

INTRODUCTION TO DEEP LEARNING

UPC TelecomBCN Barcelona (4th edition). Spring Edition.



Instructors:



Xavier
Giró-i-Nieto



Ferran
Marqués



Ramon
Morros



Montse
Pardàs



Javier
Ruiz



Elisa
Sayrol



Verònica
Vilaplana

Teaching Assistants:



Gerard
Gallego



Albert
Mosella

Day 6 Lecture 3

Attention Mechanisms



Xavier Giro-i-Nieto



@DocXavi



xavier.giro@upc.edu



Associate Professor
Universitat Politècnica de Catalunya



Versió extesa



Deep Learning 4

Mecanismes d'Atenció

DEEP AND REINFORCEMENT LEARNING
UPC TelecomBCN Barcelona (3rd edition). Autumn 2020.



Instructors:

Margarita Camara	Víctor Campero	Xavier Giro-i-Nieto	Juan José Nieto	Joop Vilal	Guests:	Fernan Alet (MRT)	Carlos Floresca (Cavallart)	Ignasi Olivers (Berkoleys)	Jonki Tarrés (KSC)	Oriol Vinyals (DeepMind)
------------------	----------------	---------------------	-----------------	------------	---------	-------------------	-----------------------------	----------------------------	--------------------	--------------------------

[course site]



Xavier Giro-i-Nieto
Associate Professor
Universitat Politècnica de Catalunya

 [@DocXavi](#)
 xavier.giro@upc.edu



Xavier Giró (en català)
[UPC AA2 2020](#)

Acknowledgments



Amaia Salvador

PhD (2019)
Universitat Politècnica de Catalunya



Marta R. Costa-jussà

Associate Professor
Universitat Politècnica de Catalunya

DEEP LEARNING FOR COMPUTER VISION
Summer Seminar UPC TelecomBCN, 4 - 8 July 2016

Instructors: [list of names]
Organizers: [list of organizations]

+ info: [TelecomBCN.DeepLearning.Barcelona](https://telecombcn-dl.github.io/2016-id/)
[course site]

Day 4 Lecture 6

Attention Models



[Amaia Salvador](#)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
Barcelona School of Informatics
Department of Signal Theory
and Communications
Image Processing Group

[DLCV 2016]

INTRODUCTION TO DEEP LEARNING
Winter School at UPC TelecomBCN Barcelona, 23-30 January 2018

Instructors: [list of names]
Organizers: [list of organizations]
Supporters: [list of supporters]

+ Info: <https://telecombcn-dl.github.io/2018-id/>
[course site]

#DLUPC

Day 3 Lecture 4
Attention-based mechanisms

SLIDES ADAPTED FROM Graham NEUBIG's lectures

Marta R. Costa-jussà
marta.mari@upc.edu

Ramón y Cajal Researcher
Universitat Politècnica de Catalunya
Technical University of Catalonia

UPC

[IDL 2018]

Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. Attention mechanisms
6. Study case: Image captioning



RNNs vs Attention

Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. Attention mechanisms
6. Study case: Image captioning



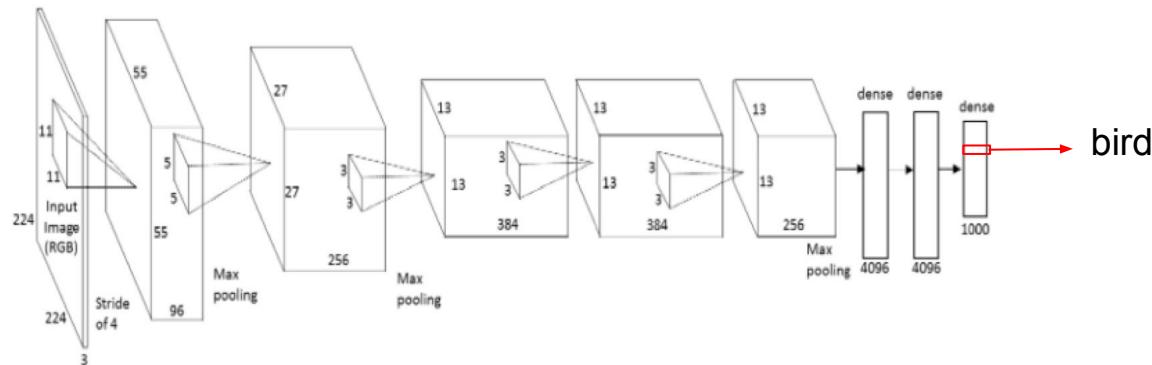
RNNs vs Attention

Motivation: Vision

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



Image:
 $H \times W \times 3$



Motivation: Vision

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

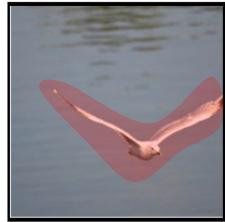
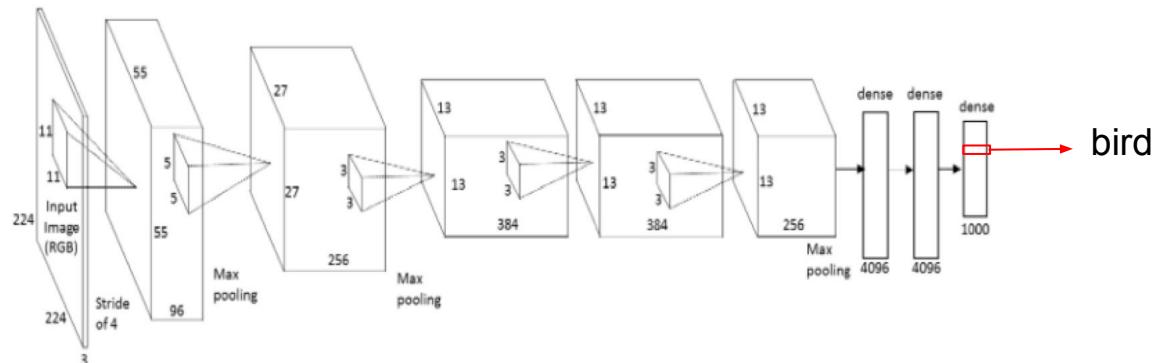


Image:
 $H \times W \times 3$



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

Motivation: Automatic Speech Recognition (ASR)

Transcribed letters correspond with just a few phonemes.

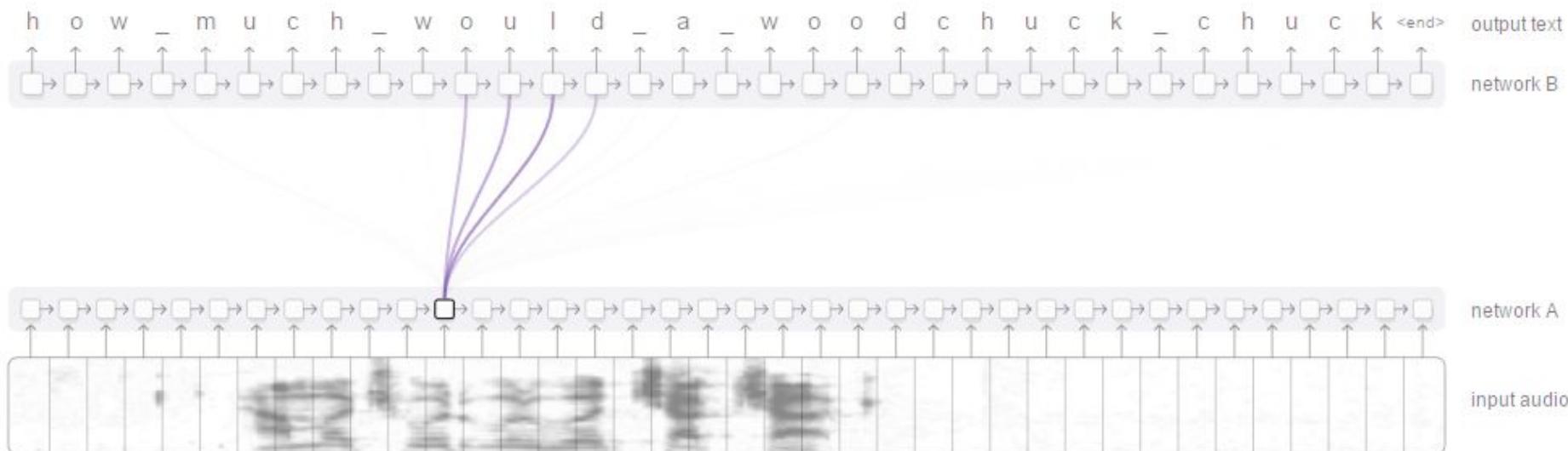
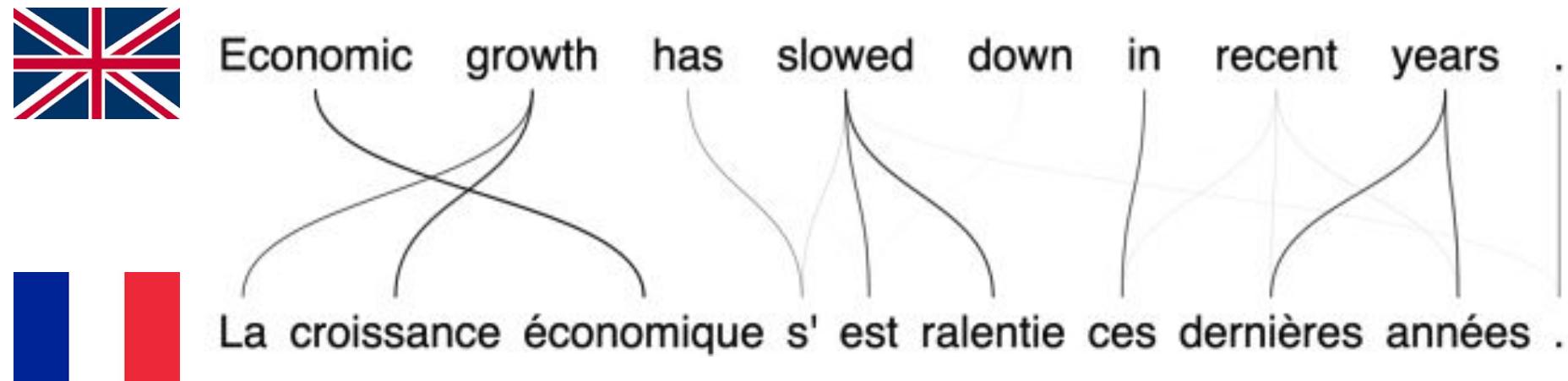


Figure: distill.pub

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. [Listen, attend and spell: A neural network for large vocabulary conversational speech recognition](#). ICASSP 2016.

Motivation: Neural Machine Translation (NMT)

The translated words are often related to a subset of the words from the source sentence.



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

Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. Attention mechanisms
6. Study case: Image captioning

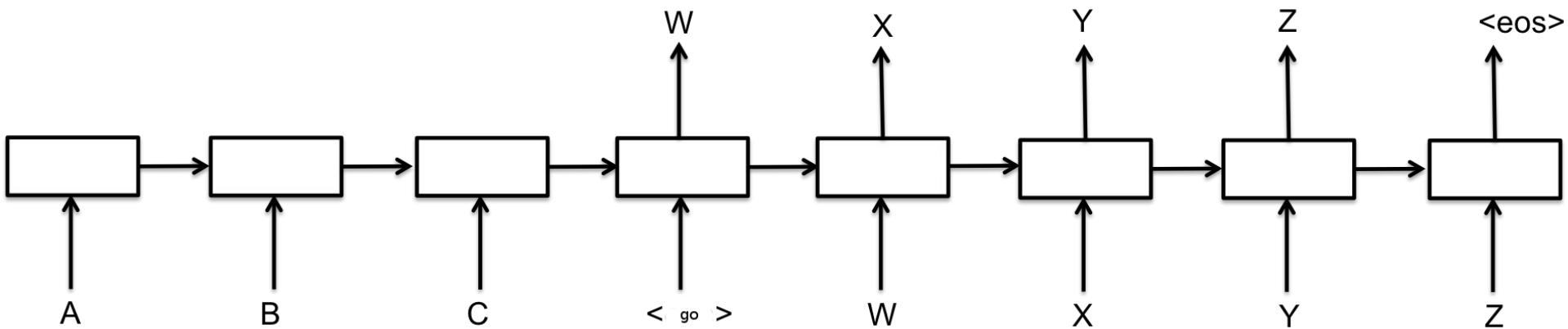


RNNs vs Attention

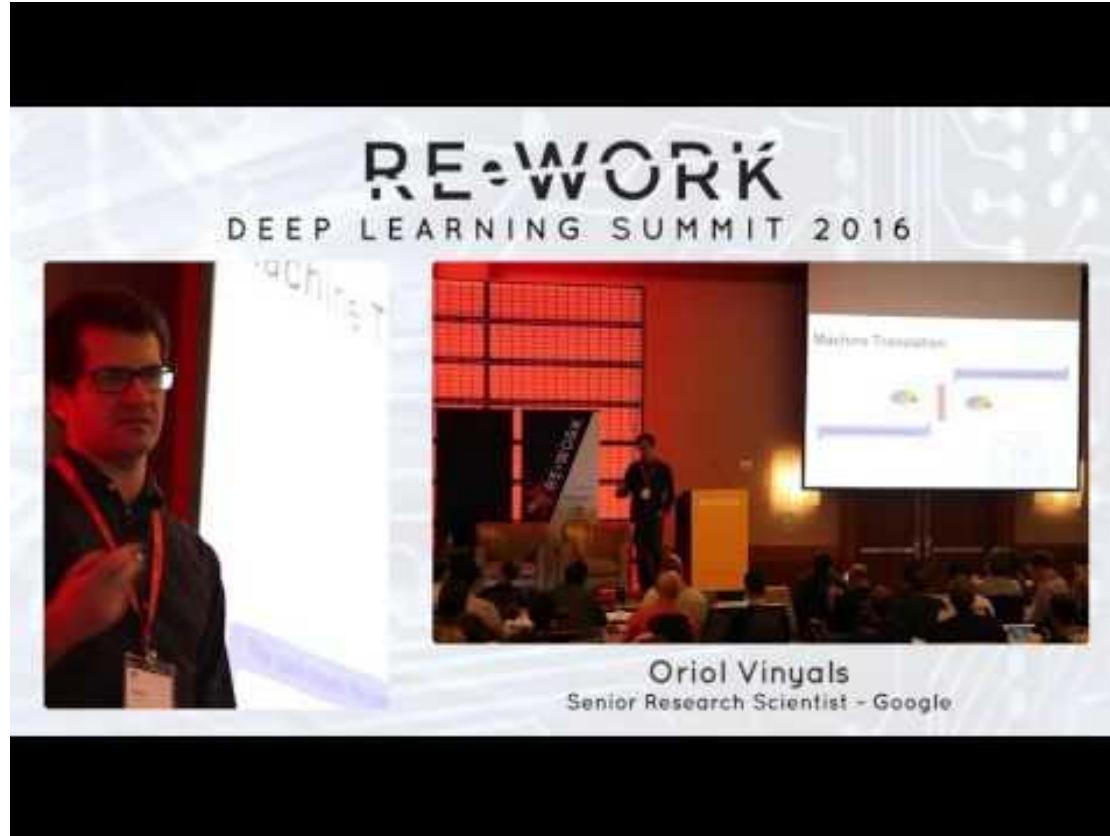
Seq2Seq with RNN

The **Seq2Seq** scheme can be implemented with RNNs:

- trigger the output generation with an input **<go>** symbol.
- the predicted word at timestep t , becomes the input at $t+1$.



