Master in Computer Vision Barcelona

[http://pagines.uab.cat/mcv/]

Xavier Giro-i-Nieto

@DocXavi

xavier.giro@upc.edu

Associate Professor

Universitat Politècnica de Catalunya

Institut de Robòtica i Informàtica Industrial

Module 6 - Day 6 - Lecture 2

Attention Mechanisms

24th March 2022

Video-lectures

Xavier Giró
[UPC AA2 2020]

Xavier Giró
UPC DLAI 2021

Acknowledgments

Amaia Salvador

PhD 2019
Universitat Politècnica de Catalunya

Marta R. Costa-jussà

Associate Professor
Universitat Politècnica de Catalunya

[DLCV 2016]

[IDL 2018]

Outline

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

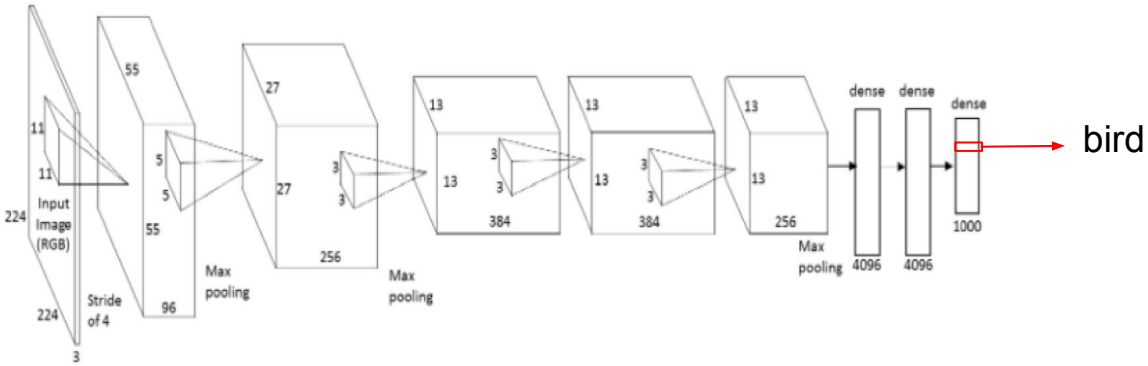
Outline

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

Motivation: Vision

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

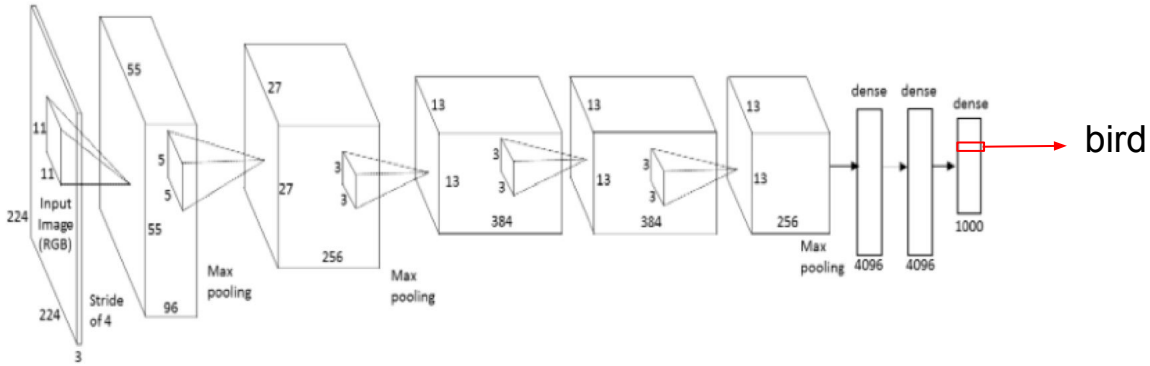
Image:
H x W x 3



Motivation: Vision

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

Image:
H x W x 3



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

Motivation: Automatic Speech Recognition (ASR)

Transcribed letters correspond with just a few phonemes.

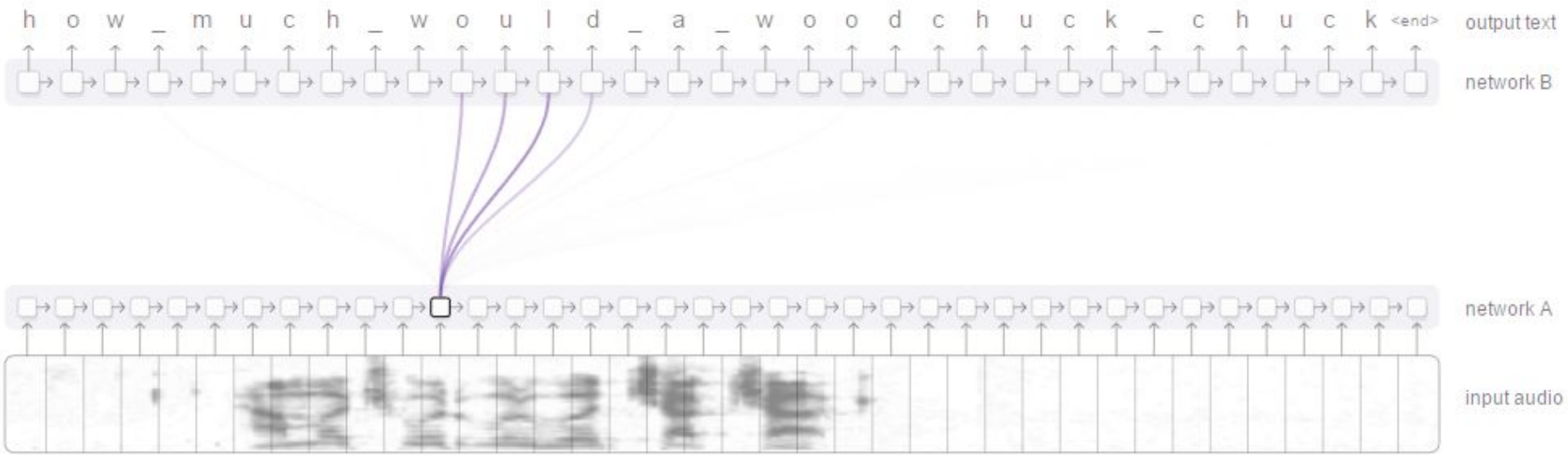
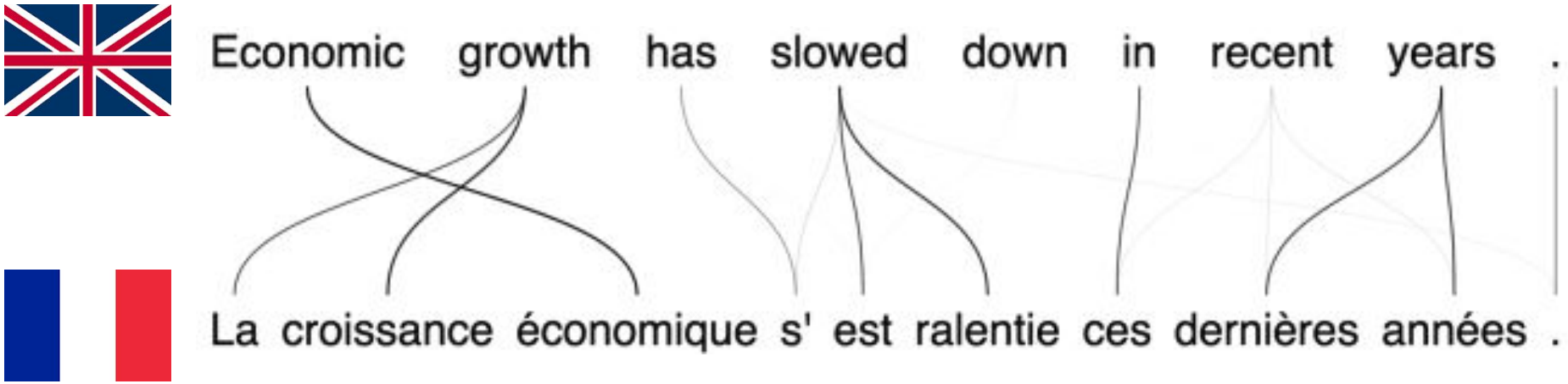


Figure: distill.pub

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)

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

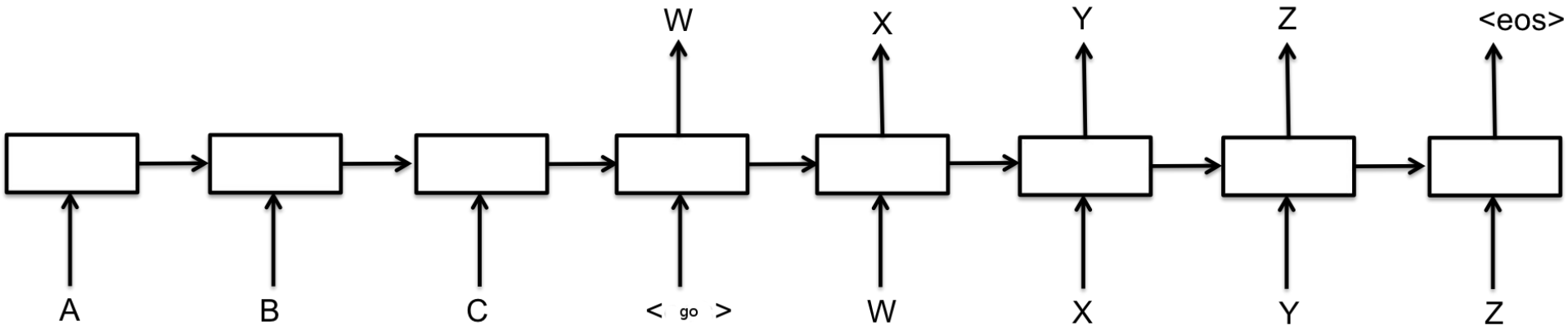
Outline

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

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*.

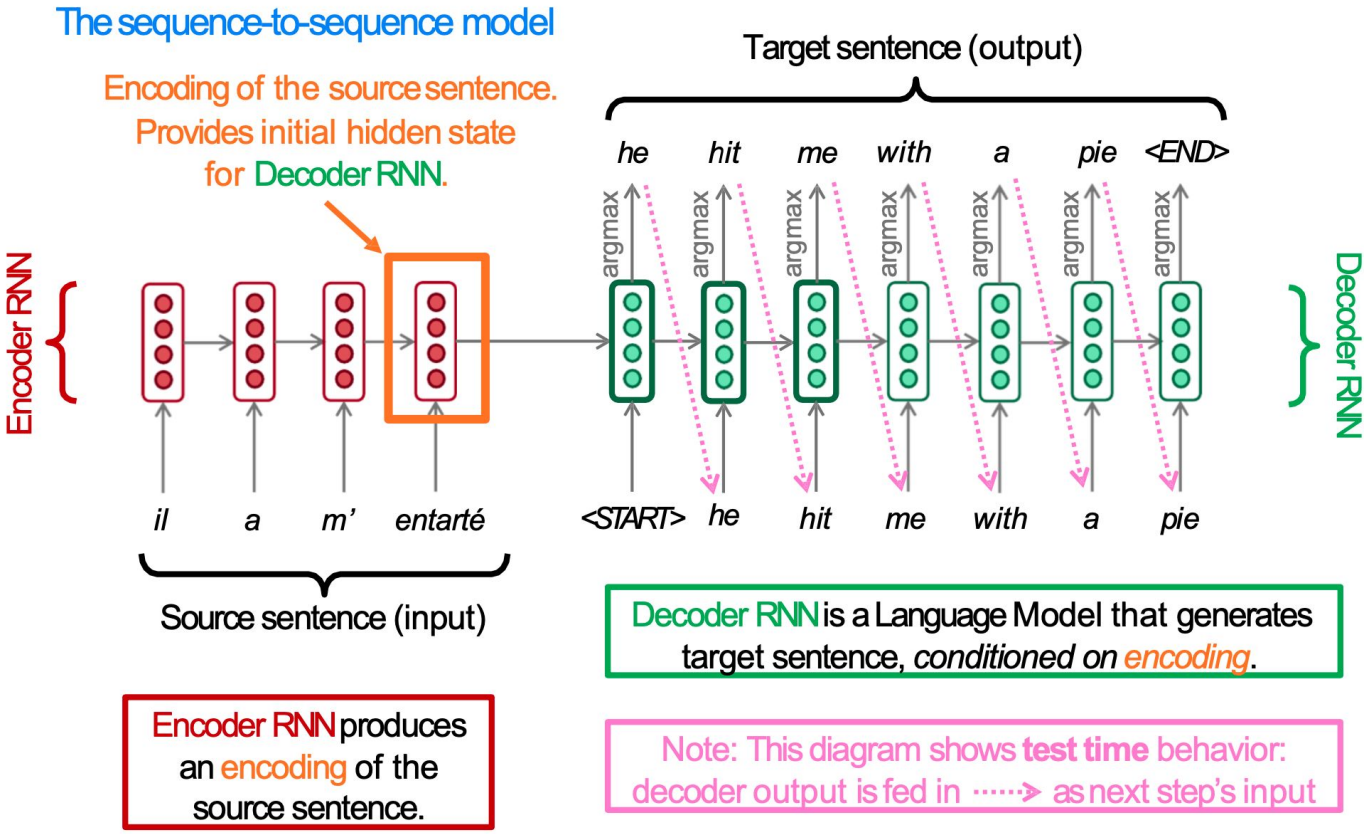


Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." NIPS 2014.

