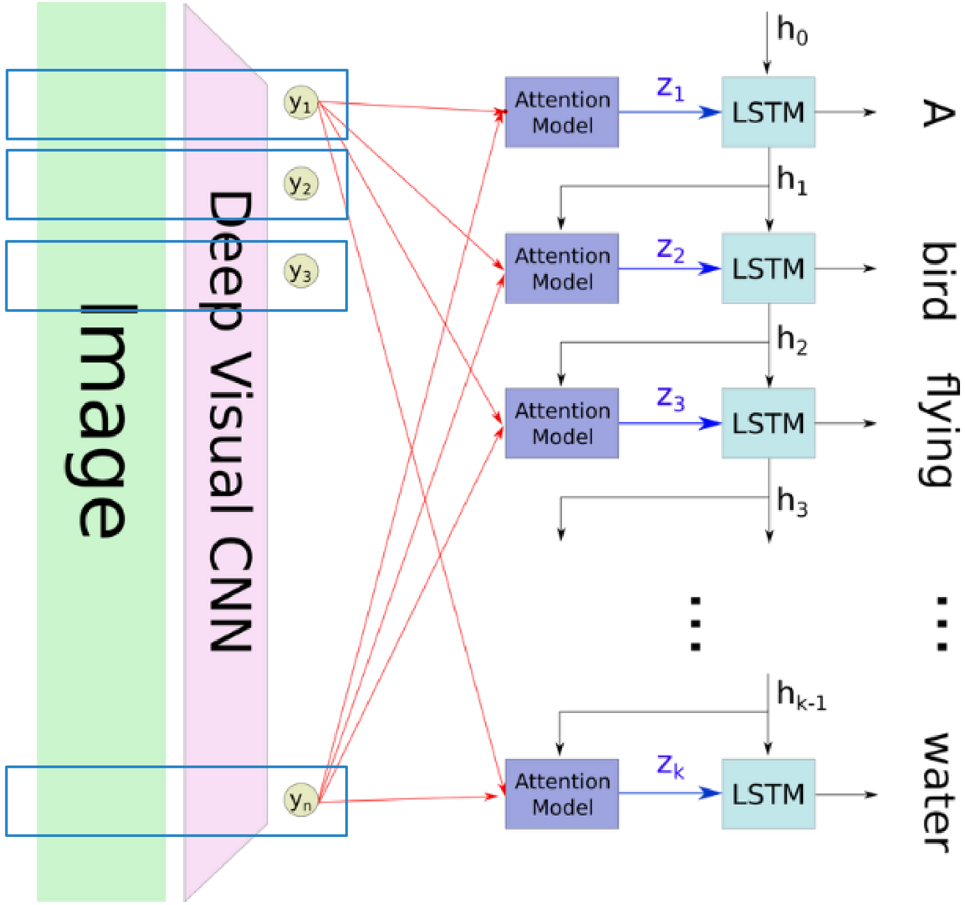
A woman holding a clock in her hand.

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

Attention-Based Captioning – The Model

Encodings associated to *n* image regions

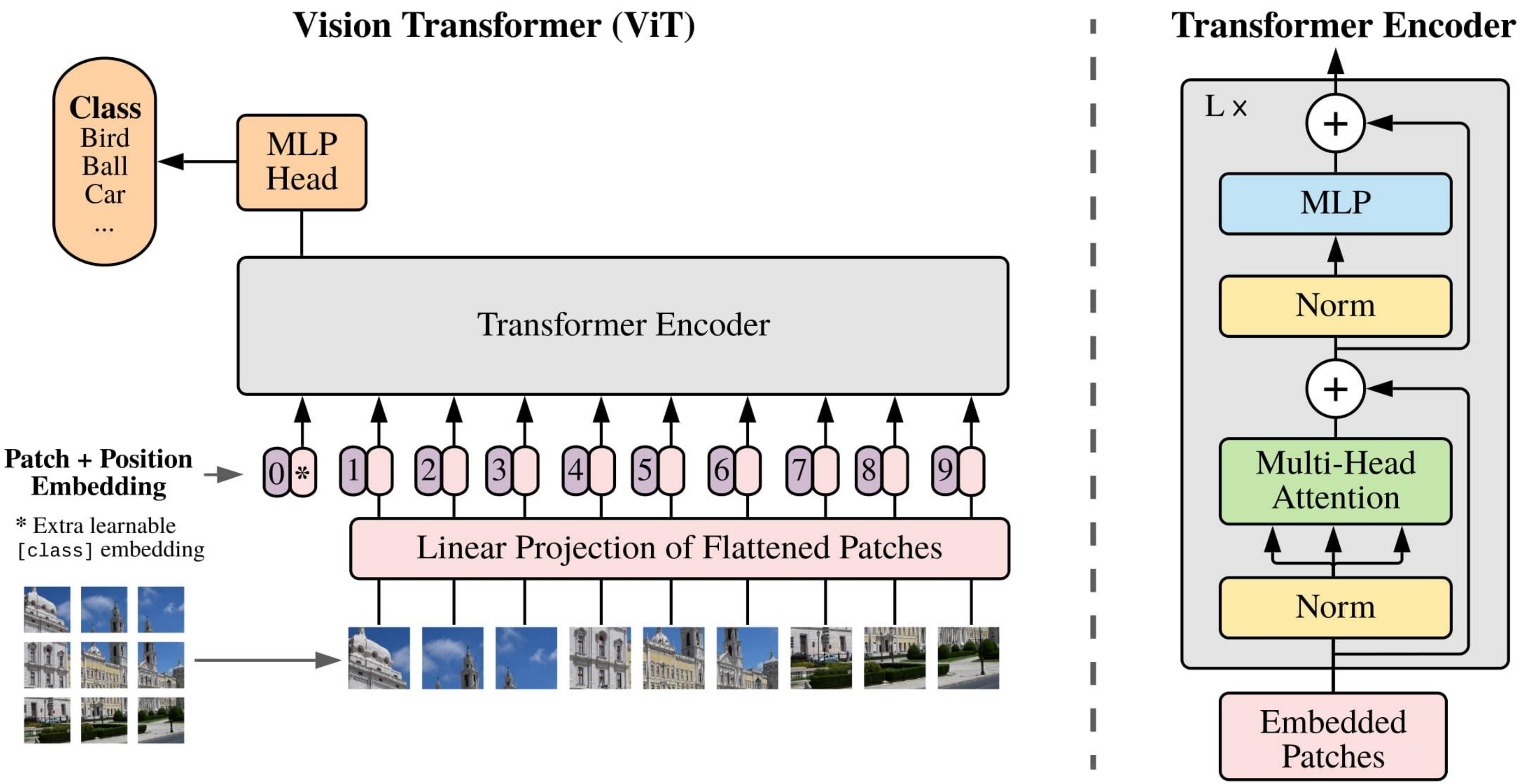
From convolutional layers rather than from fully connected



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

The Vision Transformer (ViT)

A. Dosovitskiy et al, ICLR 2021



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
- Transformers as low-inductive bias architectures
  - Need huge amounts of data to generalize