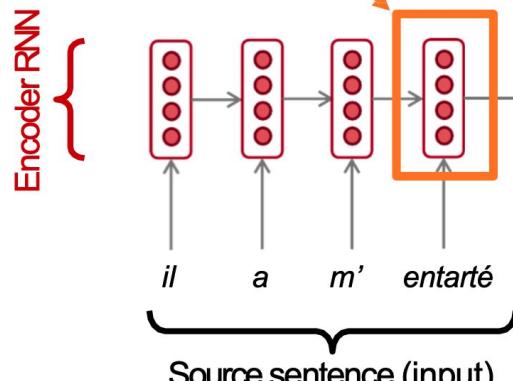
Seq2Seq with RNN



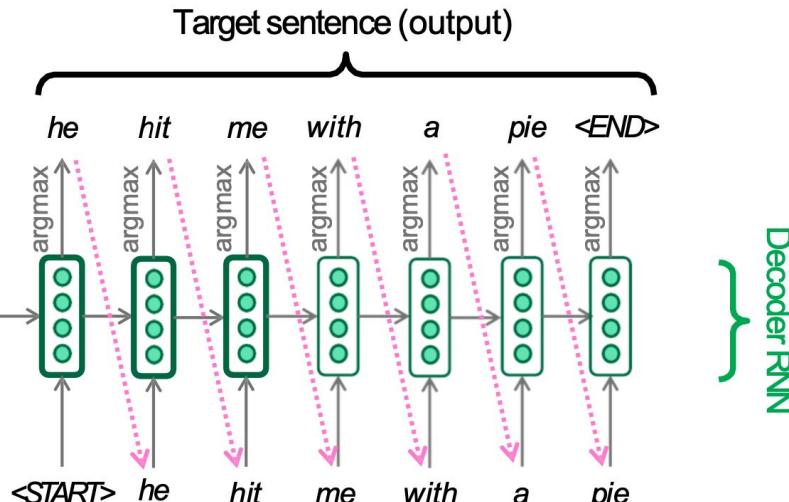
Seq2Seq for NMT

The sequence-to-sequence model

Encoding of the source sentence.
Provides initial hidden state
for Decoder RNN.



Encoder RNN produces
an encoding of the
source sentence.



Decoder RNN is a Language Model that generates target sentence, conditioned on encoding.

Note: This diagram shows test time behavior:
decoder output is fed in as next step's input

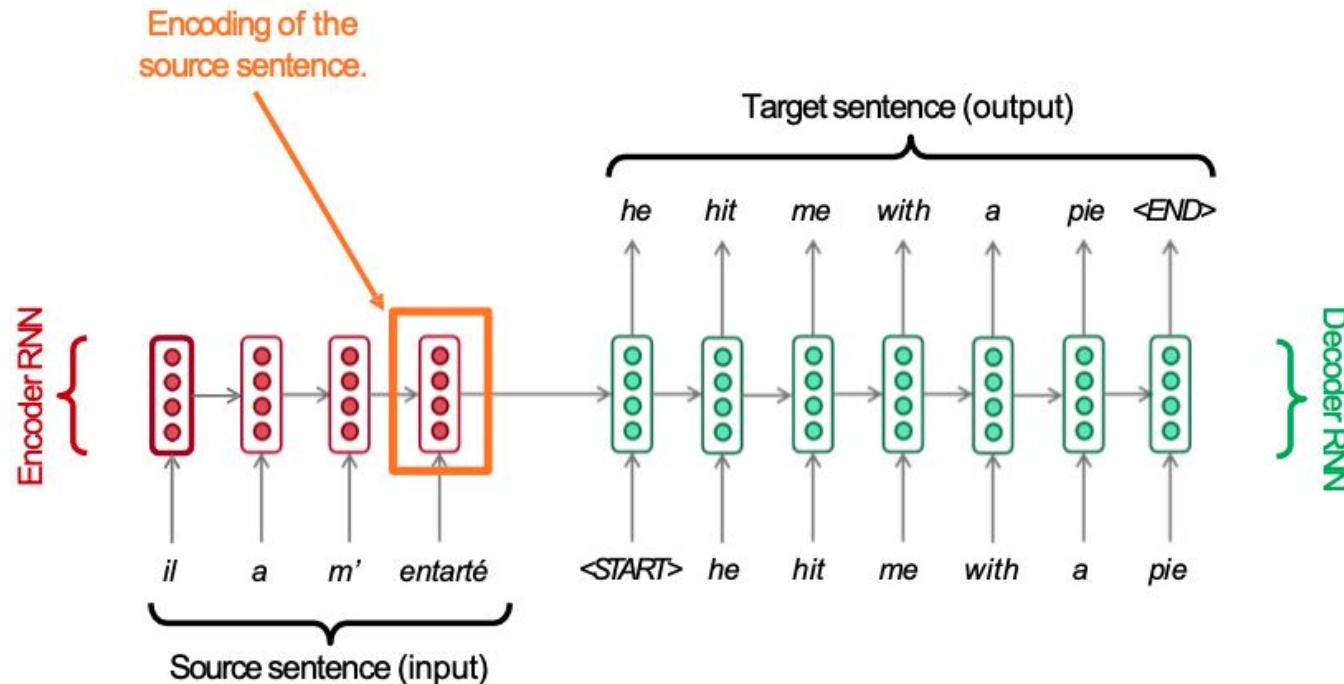
Seq2Seq for NMT



Figure: Jay Alammar, "[Visualizing A Neural Machine Translation Model](#)"

Seq2Seq for NMT

How does the length of the source sentence affect the size (d) of its encoding ?



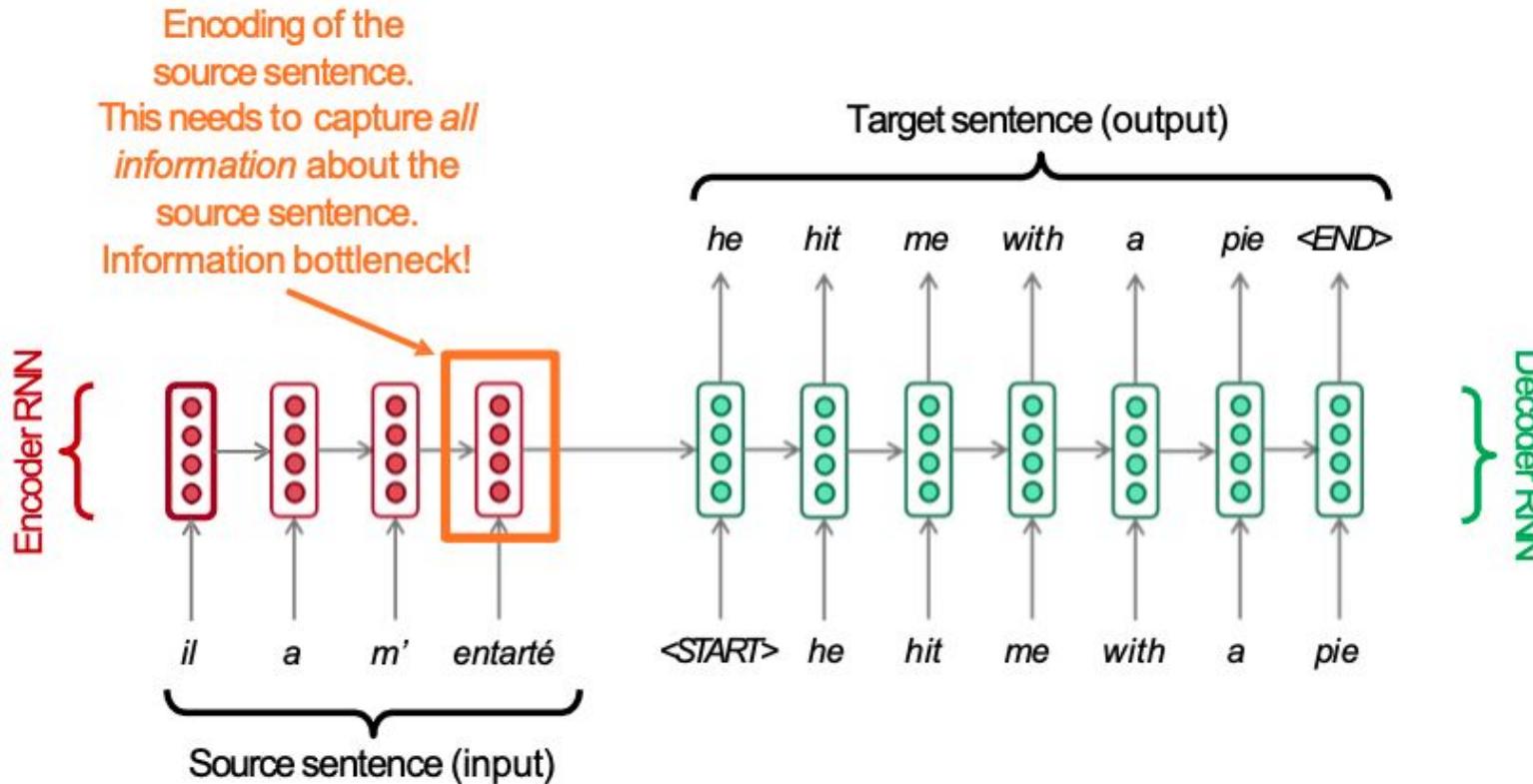
Seq2Seq for NMT

How does the length of the input sentence affect the size (d) of its encoded representation ?



"You can't cram the meaning of
a whole %&!\$# sentence into a
single \$&!#* vector!"

Seq2Seq for NMT: Information bottleneck



Outline

1. Motivation
2. Seq2Seq
- 3. Key, Query and Value**
4. Seq2Seq + Attention
5. Attention mechanisms
6. Study case: Image captioning

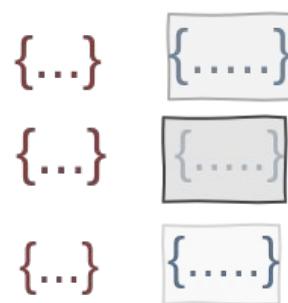


RNNs vs Attention

Query, Keys & Values

Databases store information as pair of **keys and values (K,V)**.

Example:



<Keys> <Values>

myfile.pdf

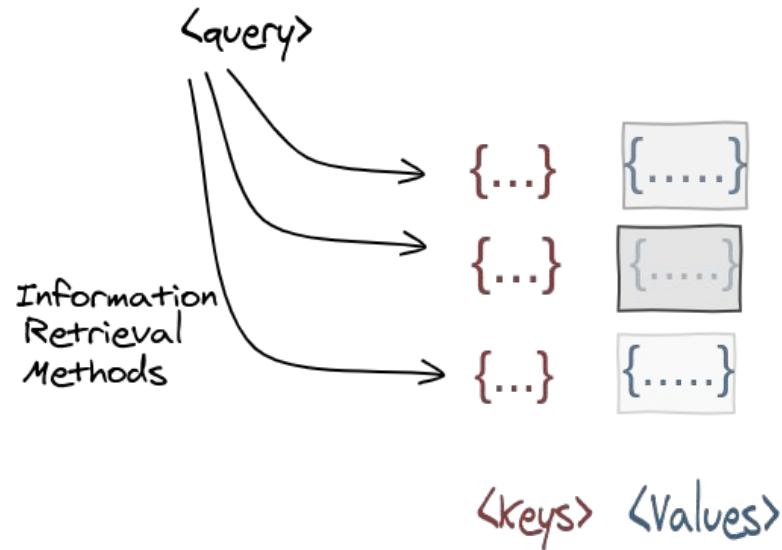


<Key>

<Value>

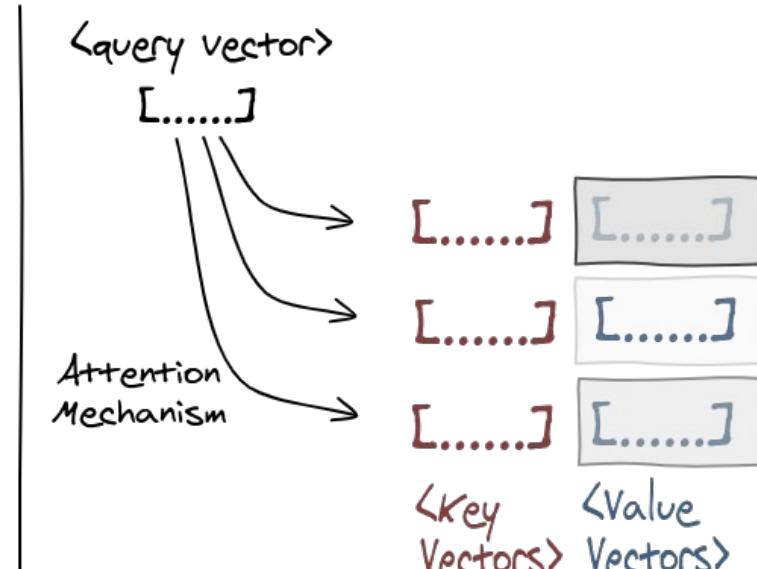
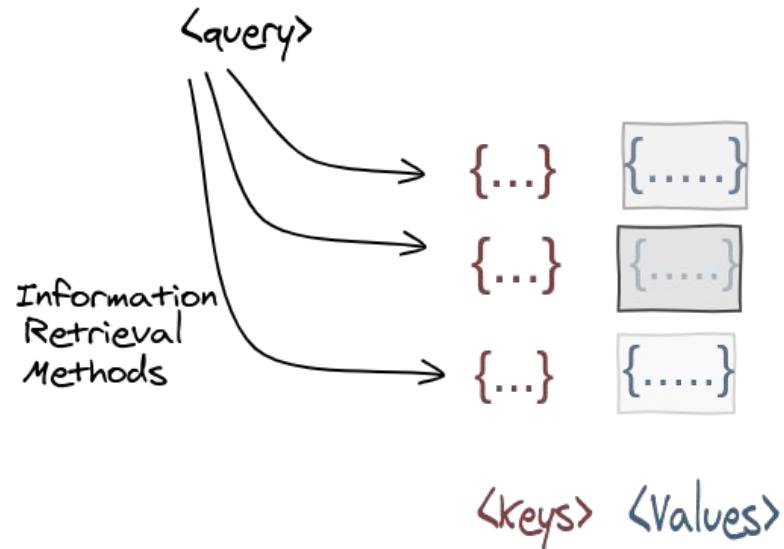
Query, Keys & Values

The (K,Q,V) terminology used to retrieve information from databases is adopted to formulate attention.