Seq2Seq with RNN

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." NIPS 2014.

Seq2Seq for NMT



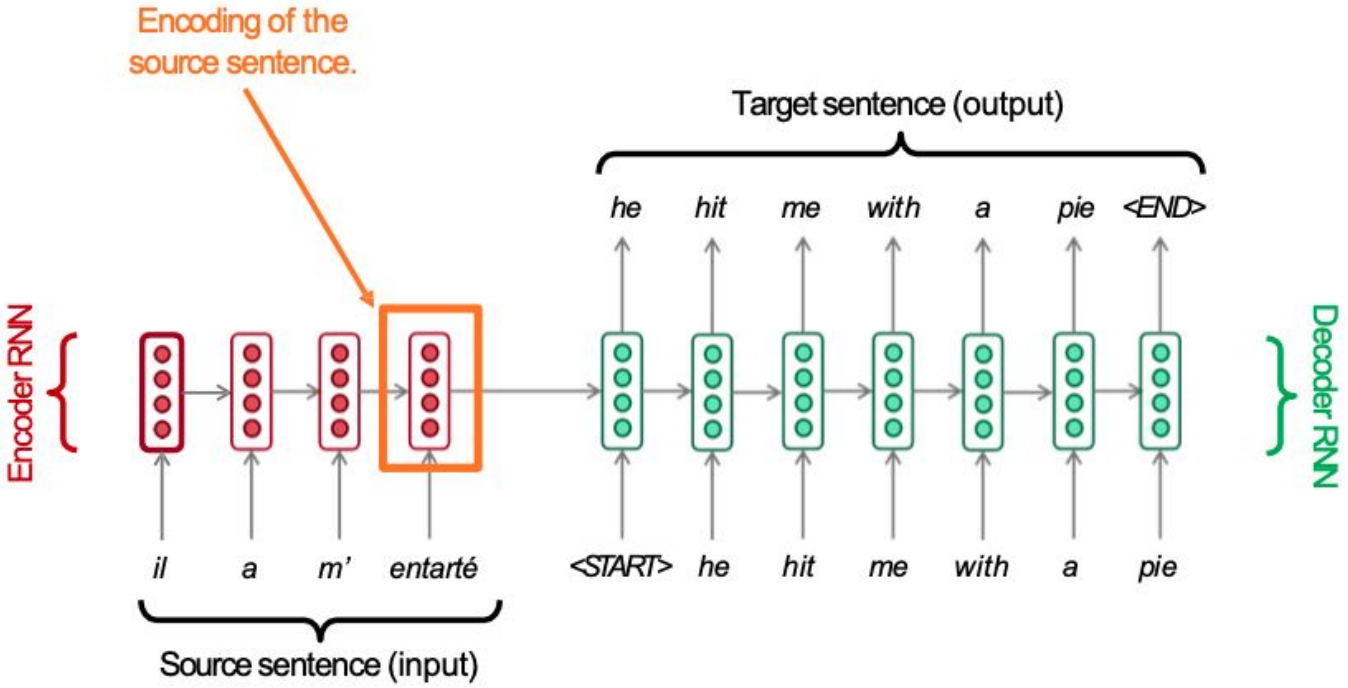
Slide: Abigail See, Matthew Lamm (Stanford CS224N), 2020

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 ?



Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

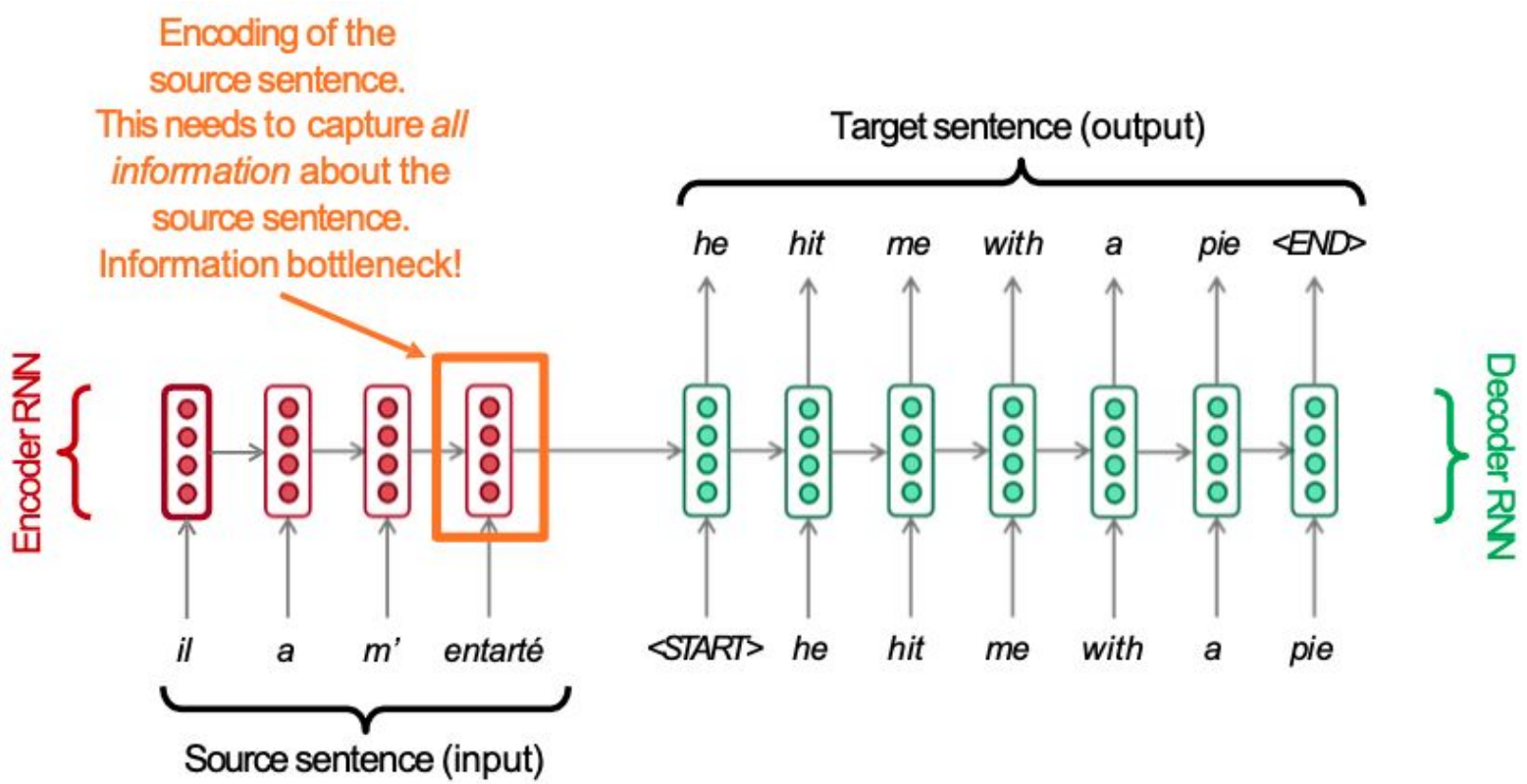
Seq2Seq for NMT

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!"

Ray Mooney, ACL 2014 Workshop on Semantic Parsing.

Seq2Seq for NMT: Information bottleneck



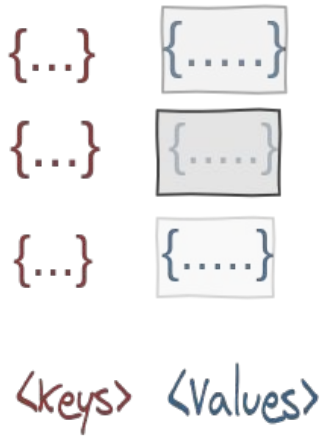
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

Outline

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

Query, Keys & Values

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



Example:

`myfile.pdf`

<Key> <Value>

Figure: Nikhil Shah, "Attention? An Other Perspective! [Part 1]" (2020)

Query, Keys & Values

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

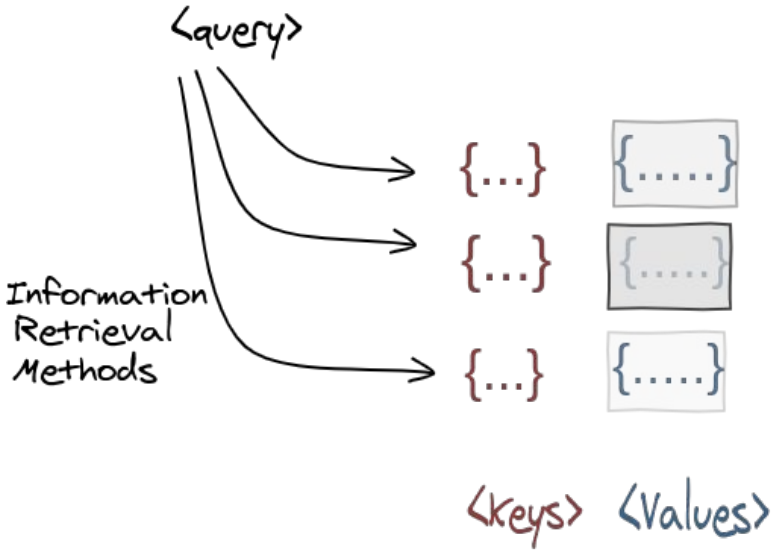


Figure: Nikhil Shah, "Attention? An Other Perspective! [Part 1]" (2020)

Query, Keys & Values

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

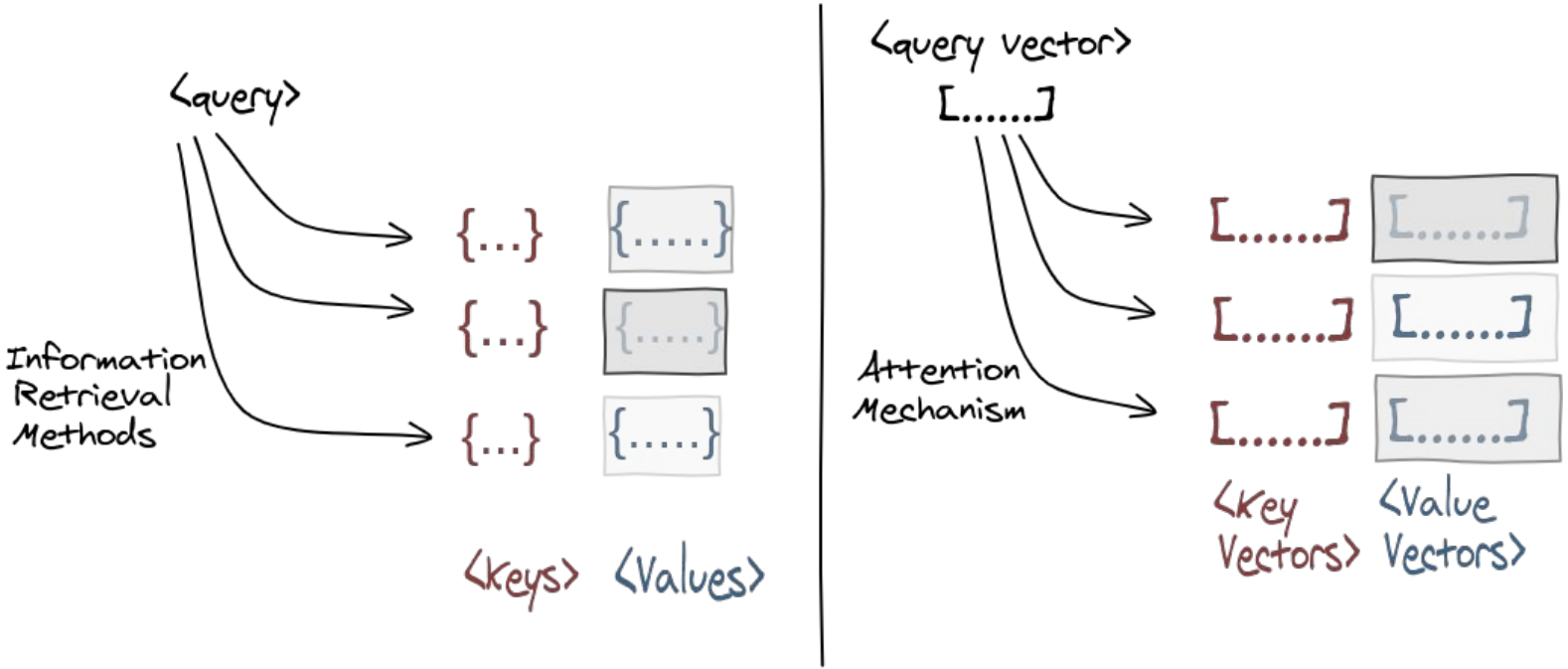
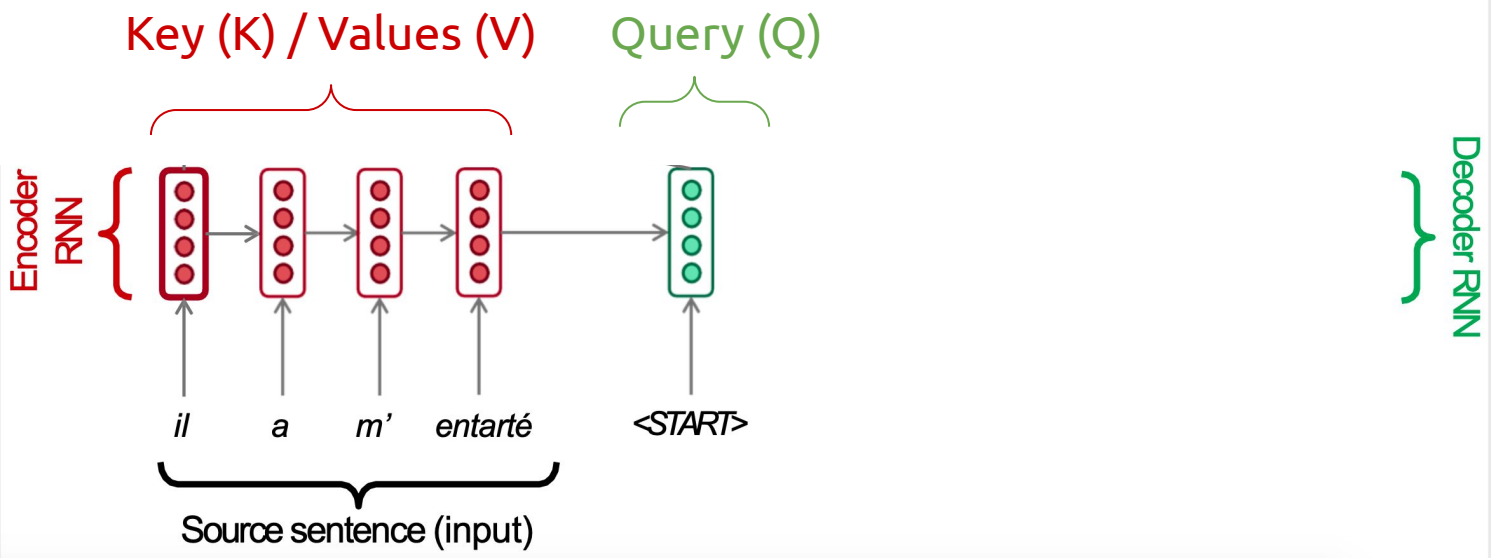