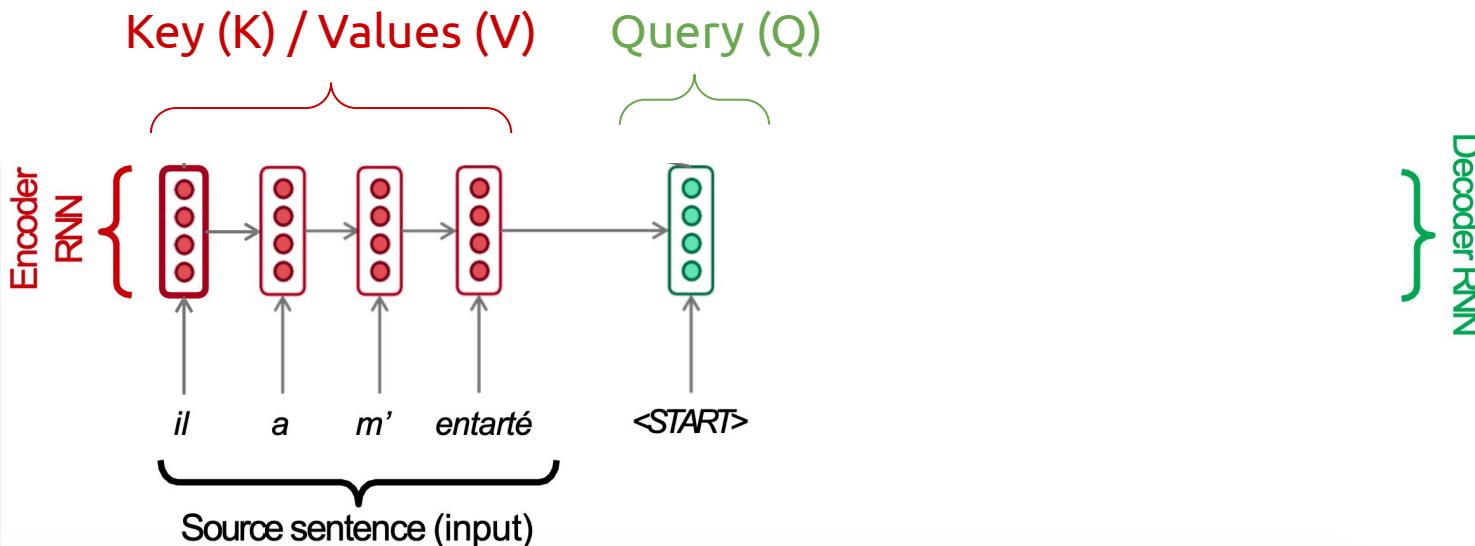
Query, Keys & Values

Attention is a mechanism to compute a context vector (c) for a **query (Q)** as a weighted sum of **values (V)**.



Query, Keys & Values

Usually, both keys and values correspond to the encoder states.



Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. **Seq2Seq + Attention**
5. Attention mechanisms
6. Study case: Image captioning



RNNs vs Attention

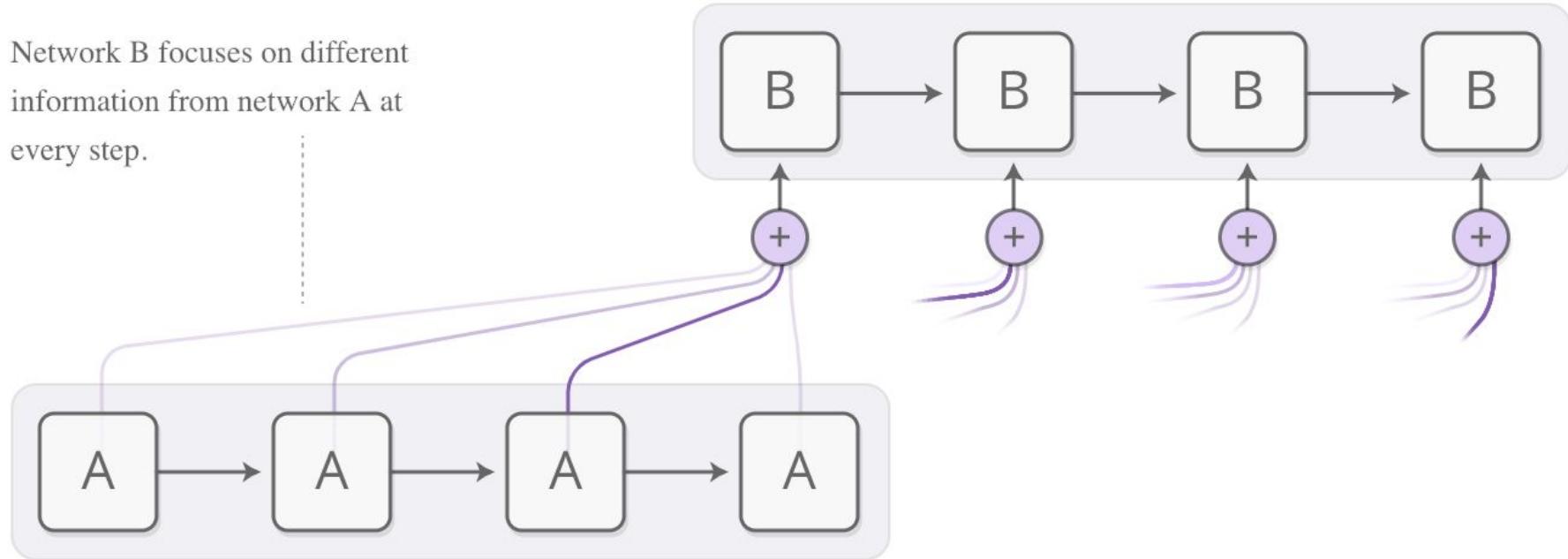
Seq2Seq + Attention



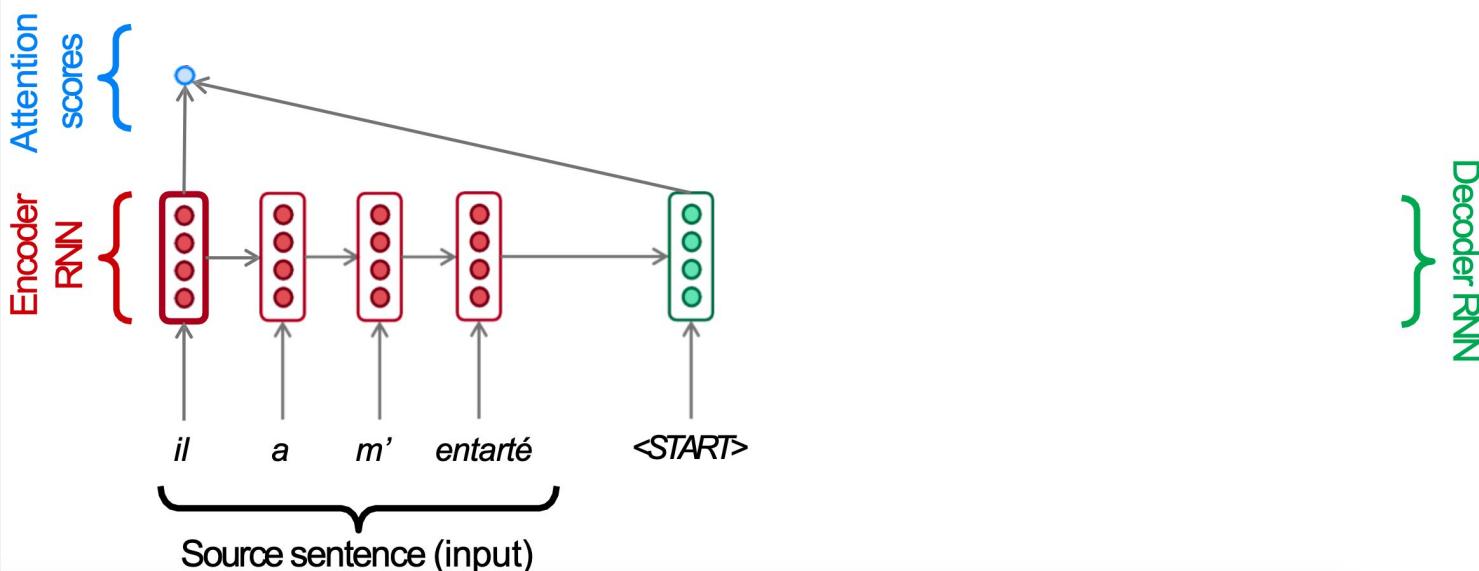
Figure: Jay Alammar, "[Visualizing A Neural Machine Translation Model](#)" (2018)

Seq2Seq + Attention

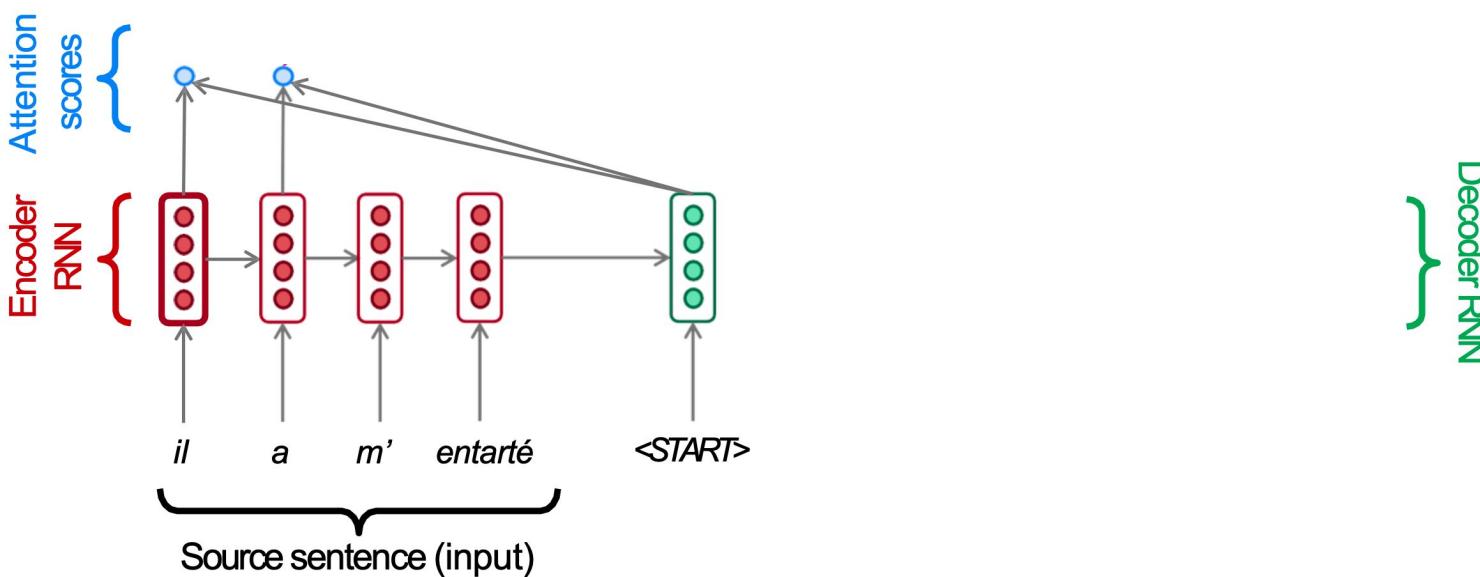
Network B focuses on different information from network A at every step.