Figure: Nikhil Shah, "Attention? An Other Perspective! [Part 1]" (2020)

Query, Keys & Values

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



Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

Outline

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

Seq2Seq + Attention

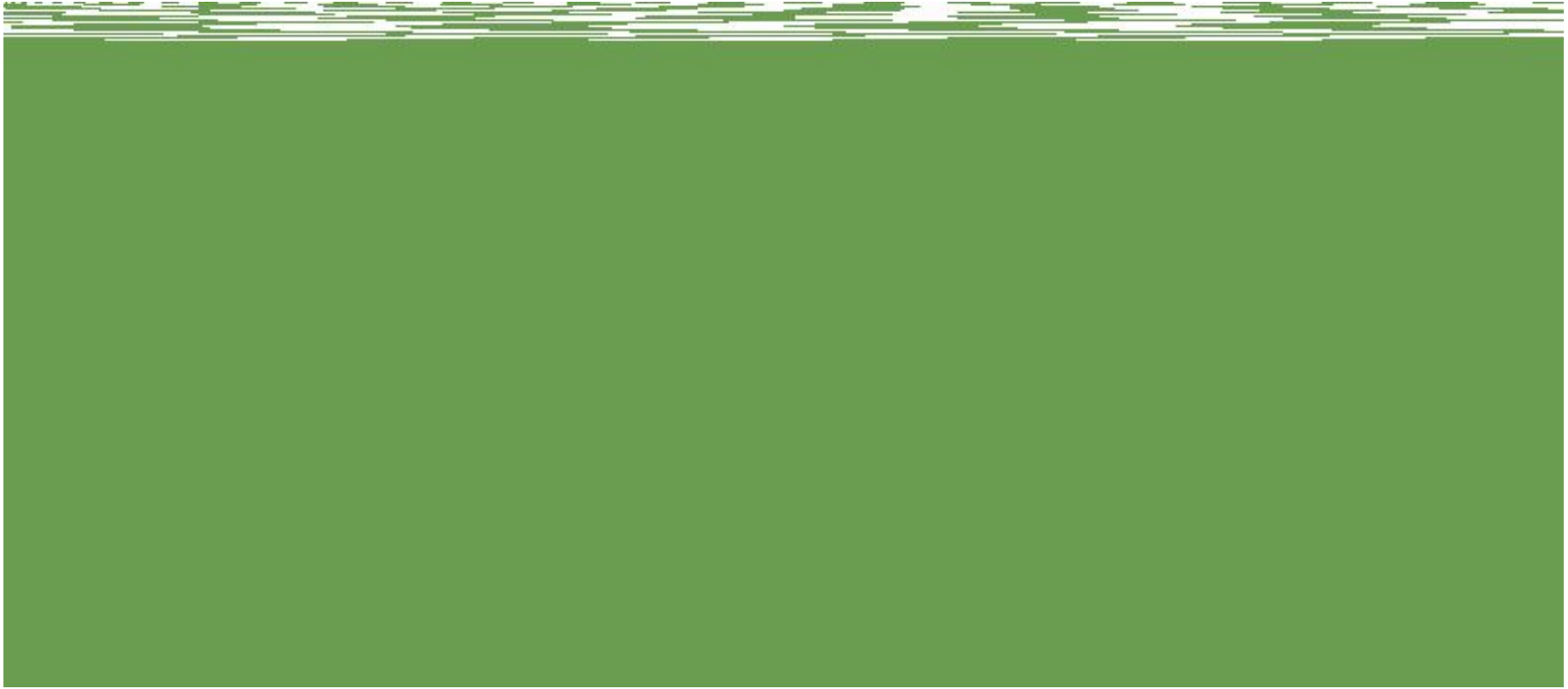
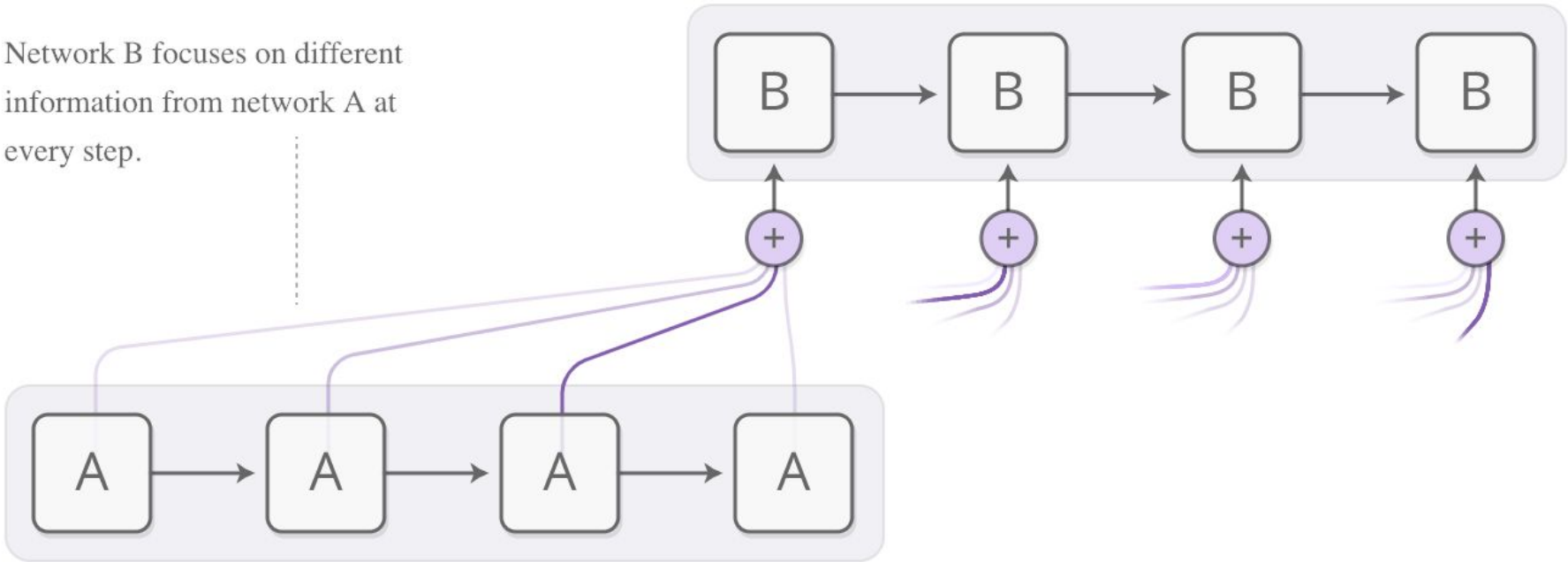


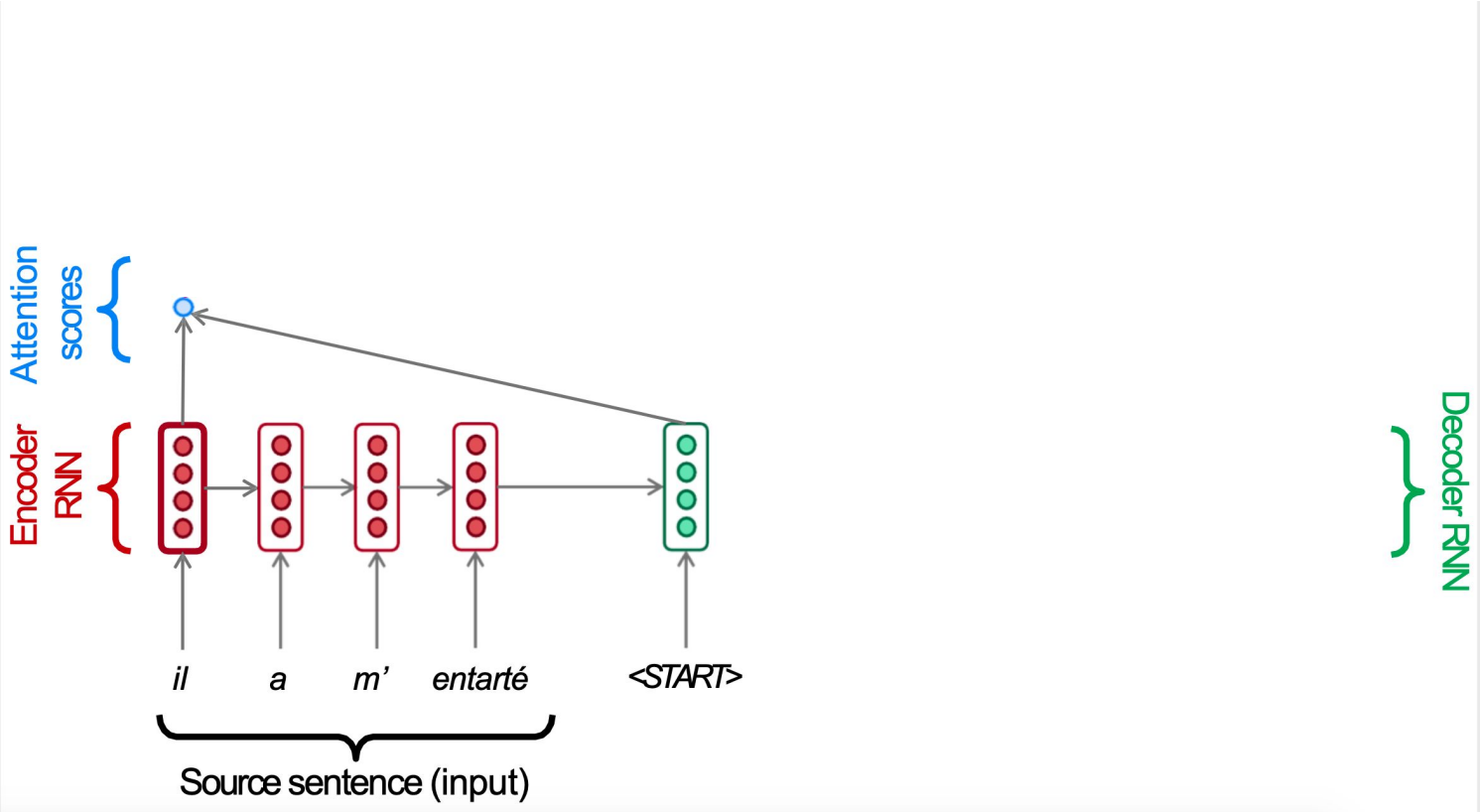
Figure: Jay Alammar, "Visualizing A Neural Machine Translation Model" (2018)

Seq2Seq + Attention



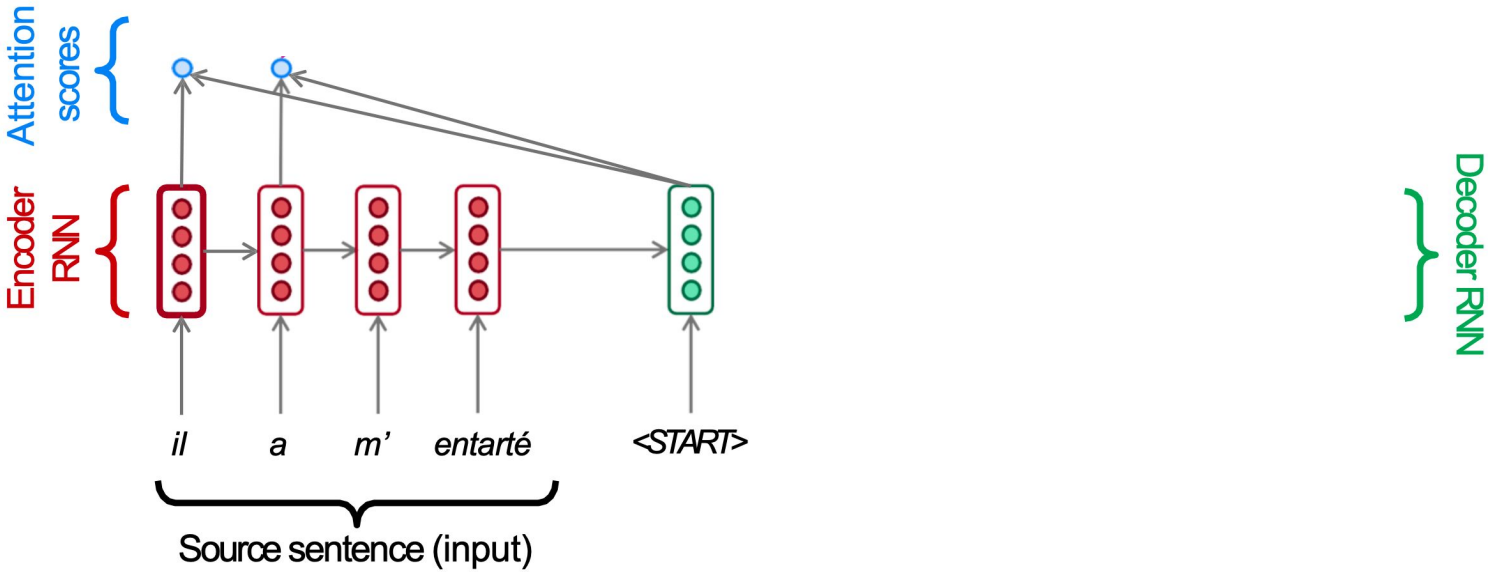
Chis Olah & Shan Cate, [“Attention and Augmented Recurrent Neural Networks”](#) distill.pub 2016.

RNN + Attention: NMT



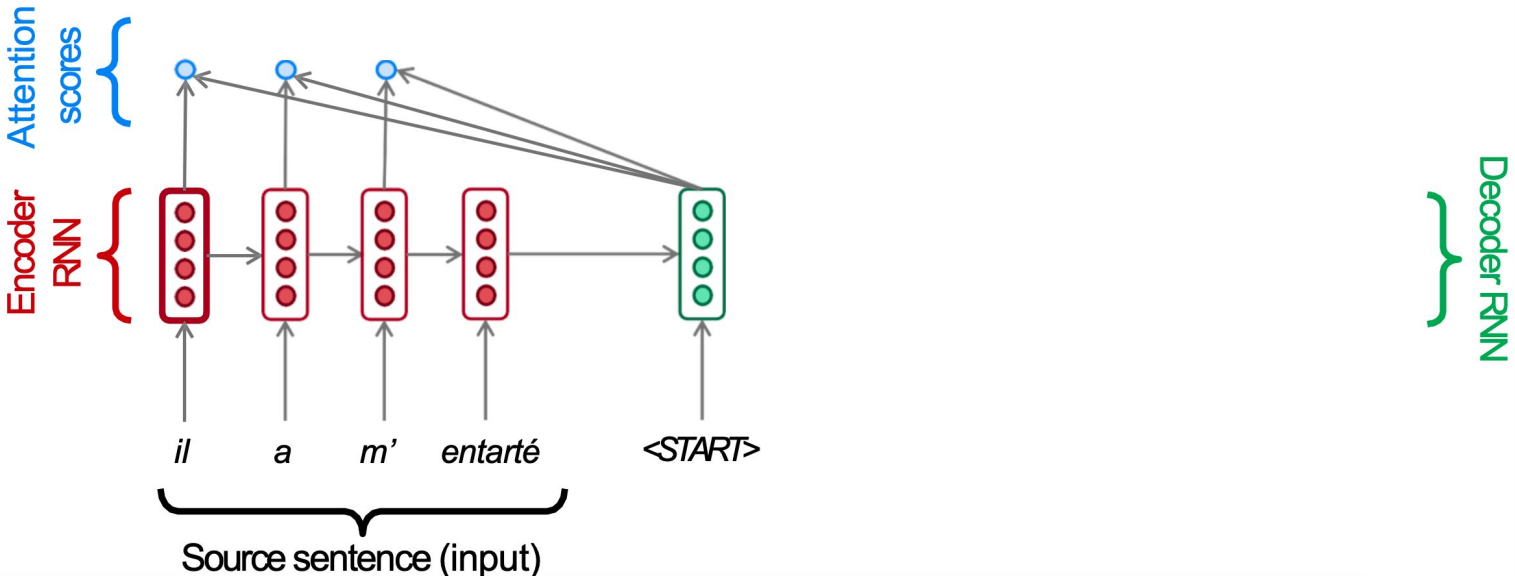
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



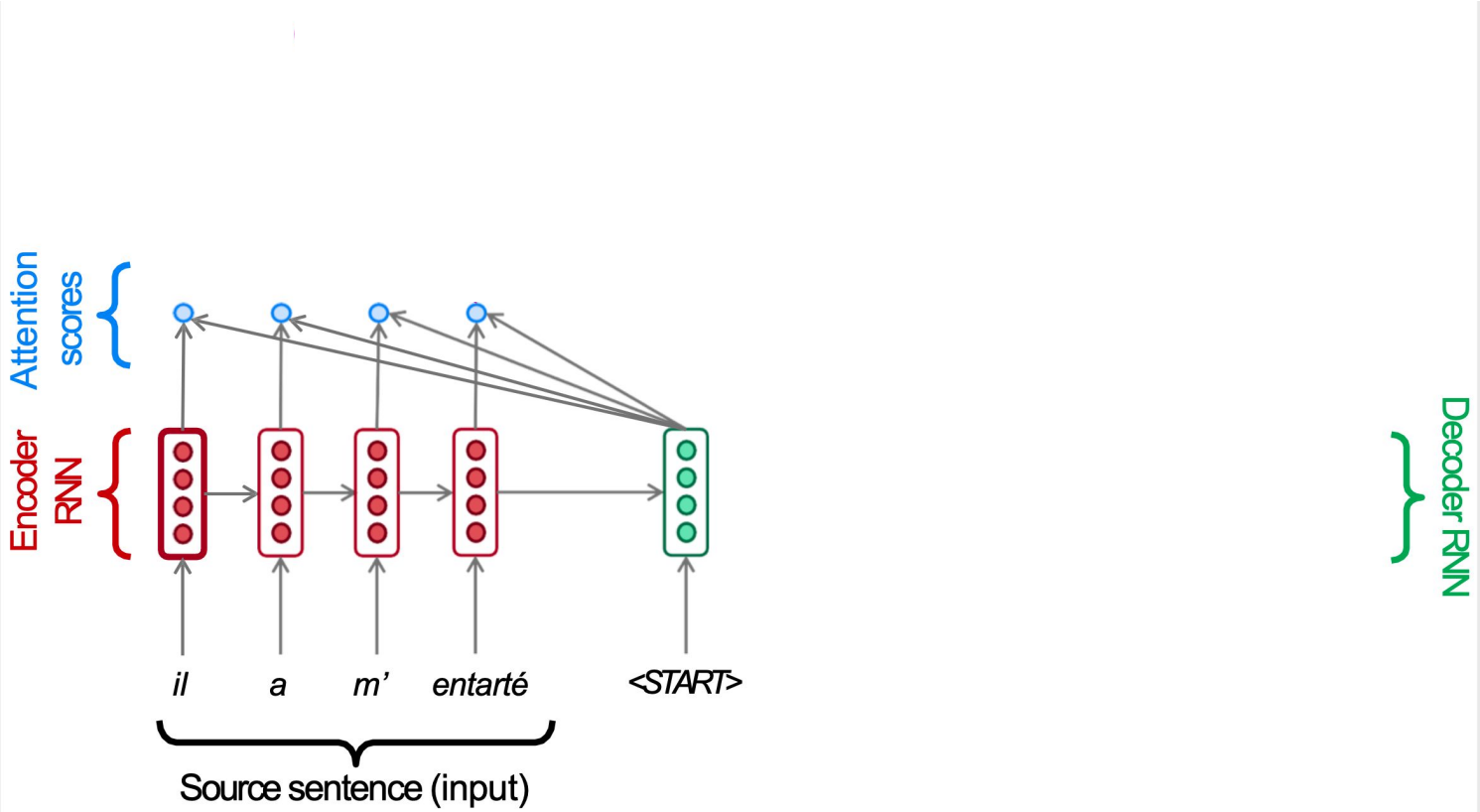
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



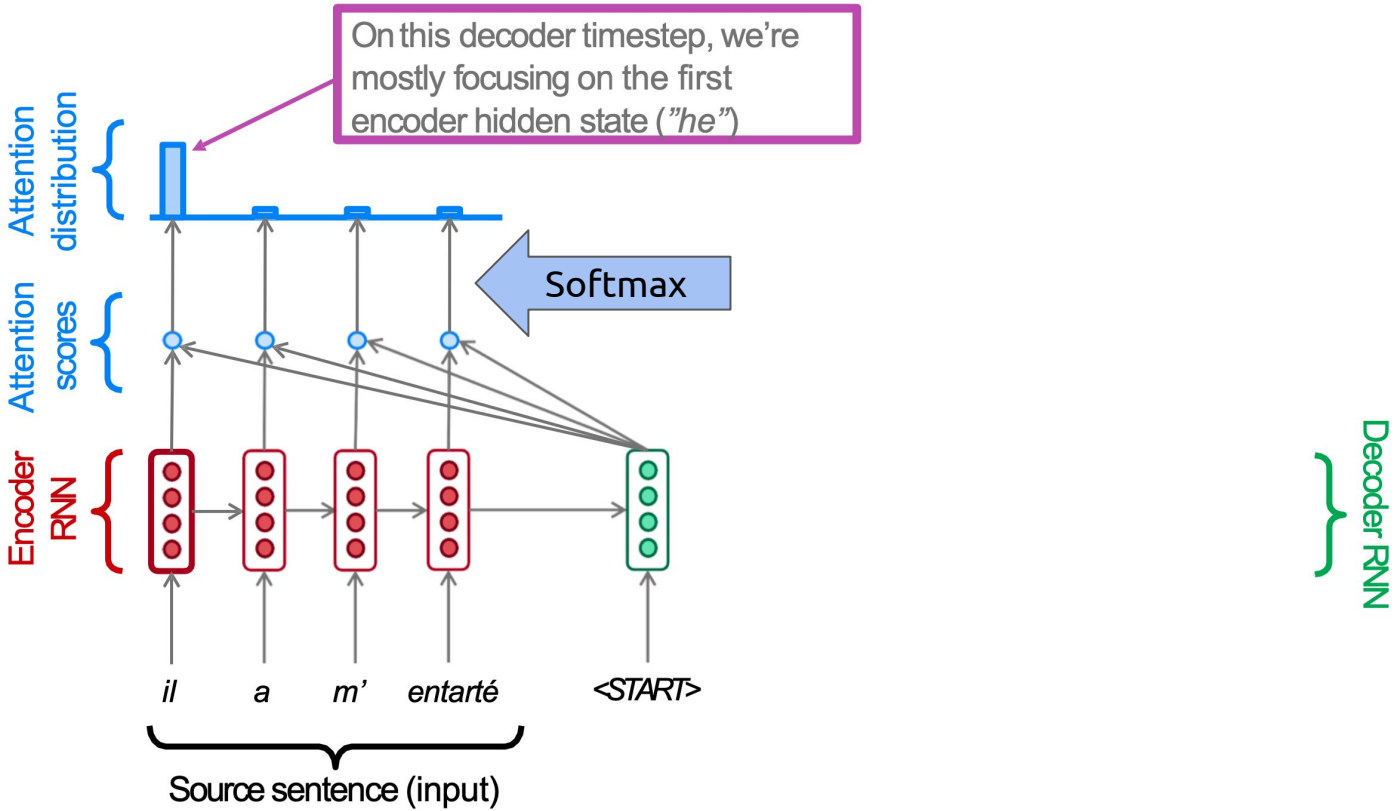
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