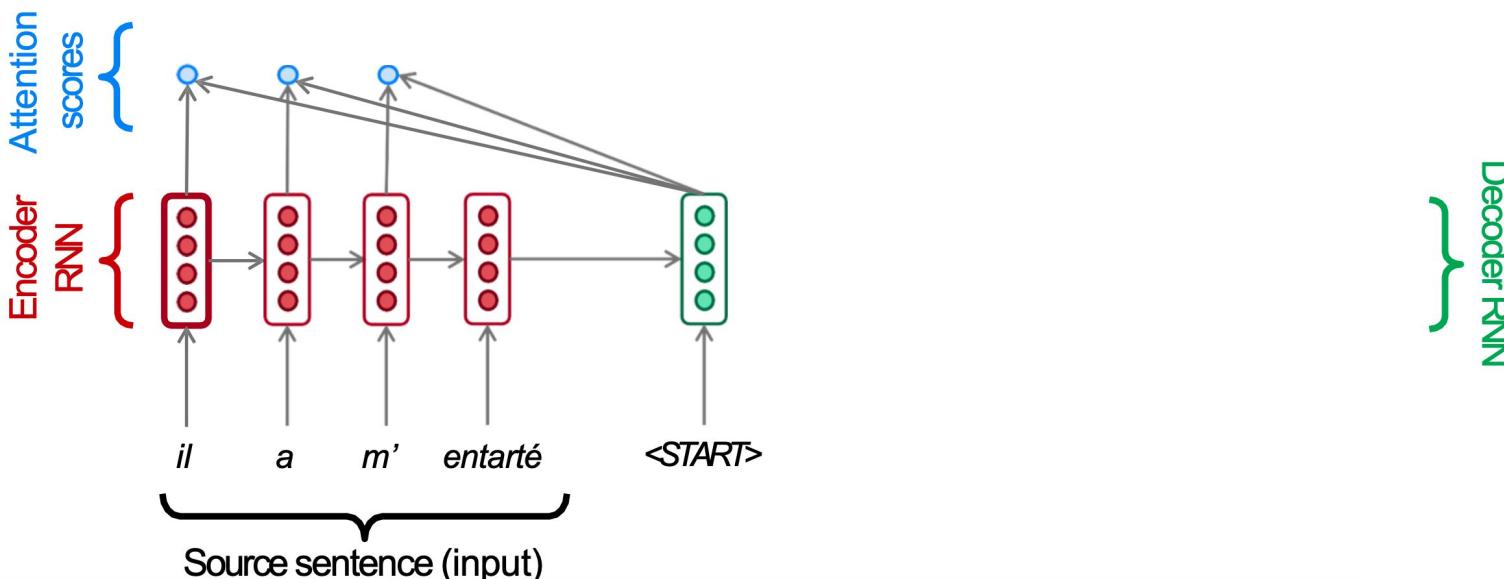
Seq2Seq + Attention: NMT



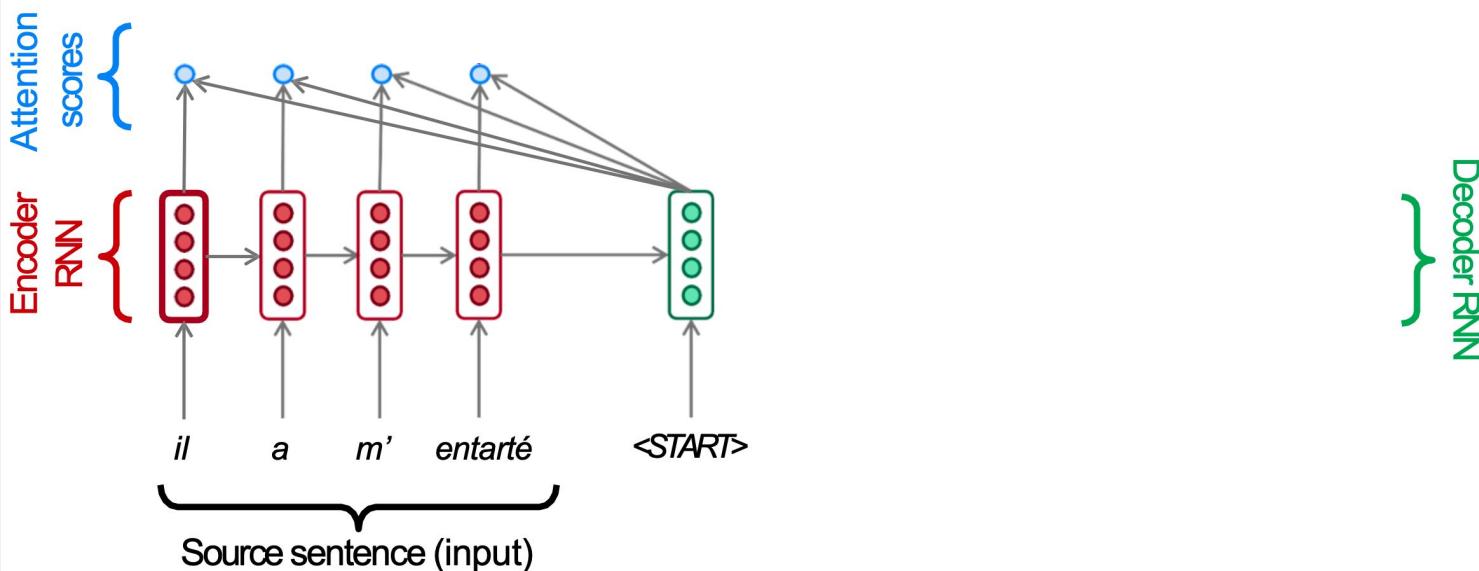
Seq2Seq + Attention: NMT



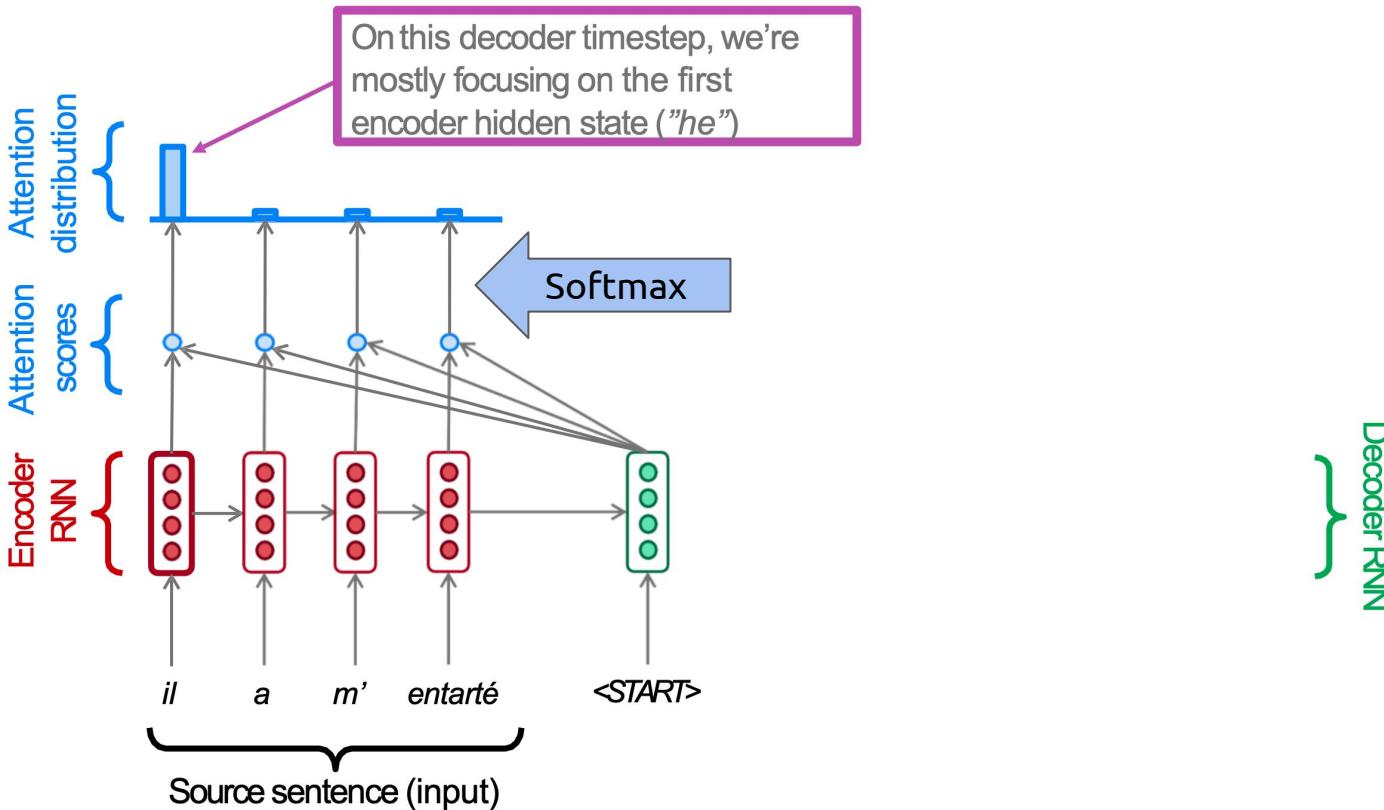
Seq2Seq + Attention: NMT



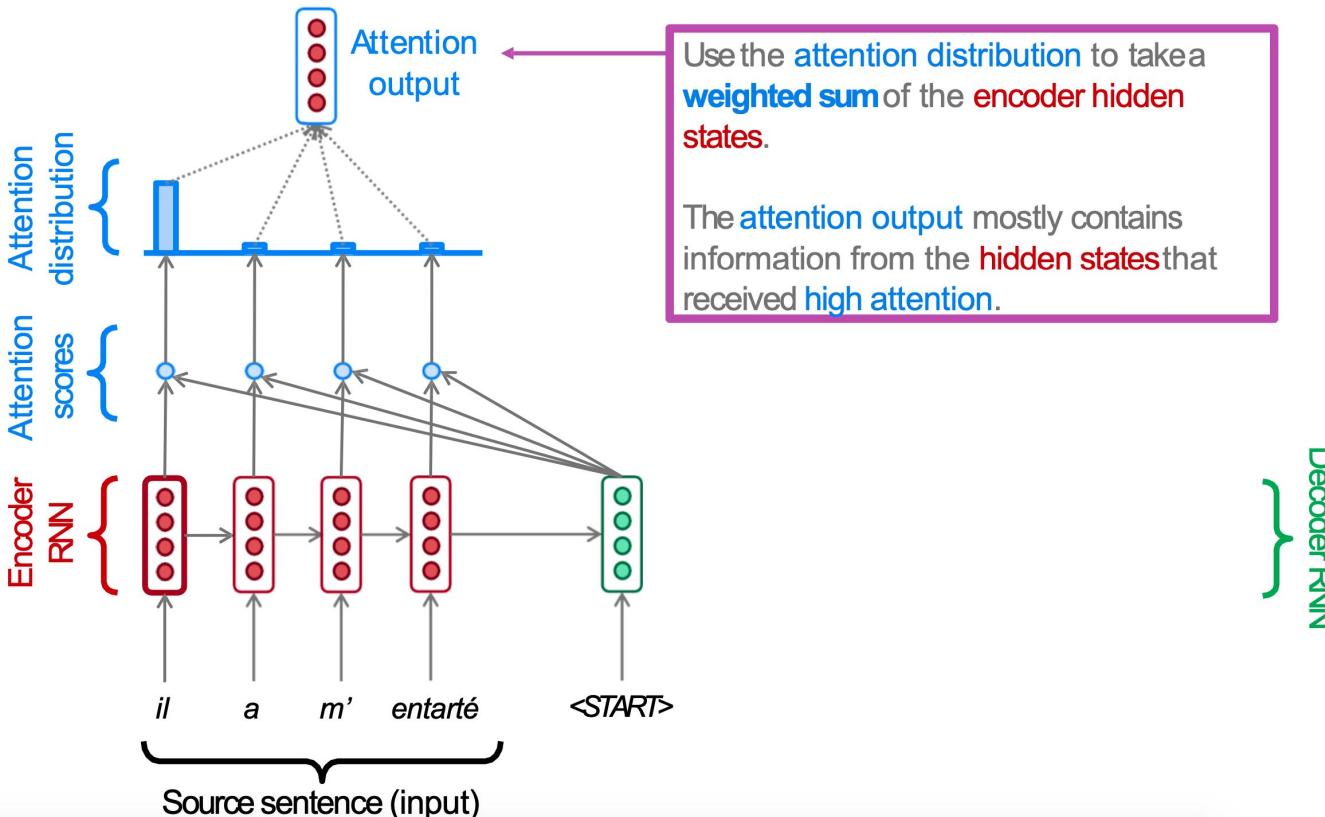
Seq2Seq + Attention: NMT



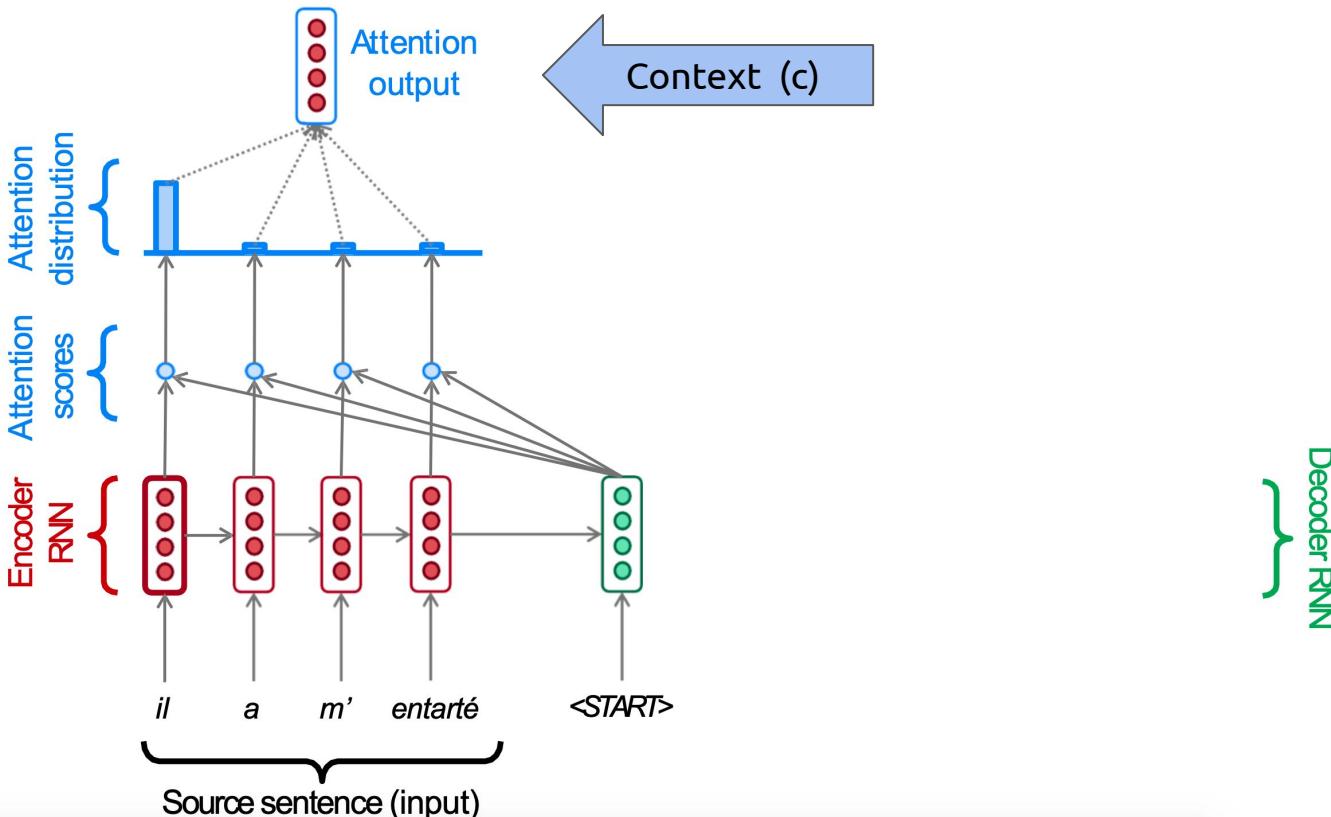
Seq2Seq + Attention: NMT



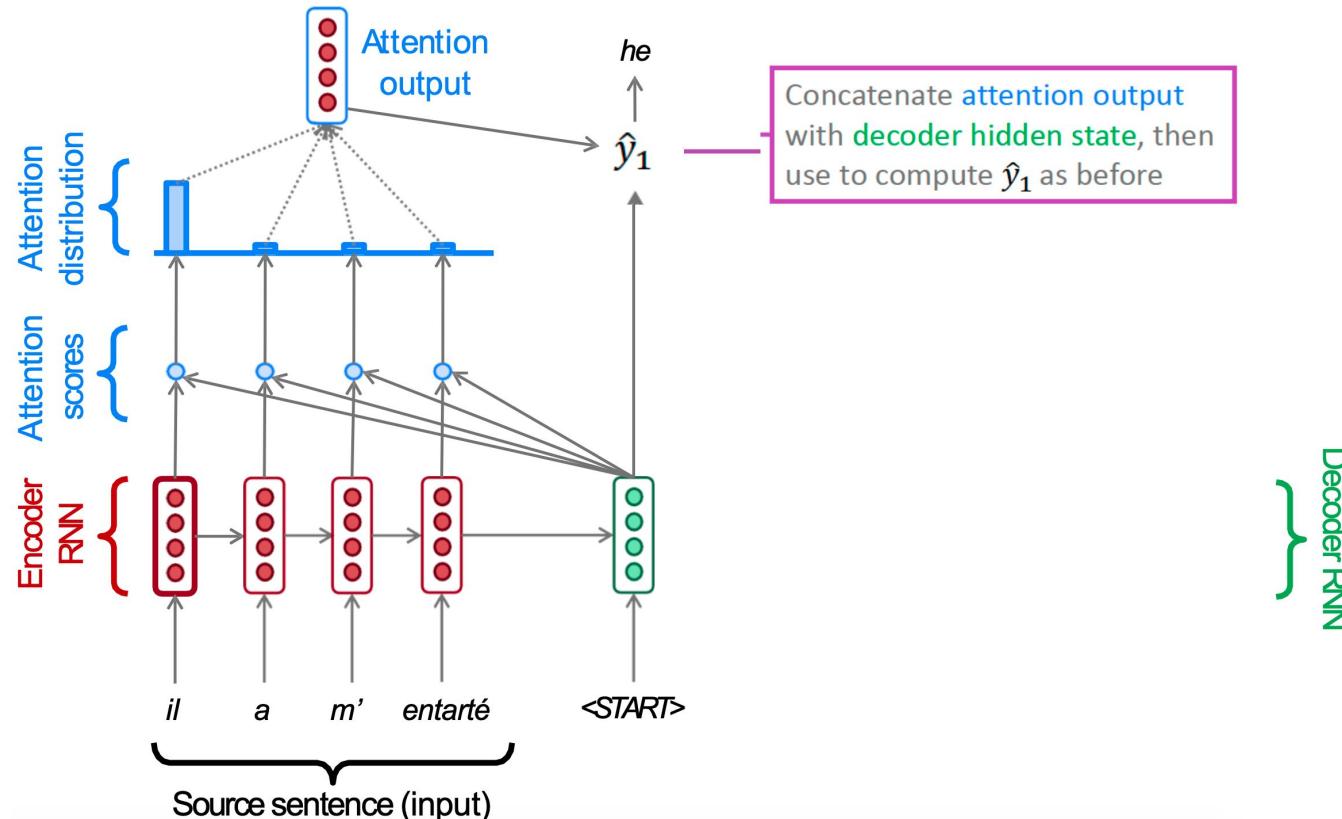
Seq2Seq + Attention: NMT



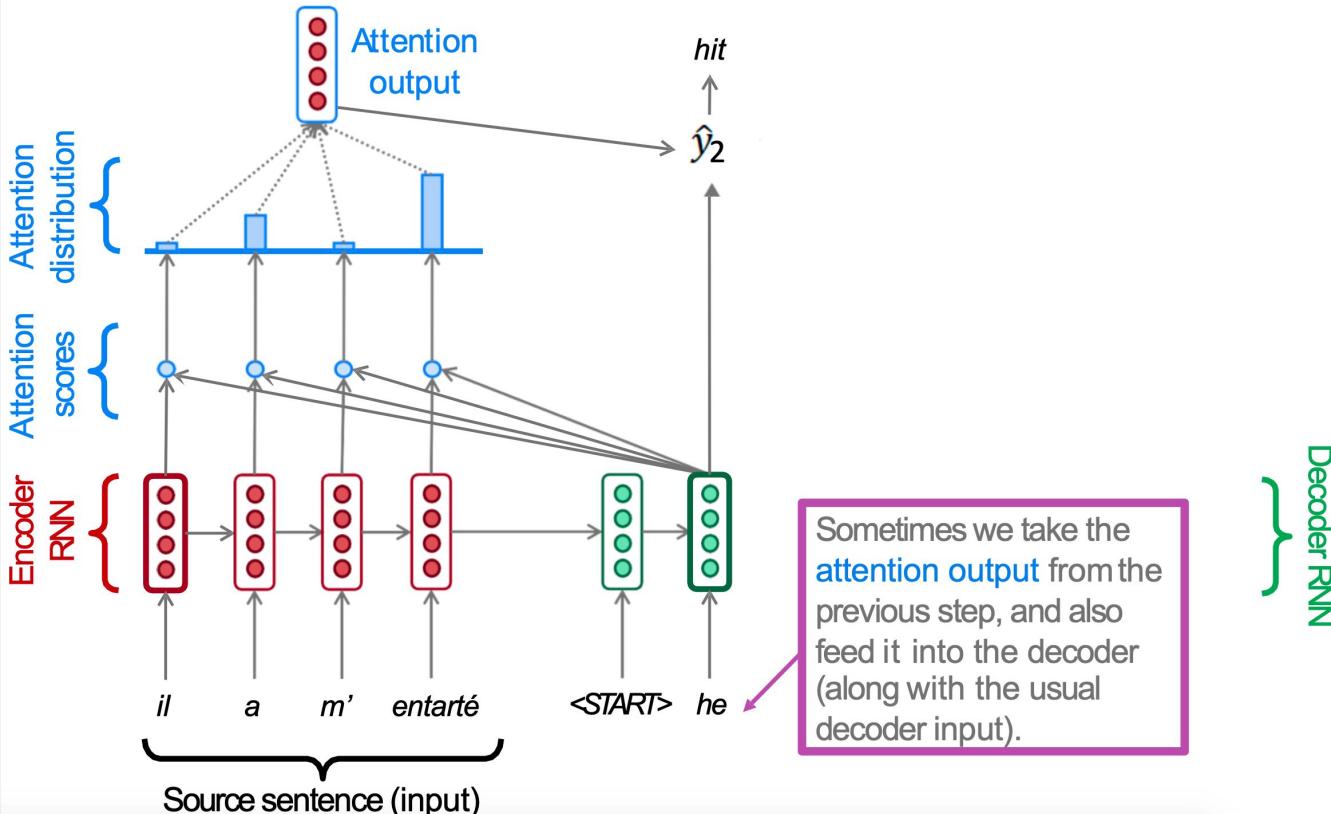
Seq2Seq + Attention: NMT



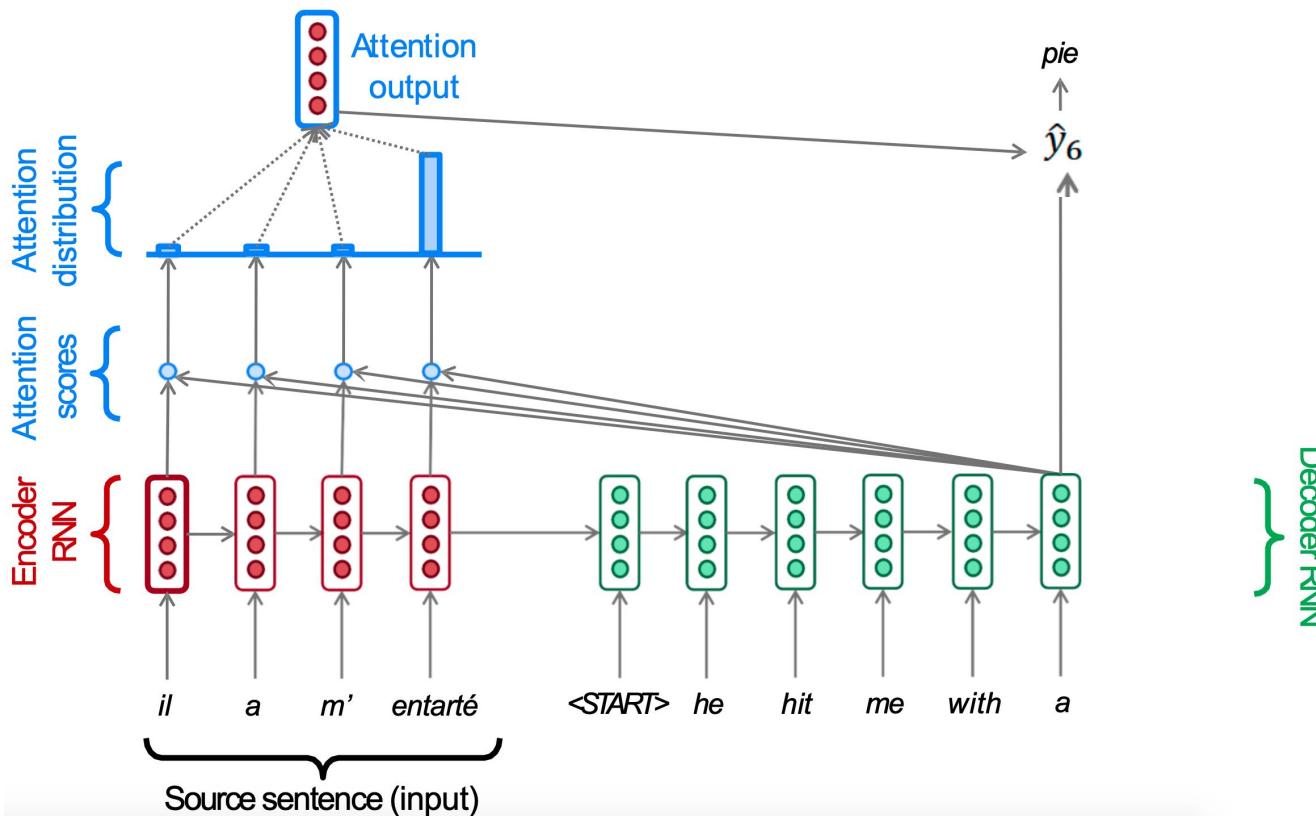
Seq2Seq + Attention: NMT



Seq2Seq + Attention: NMT



Seq2Seq + Attention: NMT



Attention matrix

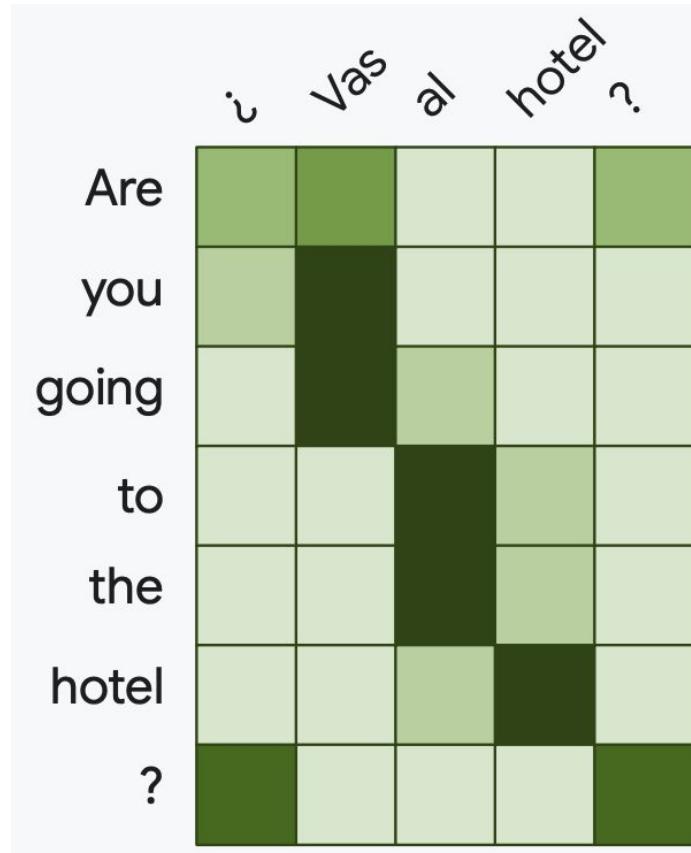
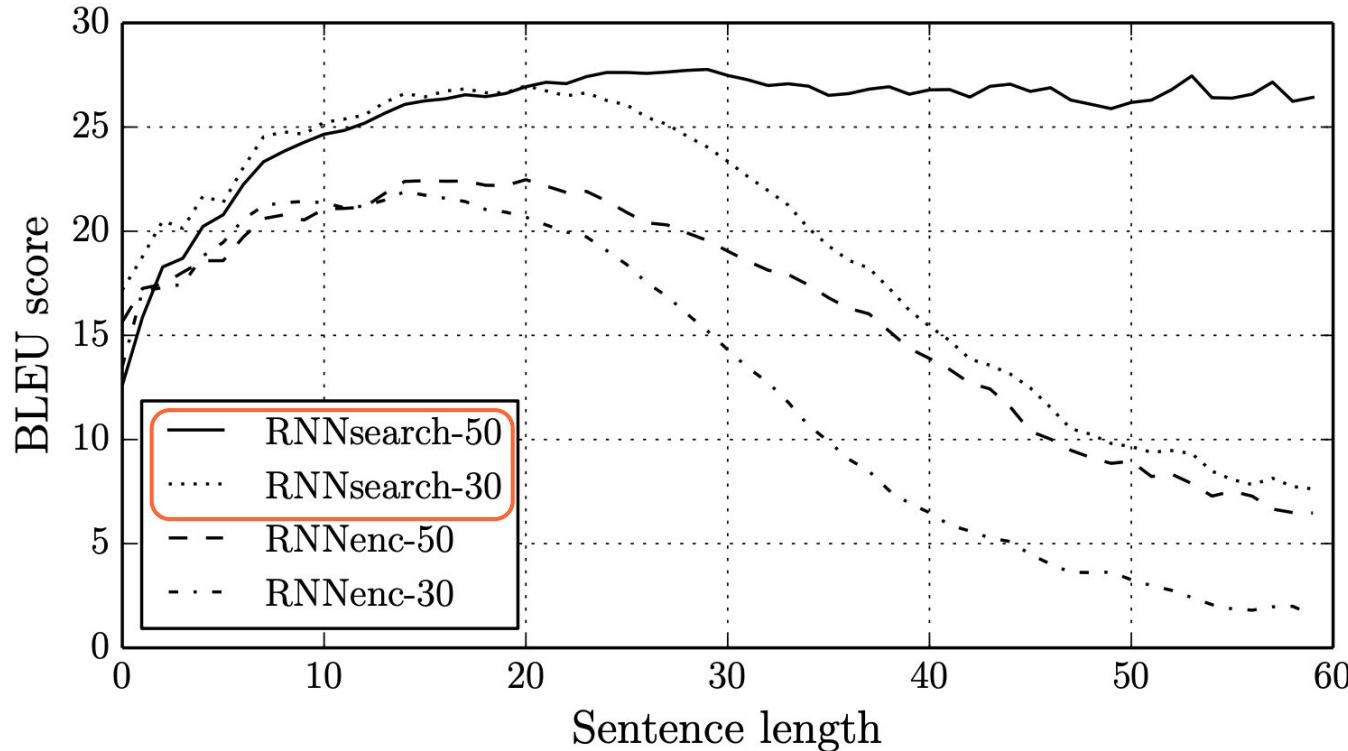


Figure: Ilharco, Ilharco, Turc, Dettmers, Ferreira, Lee, "[High Performance Natural Language Processing](#)". EMNLP 2020.