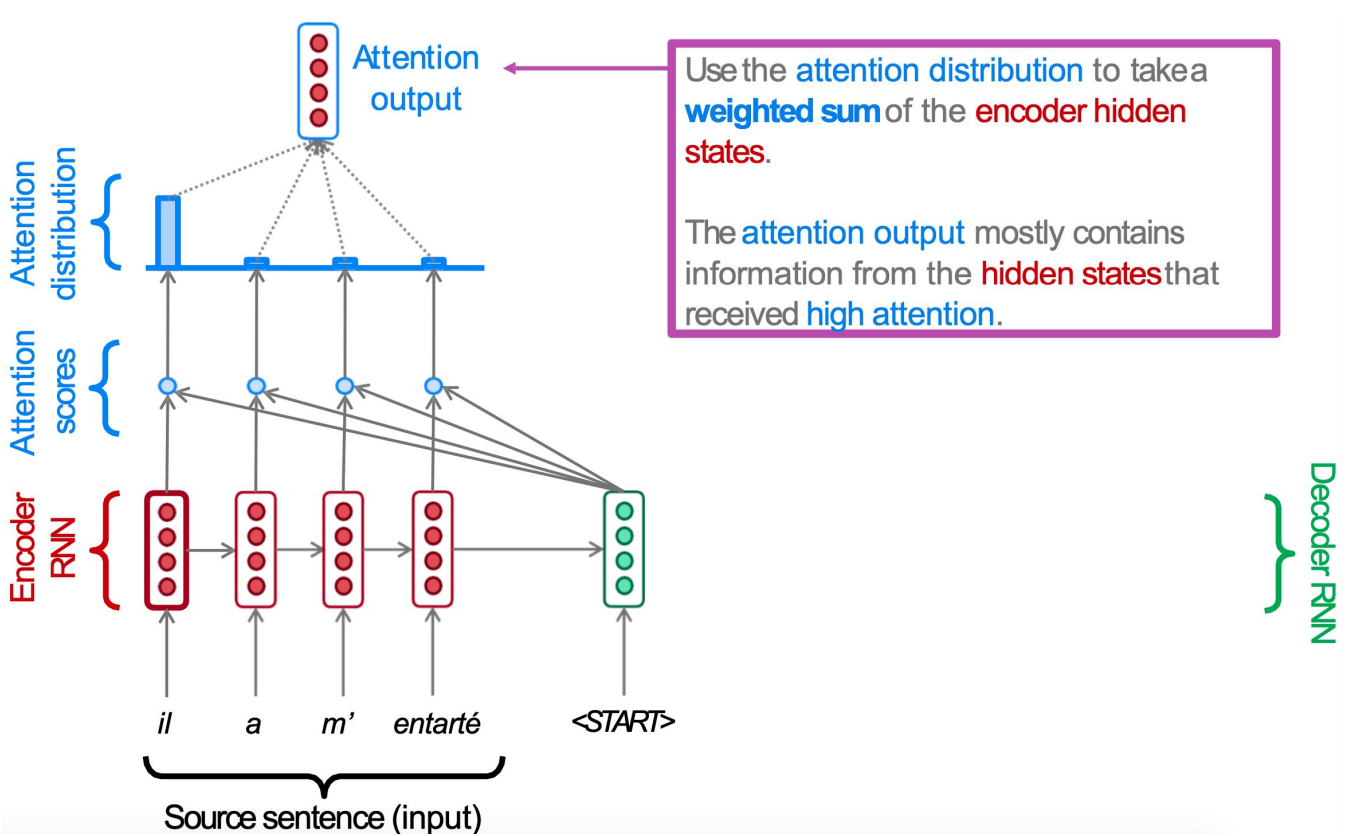
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



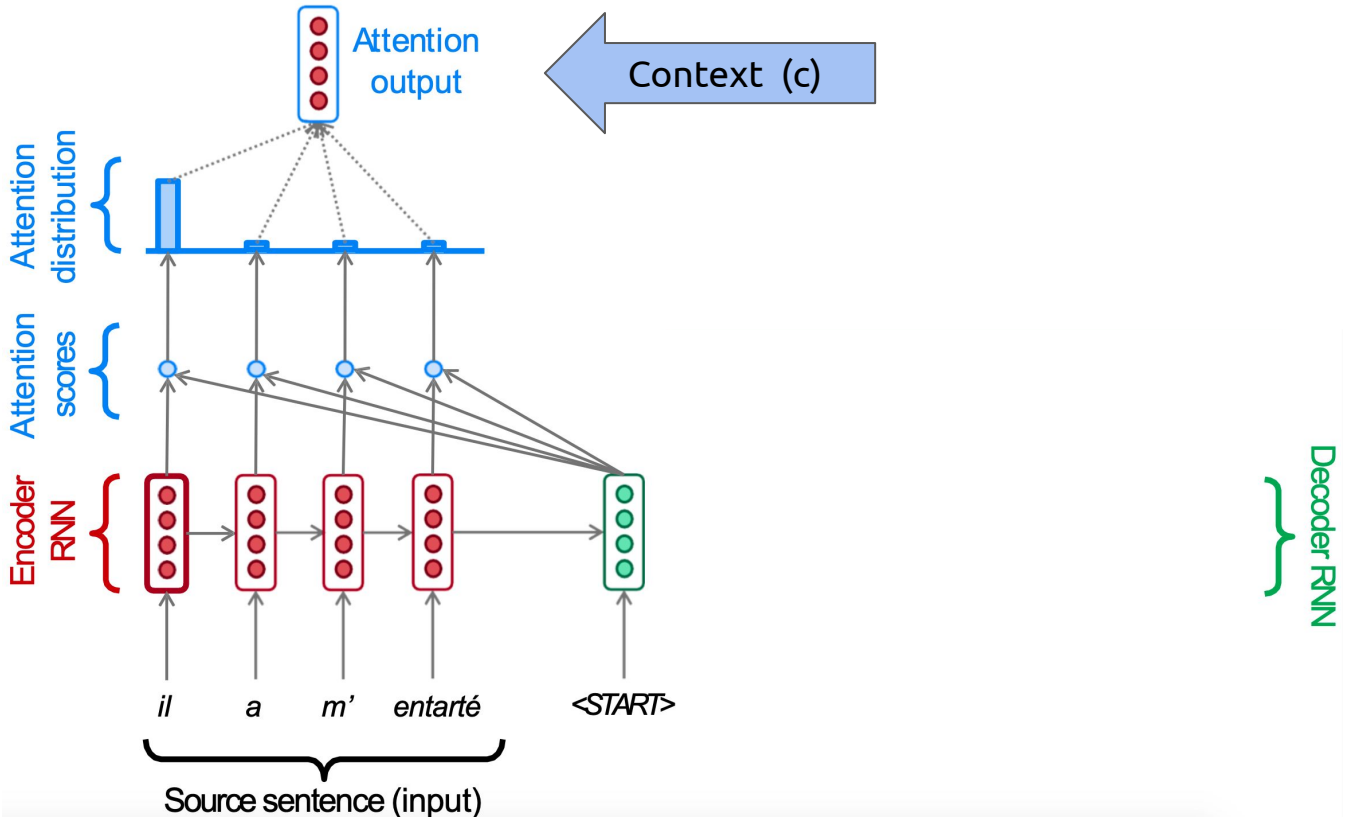
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



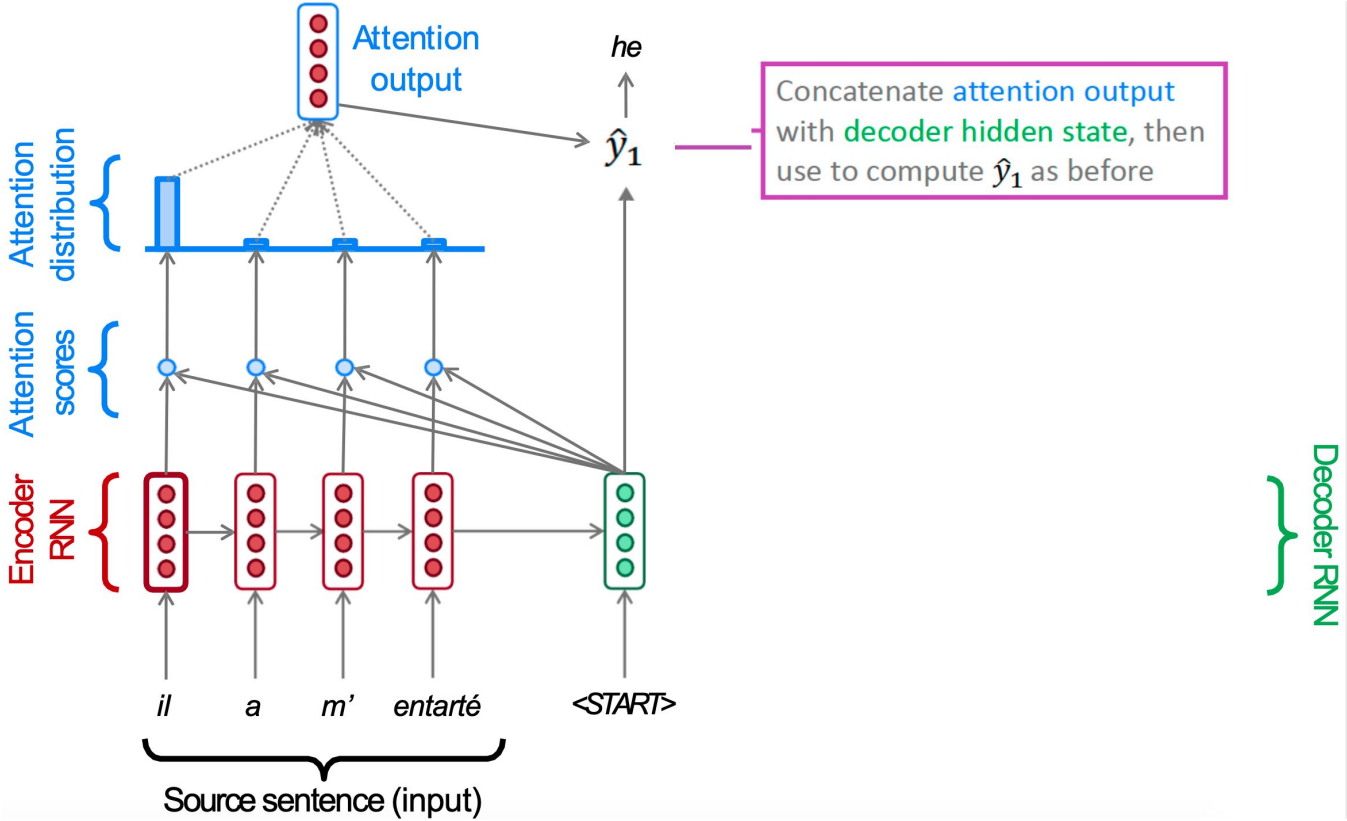
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



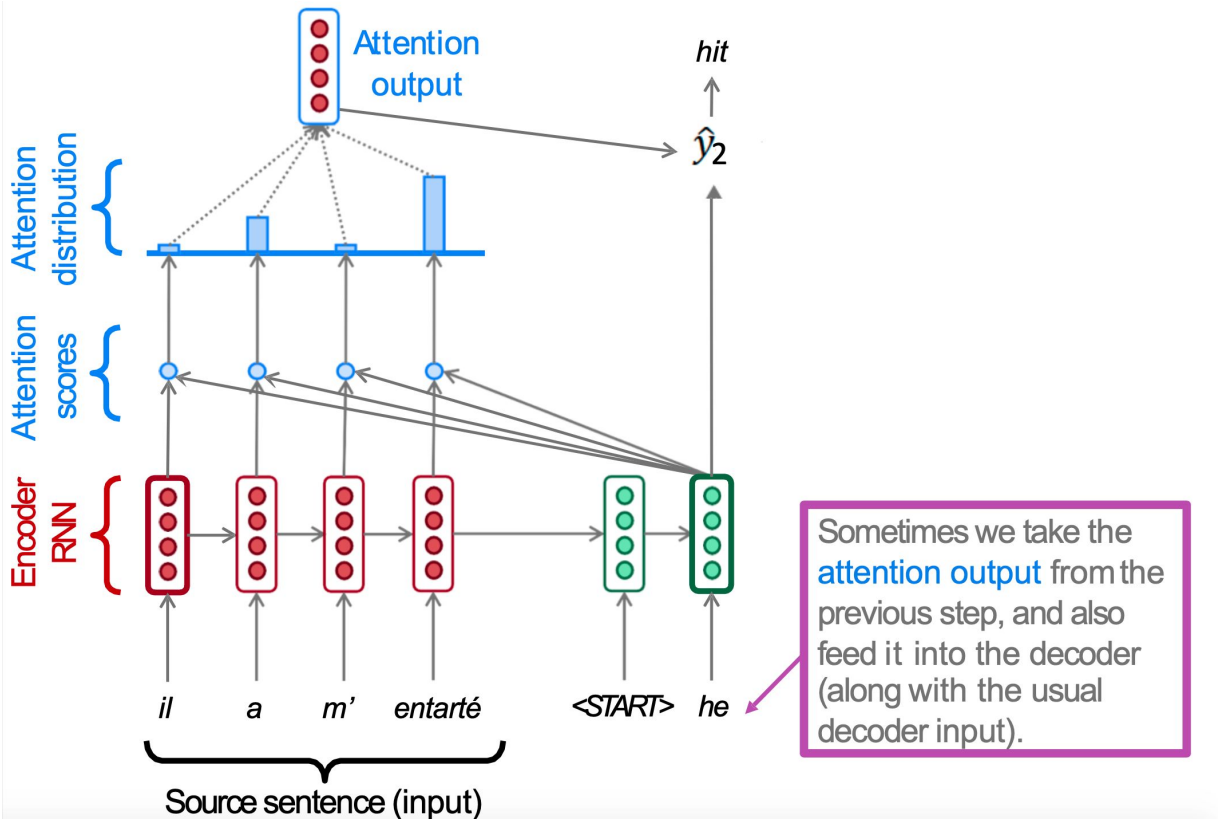
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