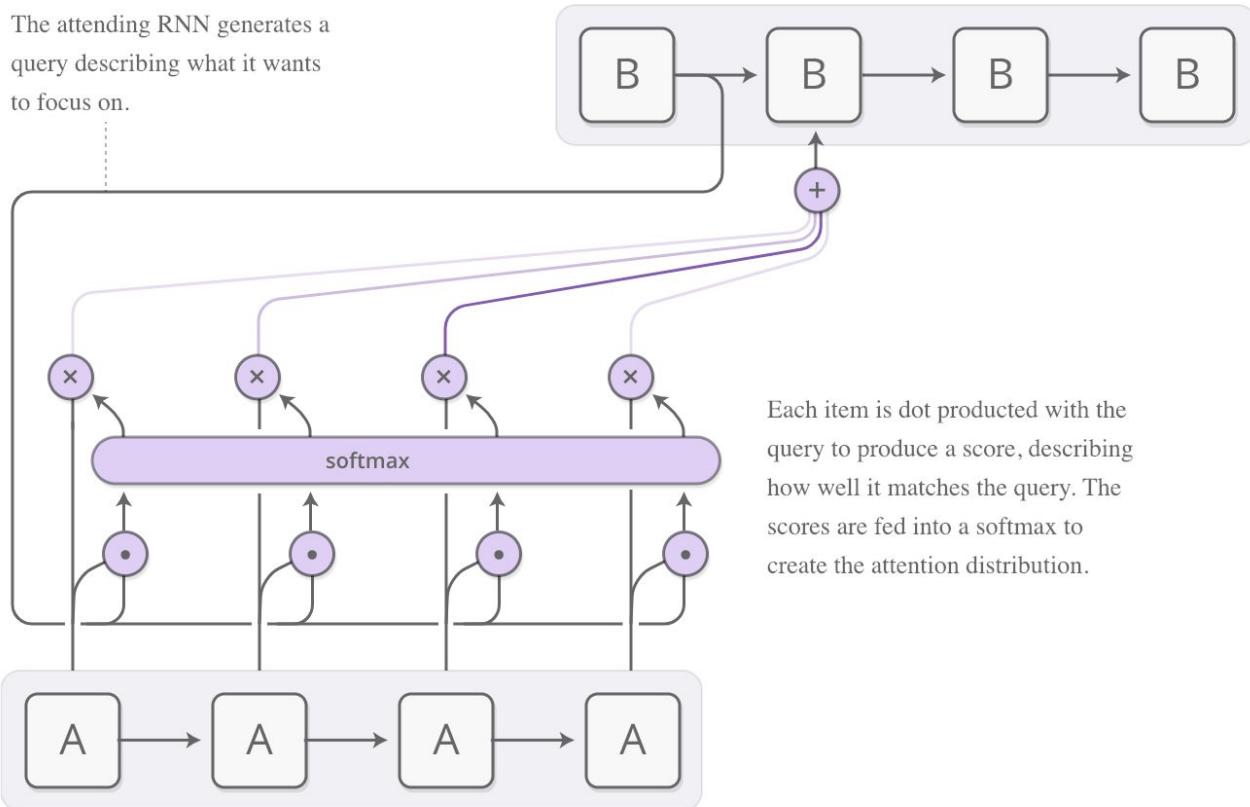
Seq2Seq + Attention: NMT



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "[Neural machine translation by jointly learning to align and translate.](#)" ICLR 2015.

Seq2Seq + Attention

The attending RNN generates a query describing what it wants to focus on.



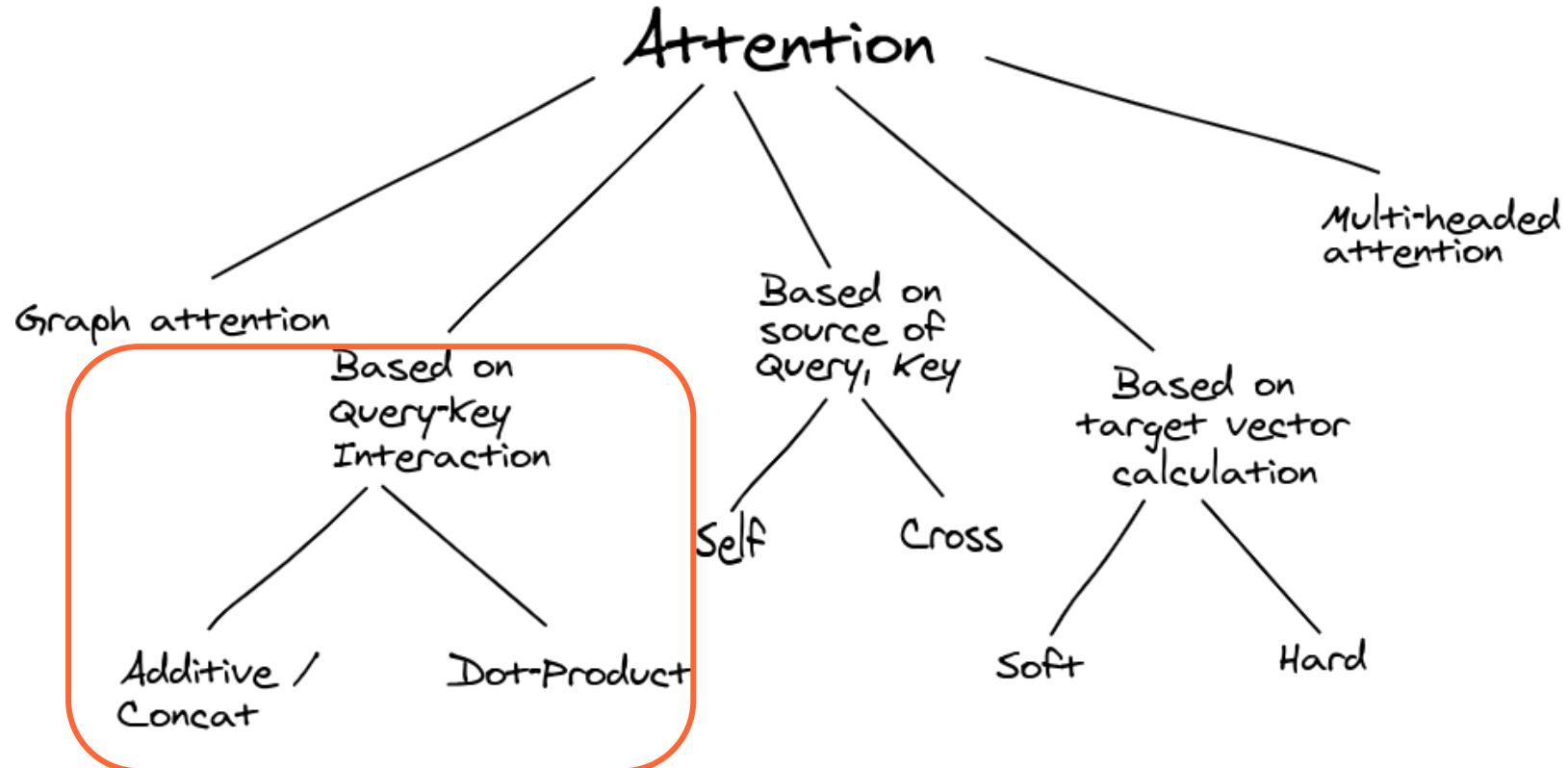
Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. **Attention mechanisms**
6. Study case: Image captioning

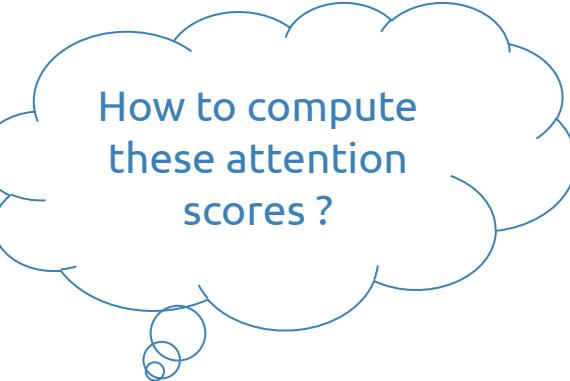
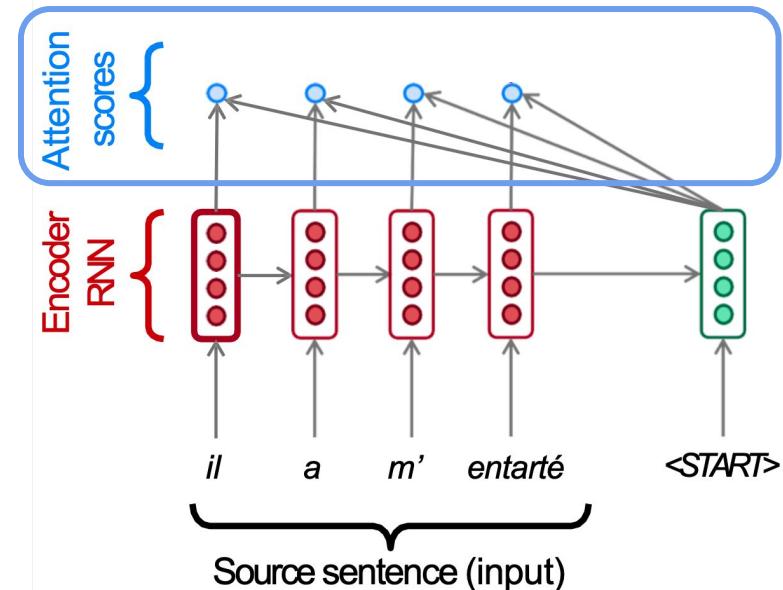


RNNs vs Attention

Attention Mechanisms



Attention Mechanisms



- Additive attention
- Multiplicative
- Scaled dot product

Outline

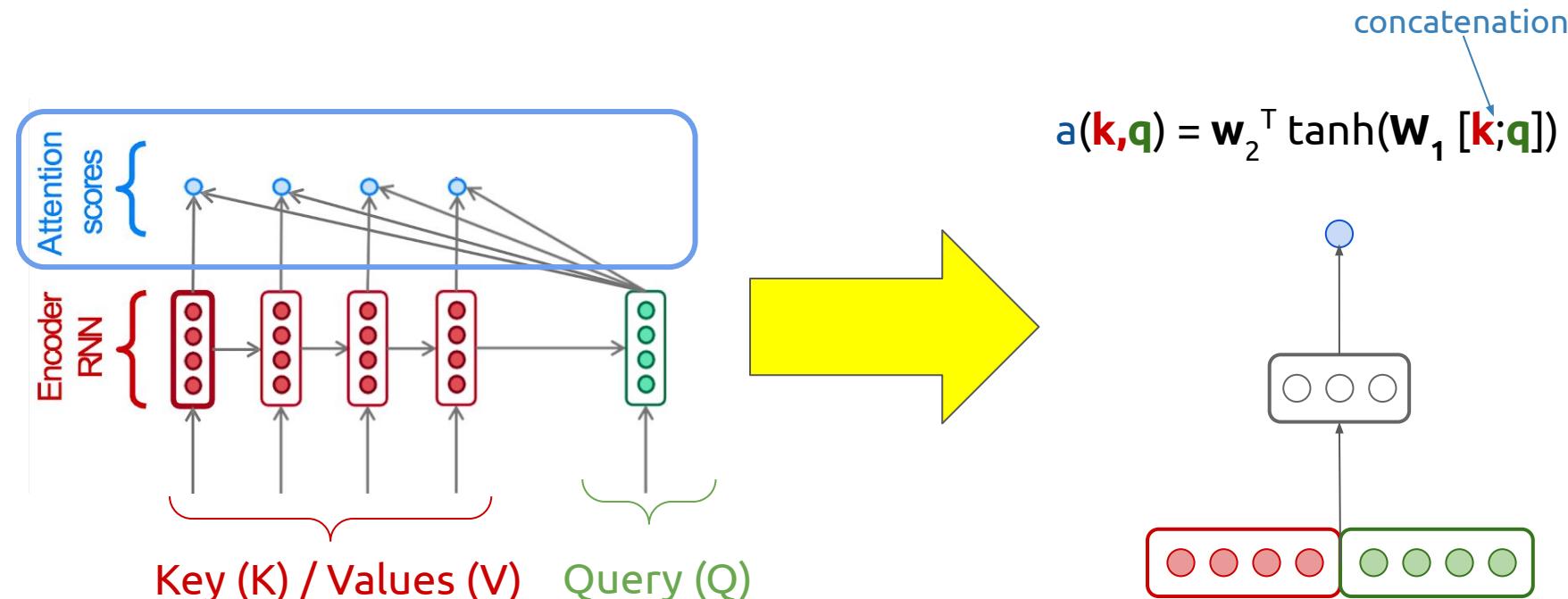
1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. **Attention mechanisms**
 - Additive



RNNs vs Attention

Attention Mechanisms: Additive Attention

The first attention mechanism used a one-hidden layer feed-forward network to calculate the attention alignment:

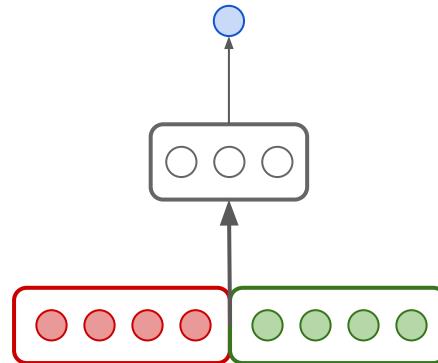


Attention Mechanisms: Additive Attention

The “additive” adjective corresponds to considering the MLP as two different ones, one focus on the query, another on the keys.