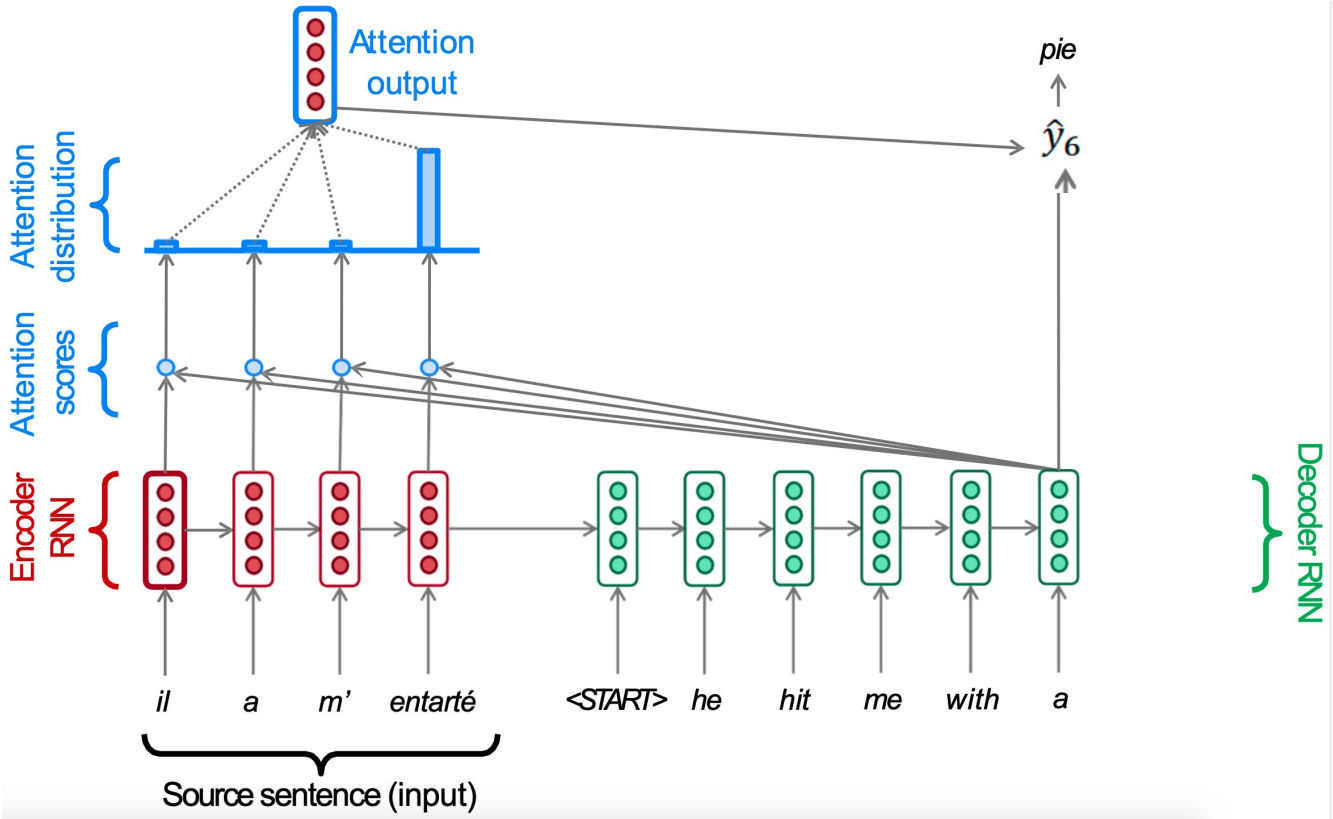
Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

RNN + Attention: NMT



Slide: Marta R. Costa-jussà

RNN + Attention: NMT

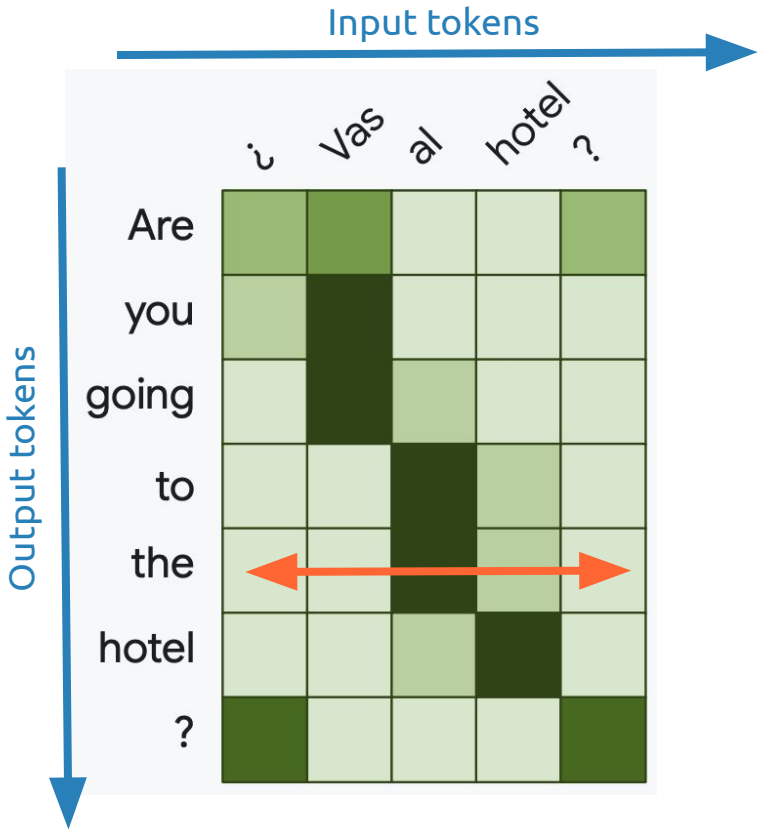


Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

Attention matrix

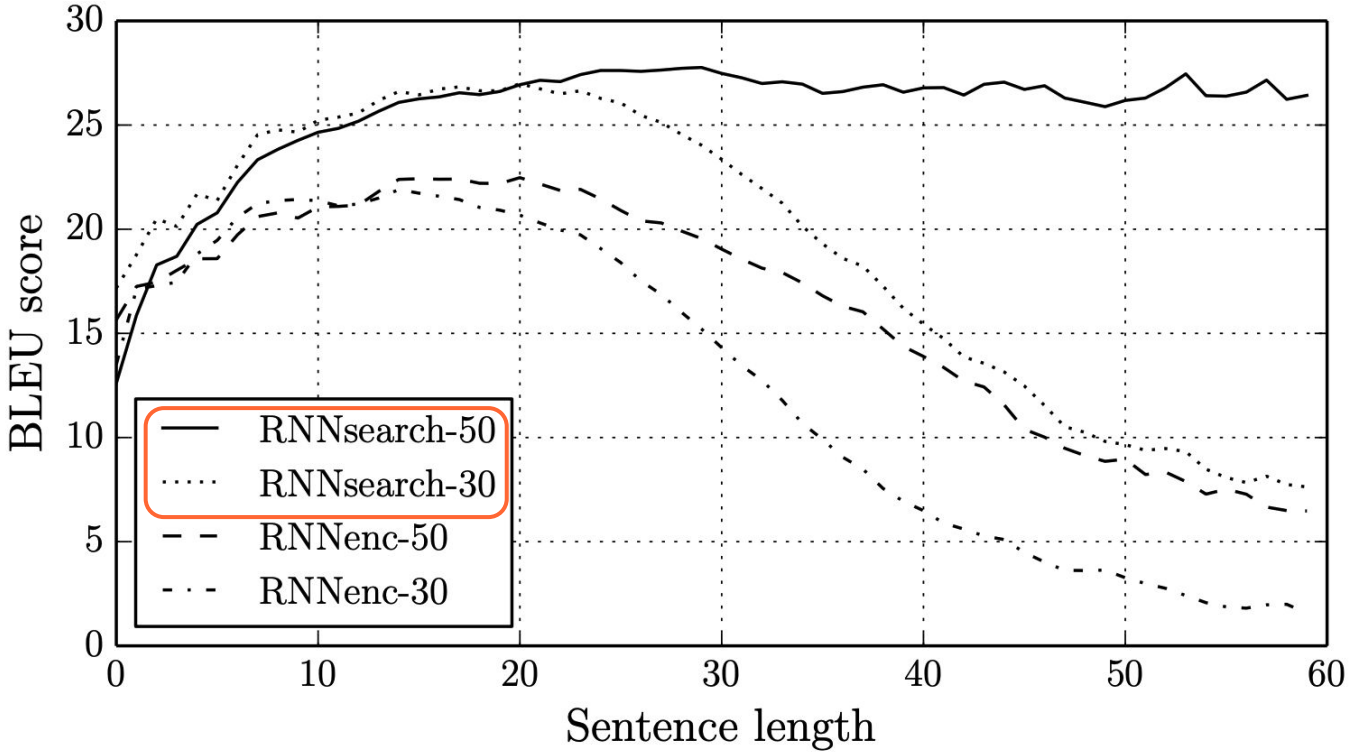
The **attention matrix** provides a visual representation of which input tokens are attended to produce each output token.

In which direction is the softmax normalization applied over the attention matrix: rows or columns ?



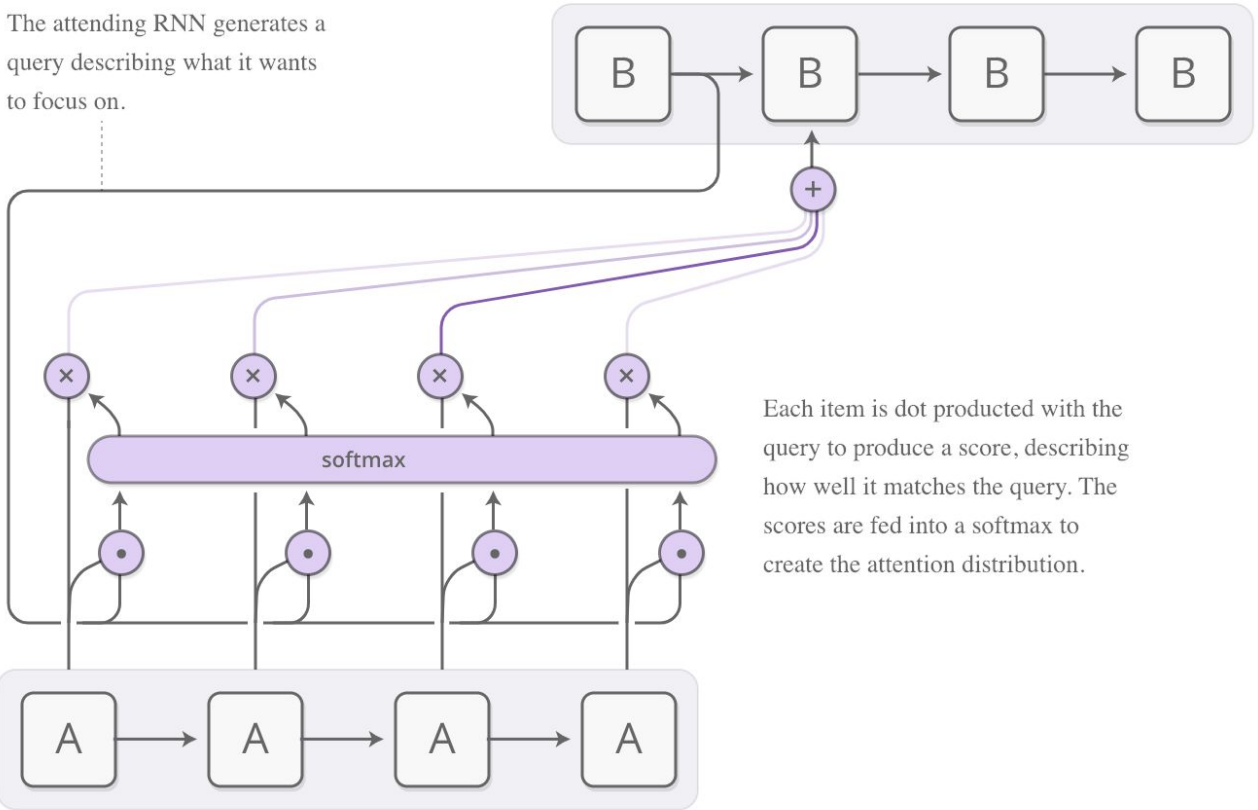
Example: Ilharco, Ilharco, Turc, Dettmers, Ferreira, Lee, "High Performance Natural Language Processing". EMNLP 2020.

Seq2Seq + Attention: NMT



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

Seq2Seq + Attention

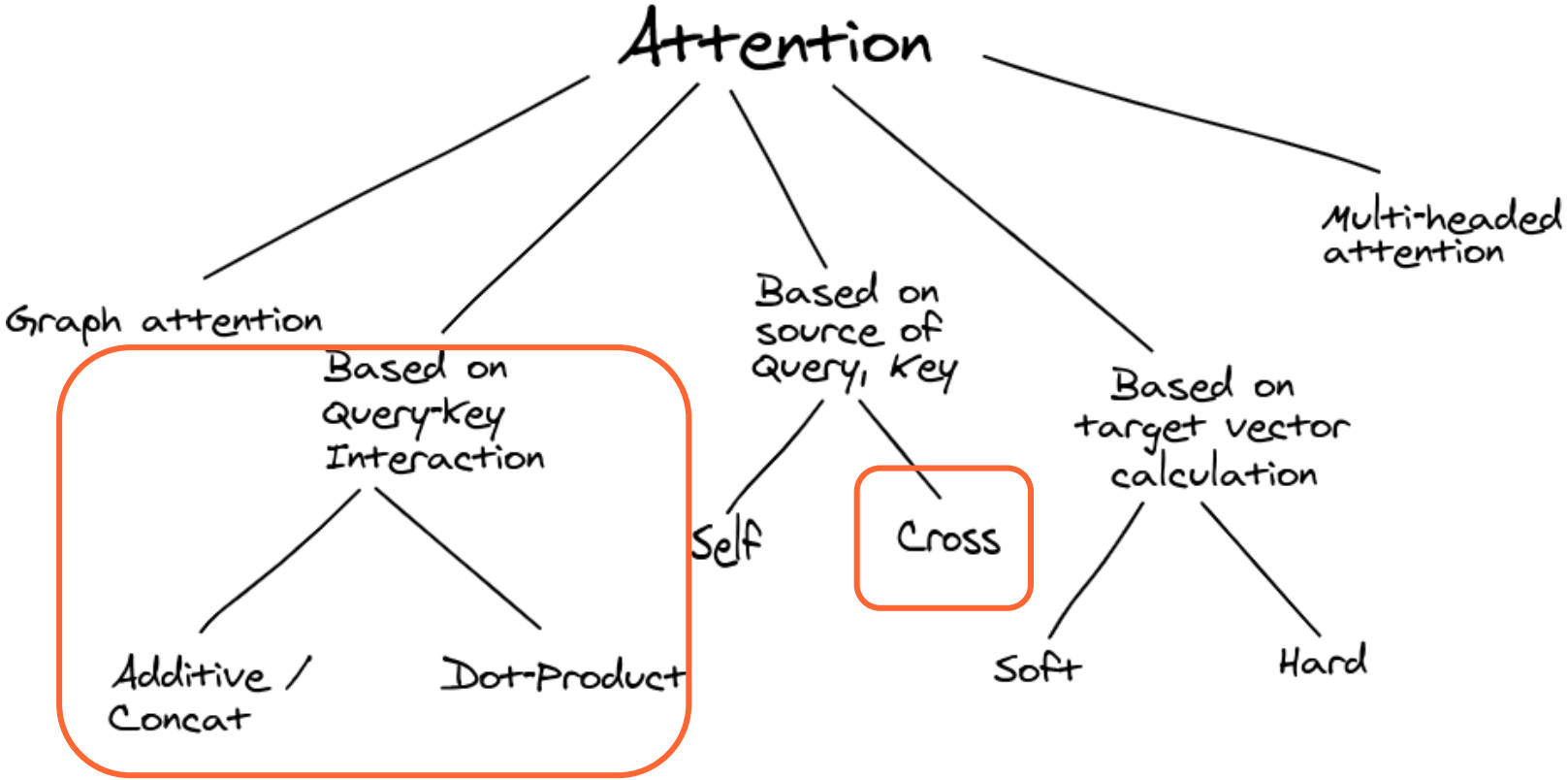


Chis Olah & Shan Cate, [“Attention and Augmented Recurrent Neural Networks”](#) distill.pub 2016.

Outline

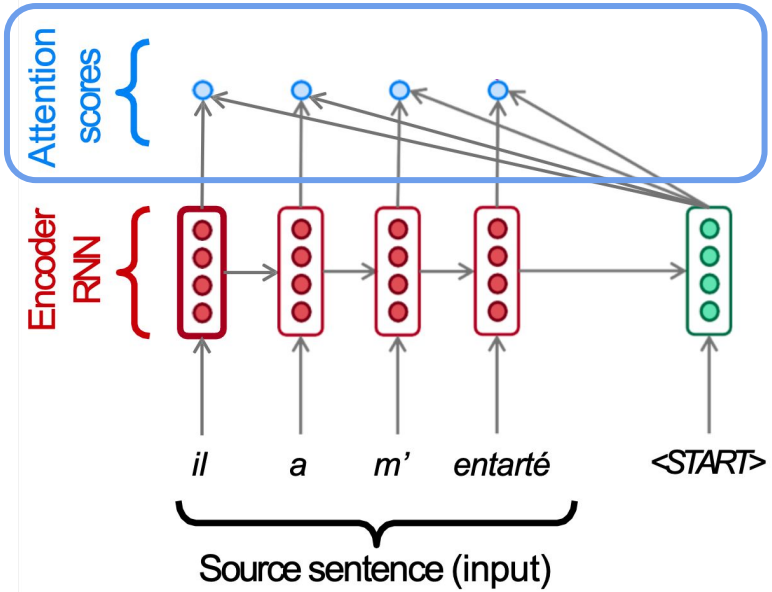
1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. **Attention mechanisms**
6. Study case: Image captioning

Attention Mechanisms



Nikhil Sha, "Attention ? An other Perspective!". 2020.

Attention Mechanisms



How to compute these attention scores ?

- Additive attention
- Multiplicative
- Scaled dot product

Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

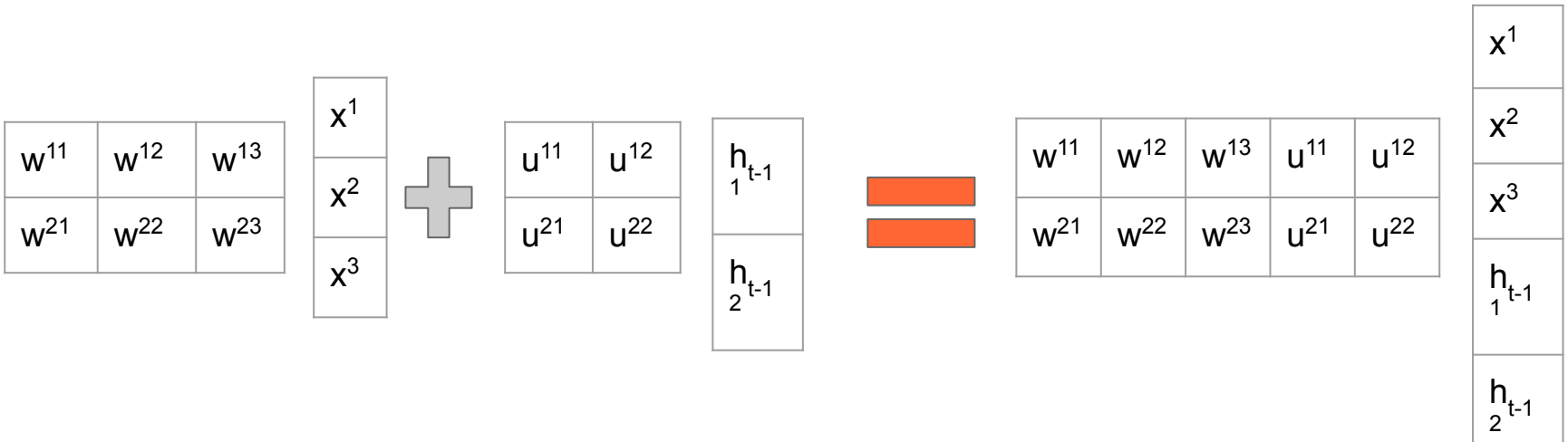
Outline

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

Reminder from RNN

Prove that a concatenation of vectors $\mathbf{h}_{t-1}$ and $\mathbf{x}_t$ is equivalent to the following expression:

$$h_t = f(W \cdot x_t + U \cdot h_{t-1} + b)$$

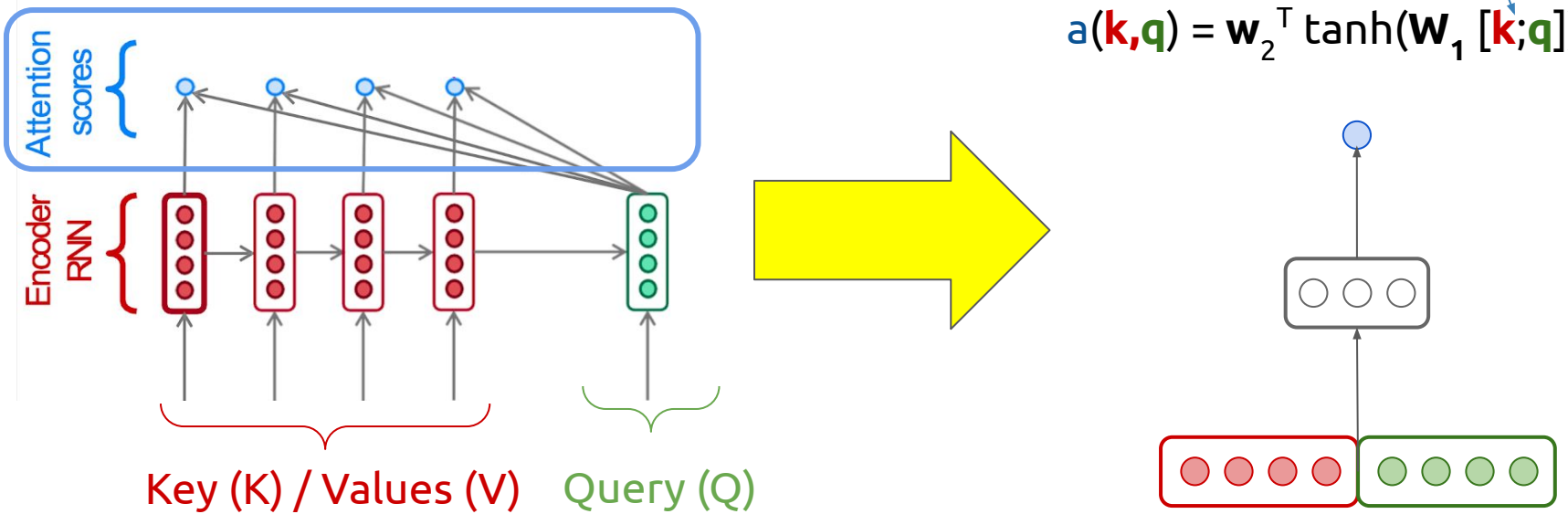


Attention Mechanisms: Additive Attention

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

concatenation

$$a(\mathbf{k},\mathbf{q}) = \mathbf{w}_2^T \tanh(\mathbf{W}_1 [\mathbf{k};\mathbf{q}])$$



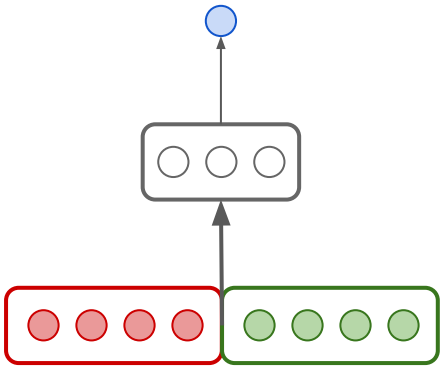
Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." ICLR 2015.

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.

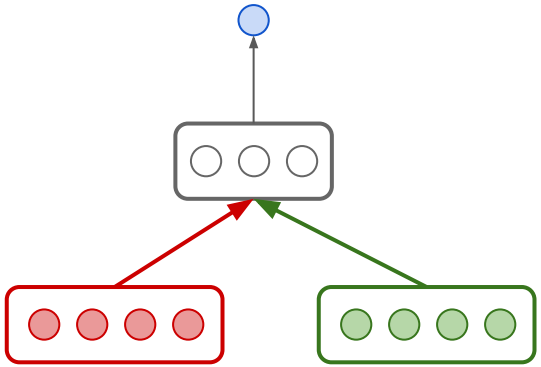
concatenation

$$a(\mathbf{q},\mathbf{k}) = \mathbf{w}_2^T \tanh(\mathbf{W}_1 [\mathbf{q};\mathbf{k}])$$



addition

$$a(\mathbf{q},\mathbf{k}) = \mathbf{w}_2^T \tanh(\mathbf{W}_q \mathbf{q} + \mathbf{W}_k \mathbf{k})$$



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

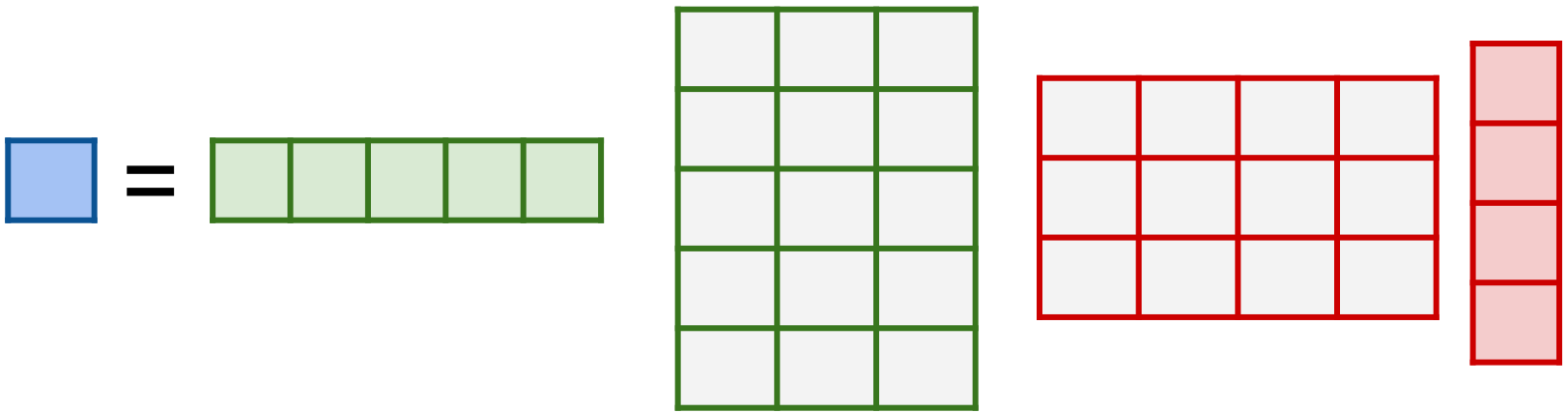
Outline

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

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.