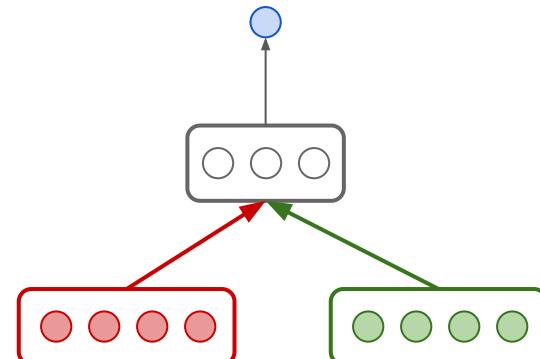
concatenation

$$a(q, k) = w_2^T \tanh(w_1 [q; k])$$



addition

$$a(q, k) = w_2^T \tanh(w_q q + w_k k)$$



Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. **Attention mechanisms**
 - **Multiplicative**

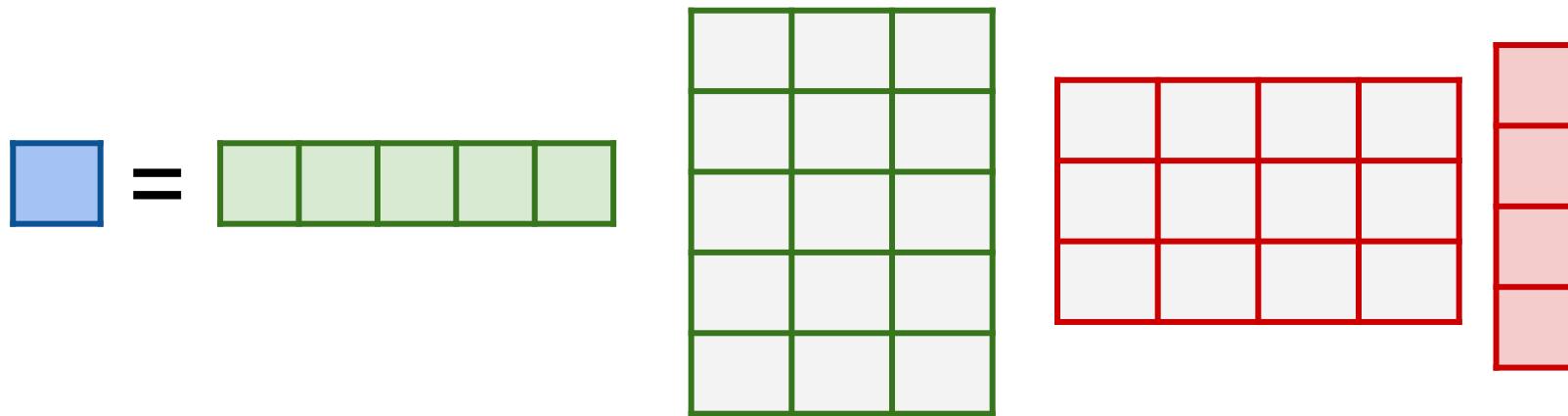


RNNs vs Attention

Attention Mechanisms: Multiplicative Attention

By discarding the tanh layer, keys and query can be directly compared with a multiplication depending on the learnable weights W_q and W_k .

$$a(q, k) = (q W_q)^T W_k k$$



Attention Mechanisms: Dot Product

If considering the projected query and keys, the multiplicative attention can be understood as a simple dot product.

$$a(\mathbf{q}, \mathbf{k}) = \mathbf{q}^T \mathbf{k}$$



Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. **Attention mechanisms**
 - Scaled dot product



RNNs vs Attention

Attention Mechanisms: Scaled dot product

Multiplicative attention performs worse than additive attention for large **dimensionality of the decoder state (d_h)**, unless a scaling factor is added:

$$a(\mathbf{q}, \mathbf{k}) = \mathbf{q}^T \mathbf{k} / \sqrt{d_h}$$

$$\boxed{\text{blue square}} = \frac{1}{\sqrt{d_h}} \begin{array}{c} \text{green row vector} \\ \text{red column vector} \end{array}$$

Attention Mechanisms: Scaled dot product

Considering a dot-product attention, how are the attention scores (a_{ij}) between each pair of input tokens x_i and x_j computed ?

- $a_{ij} = f(x_i)^T \cdot g(x_j)$
- $a_{ij} = f(x_i)^T \cdot g(y_j)$
- $a_{ij} = \text{softmax}[f(x_i)^T \cdot g(x_j)]$
- $a_{ij} = \text{softmax}[f(x_i)^T \cdot g(y_j)]$

Query (Q)
 $g(x) = W^Q x$

Key (K)
 $f(x) = W^K x$

Value (V)
 $h(x) = W^V x$

W^Q , W^K and W^V are **projection layers** shared across all words.

Attention Mechanisms: Scaled dot product

Considering a dot-product attention, how are the attention scores (a_{ij}) between each pair of input tokens x_i and x_j computed ?

- $a_{ij} = f(x_i)^\top g(x_j)$

- $a_{ij} = f(x_i)^\top g(y_j)$

- $a_{ij} = \text{softmax}[f(x_i)^\top \cdot g(x_j)]$

- $a_{ij} = \text{softmax}[f(x_i)^\top \cdot g(y_j)]$

Query (Q)
 $g(x) = W^Q x$

Key (K)
 $f(x) = W^K x$

Value (V)
 $h(x) = W^V x$

W^Q , W^K and W^V are projection layers shared across all words.

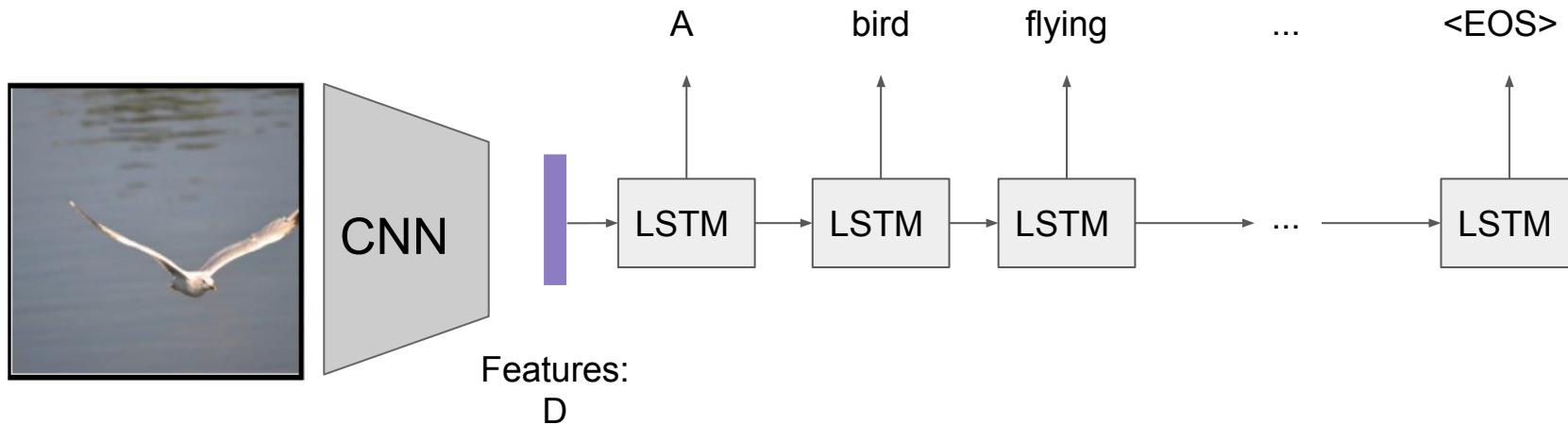
Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. Attention mechanisms
6. **Study case: Image captioning**



RNNs vs Attention

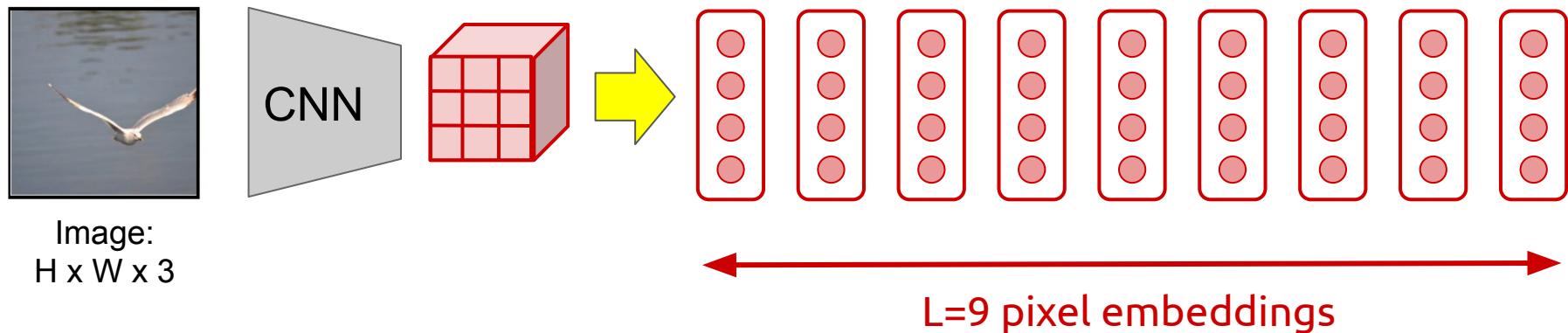
Image Captioning without Attention



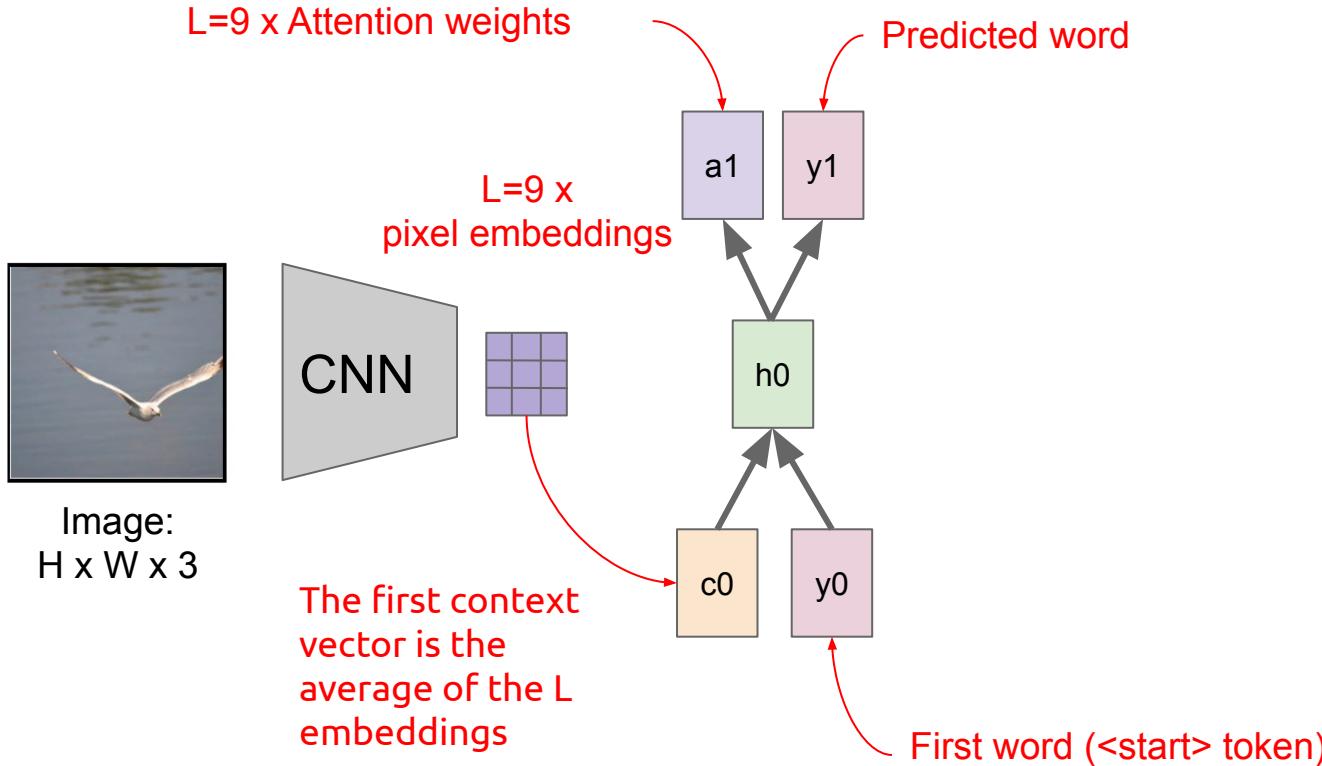
Limitation: All output predictions are based on the **final and static** output of the encoder

Image Captioning with Attention

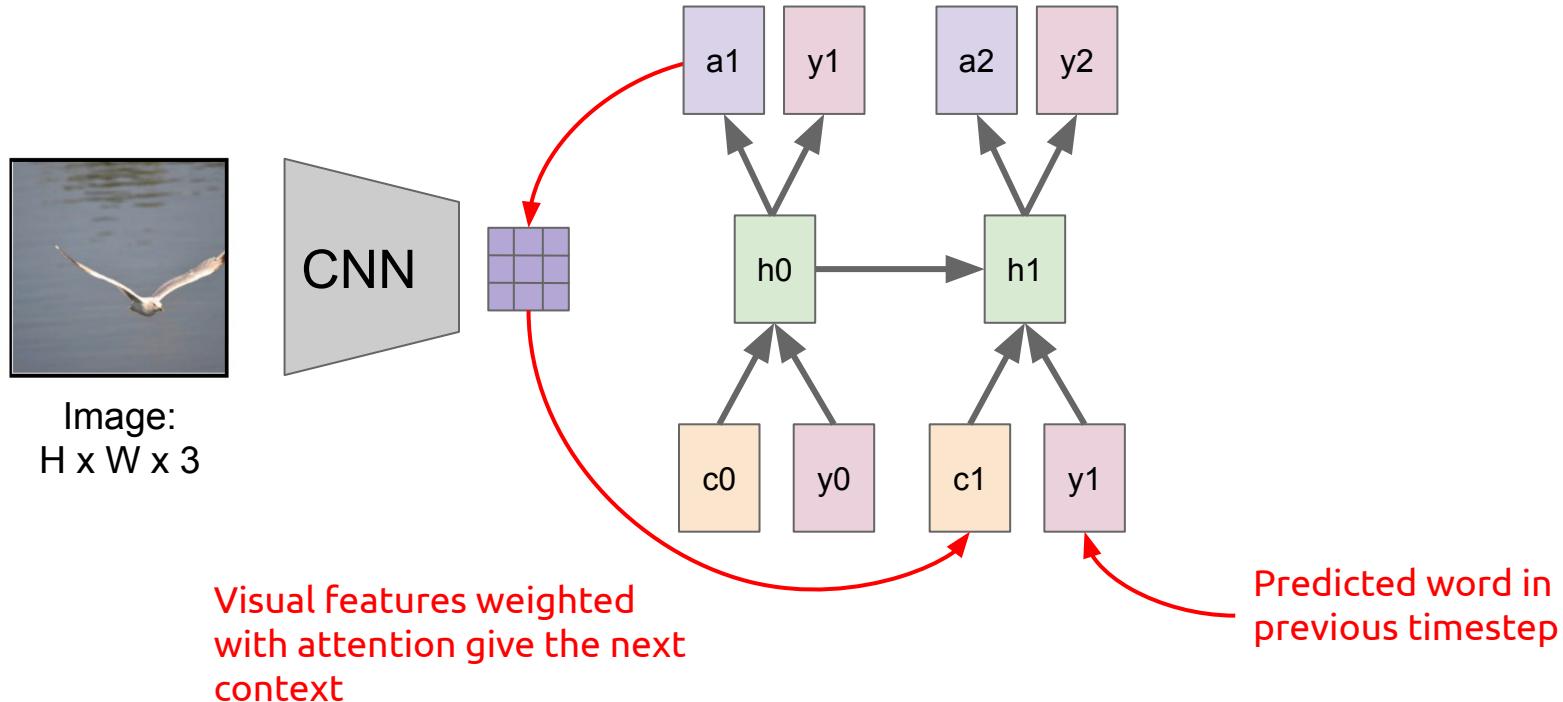
How can we encode an image as a set of representation to attend to ?



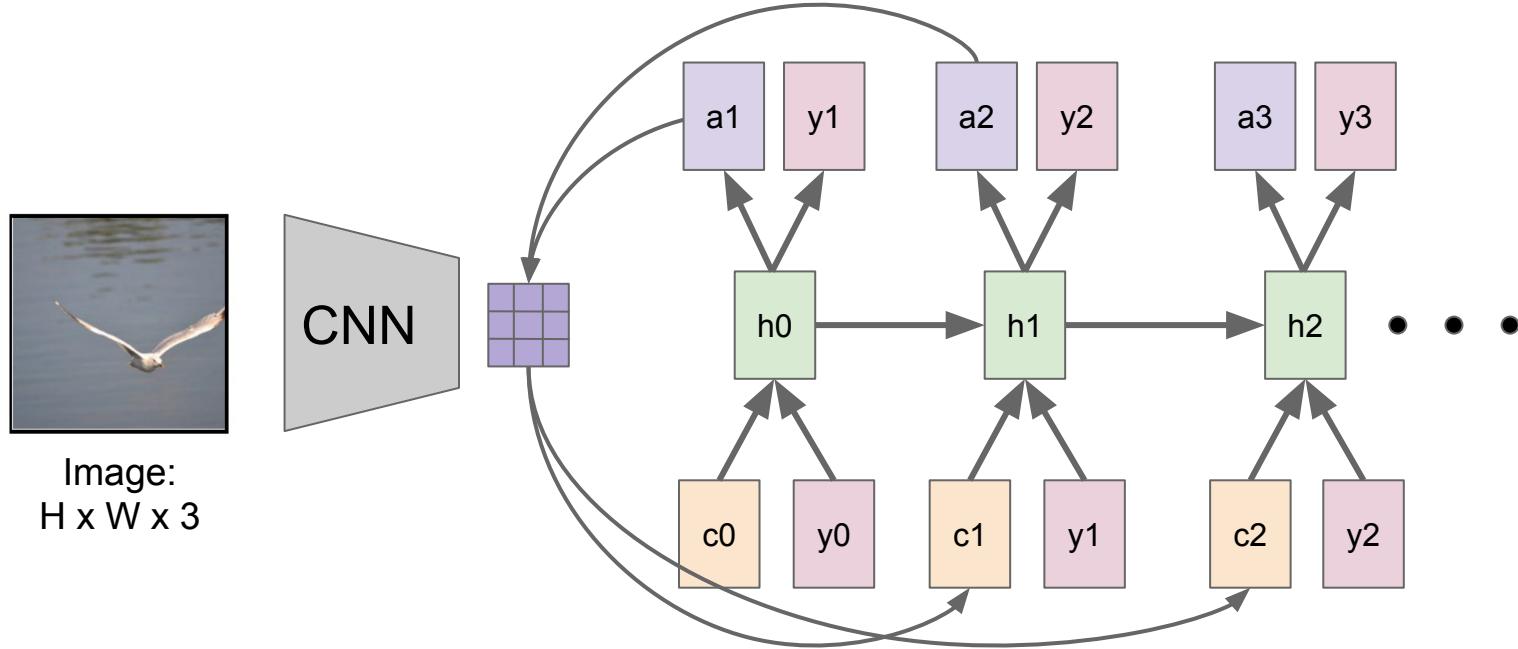
Attention for Image Captioning



Attention for Image Captioning



Attention for Image Captioning



Attention for Image Captioning

A(0.99)



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.



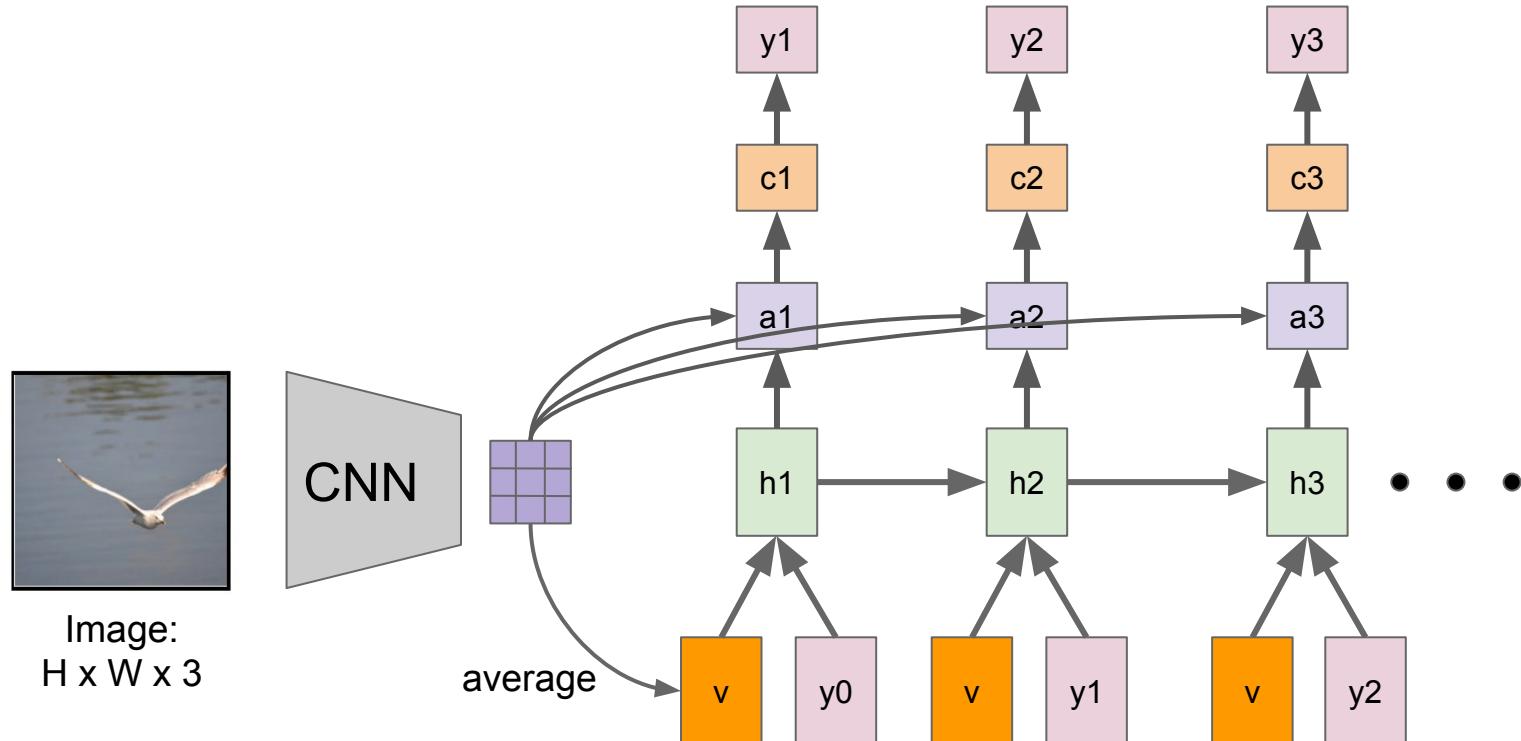
A group of people sitting on a boat in the water.



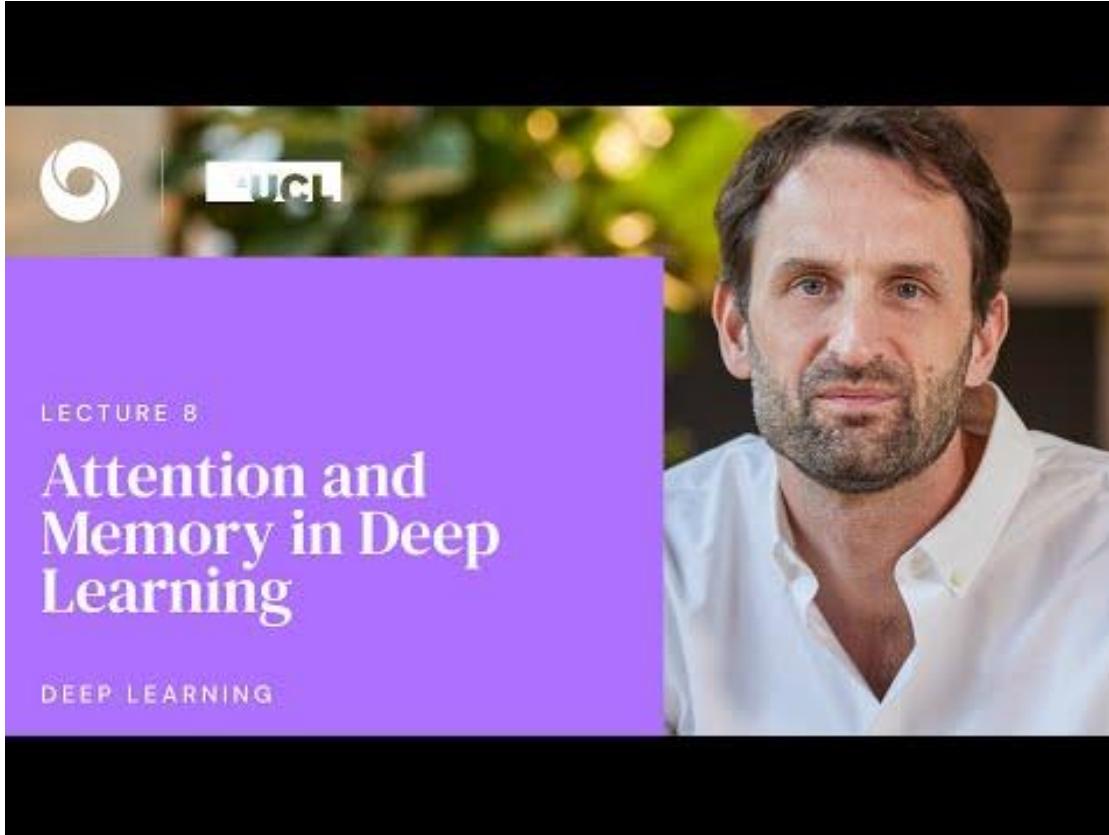
A giraffe standing in a forest with trees in the background.

Attention for Image Captioning

Side-note: attention can be computed with previous or **current** hidden state



Learn more



Alex Graves, [UCL x Deepmind 2020 \[slides\]](#)

Learn more

- Tutorials
 - Sebastian Ruder, [Deep Learning for NLP Best Practices # Attention](#) (2017).
 - Chris Olah, Shan Carter, ["Attention and Augmented Recurrent Neural Networks"](#). Distill 2016.
 - Lilian Weng, ["Attention? Attention!"](#). Lil'Log 2017.
- Demos
 - [François Fleuret](#) (EPFL)

Questions ?

Undergradese

What undergrads ask vs. what they're REALLY asking

"Is it going to be an open book exam?"

Translation: "I don't have to actually memorize anything, do I?"

"Hmm, what do you mean by that?"

Translation: "What's the answer so we can all go home."

"Are you going to have office hours today?"

Translation: "Can I do my homework in your office?"

"Can i get an extension?"

Translation: "Can you re-arrange your life around mine?"

"Is this going to be on the test?"

Translation: "Tell us what's going to be on the test."

"Is grading going to be curved?"

Translation: "Can I do a mediocre job and still get an A?"