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

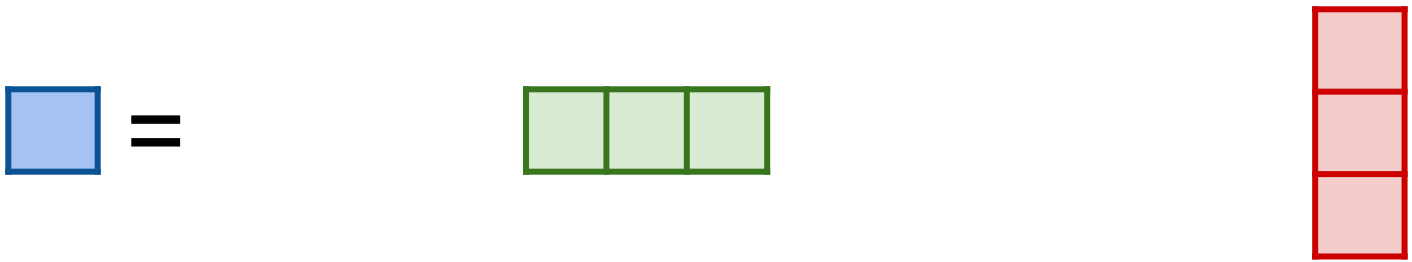


Luong, M.-T., Pham, H., & Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. EMNLP 2015

Attention Mechanisms: Dot Product

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

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



Luong, M.-T., Pham, H., & Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. EMNLP 2015

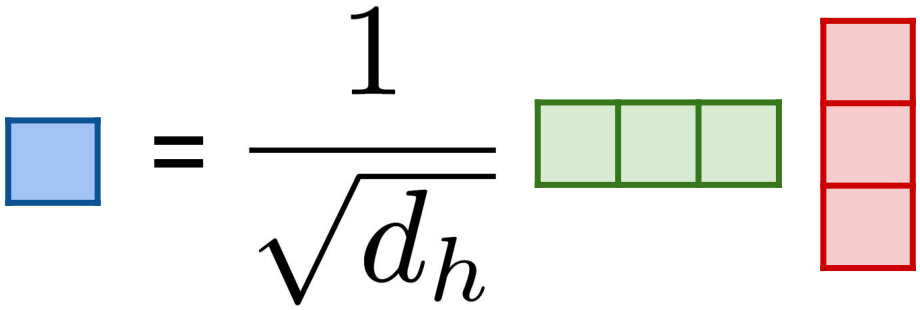
Outline

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

Attention Mechanisms: Scaled dot product

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

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



#Transformer Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.. Attention is all you need. NeurIPS 2017.

Attention Mechanisms: Scaled dot product

Considering a dot-product attention, how are the attention scores (a_{ij}) between each pair of input token x_i and output token y_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 token x_i and output token y_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)
$h(x) = W^V x$

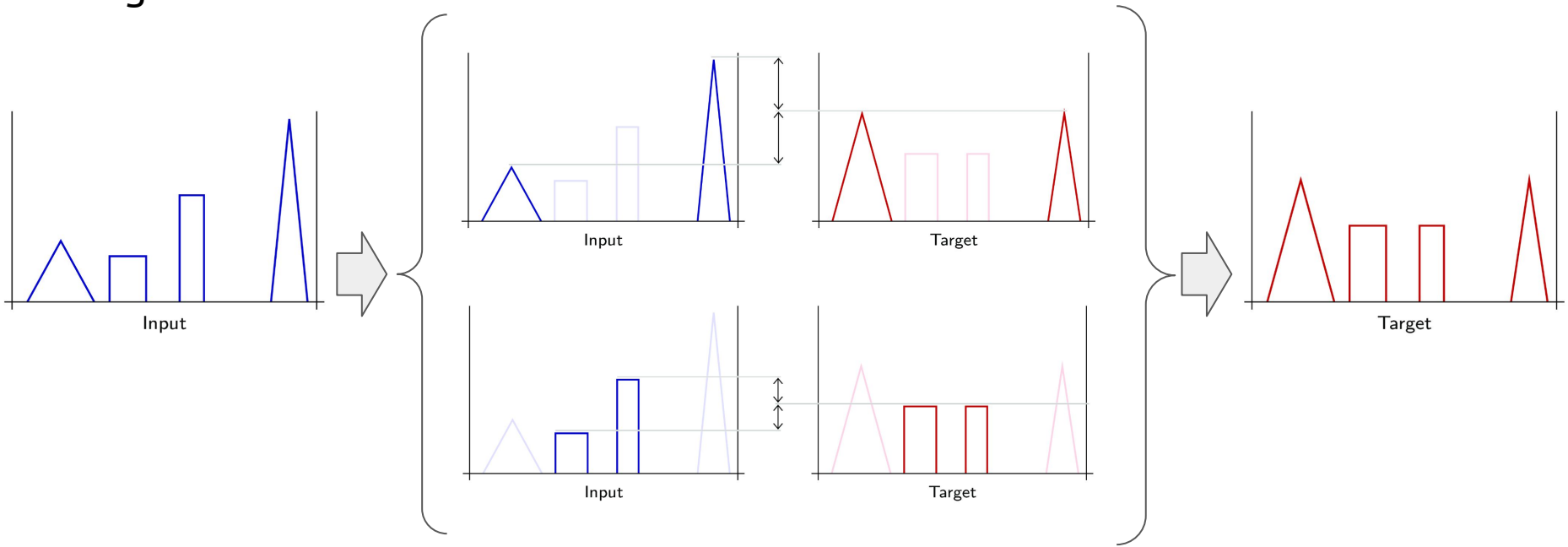
W^Q, W^K and W^V are projection layers shared across all words.

Outline

1. Motivation
2. Seq2Seq
3. Key, Query and Value
4. Seq2Seq + Attention
5. Attention mechanisms
6. **Toy Example**

Toy Example: Pulse equalization by shape group

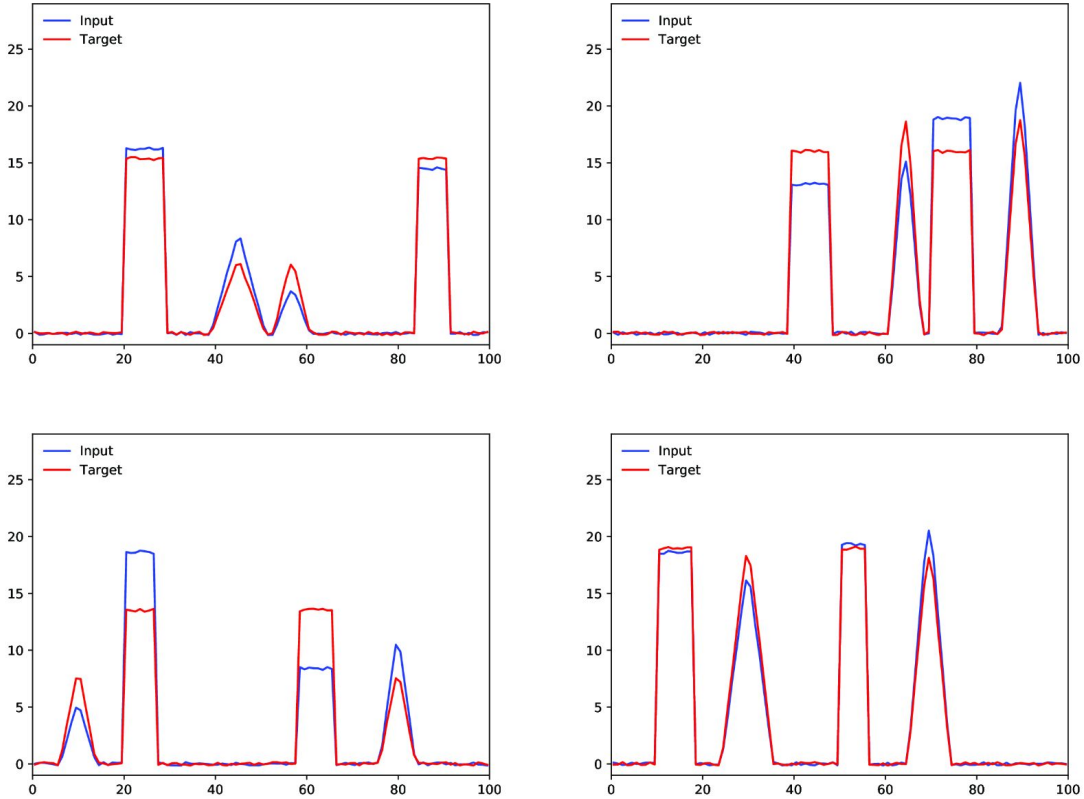
Task: Translate a 1D time series composed of two triangular impulses and to rectangular impulses so that their heights are equalized in each shape group to their average.



Source: François Floret, "Deep Learning" EPFL EE-559 2020 [tweet] [code]

Toy Example: Pulse equalization by shape group

Samples from the training dataset:



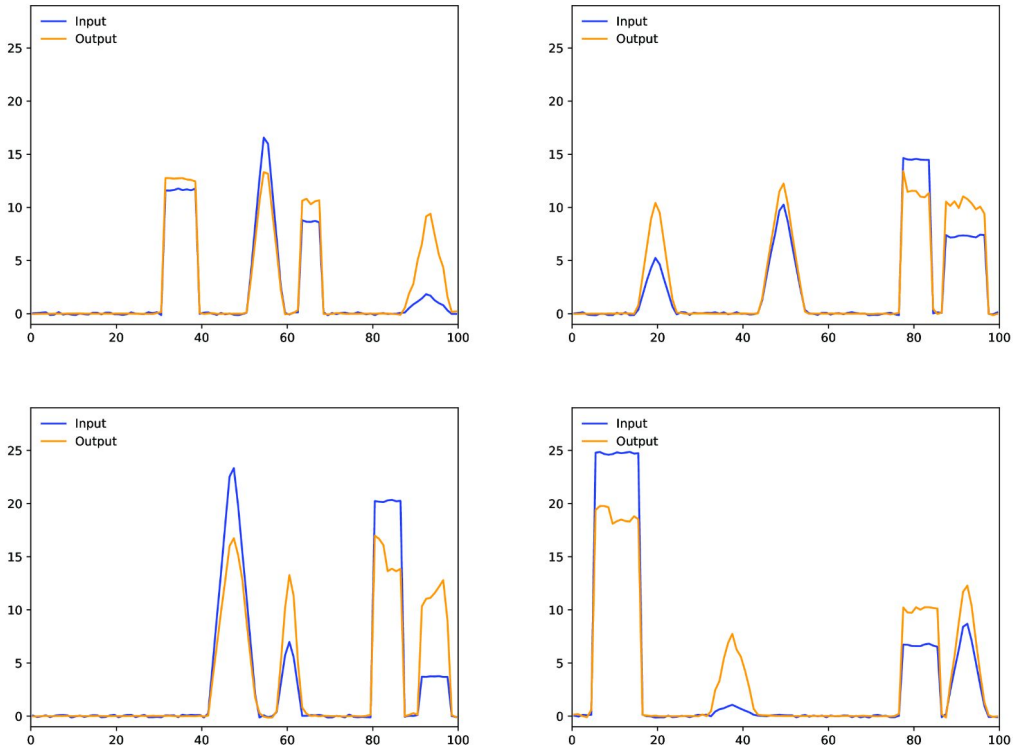
Source: François Floret, "Deep Learning" EPFL EE-559 2020 [tweet] [code]

Toy Example: Pulse equalization by shape group

Poor results with a 1D CNN network with **no attention**:

```
Sequential(
  (0): Conv1d(1, 64, kernel_size=(5,), stride=(1,), padding=(2,))
  (1): ReLU()
  (2): Conv1d(64, 64, kernel_size=(5,), stride=(1,), padding=(2,))
  (3): ReLU()
  (4): Conv1d(64, 64, kernel_size=(5,), stride=(1,), padding=(2,))
  (5): ReLU()
  (6): Conv1d(64, 64, kernel_size=(5,), stride=(1,), padding=(2,))
  (7): ReLU()
  (8): Conv1d(64, 1, kernel_size=(5,), stride=(1,), padding=(2,))
)

nb_parameters 62337
```



Source: François Floret, "Deep Learning" EPFL EE-559 2020 [tweet] [code]

Toy Example: Pulse equalization by shape group

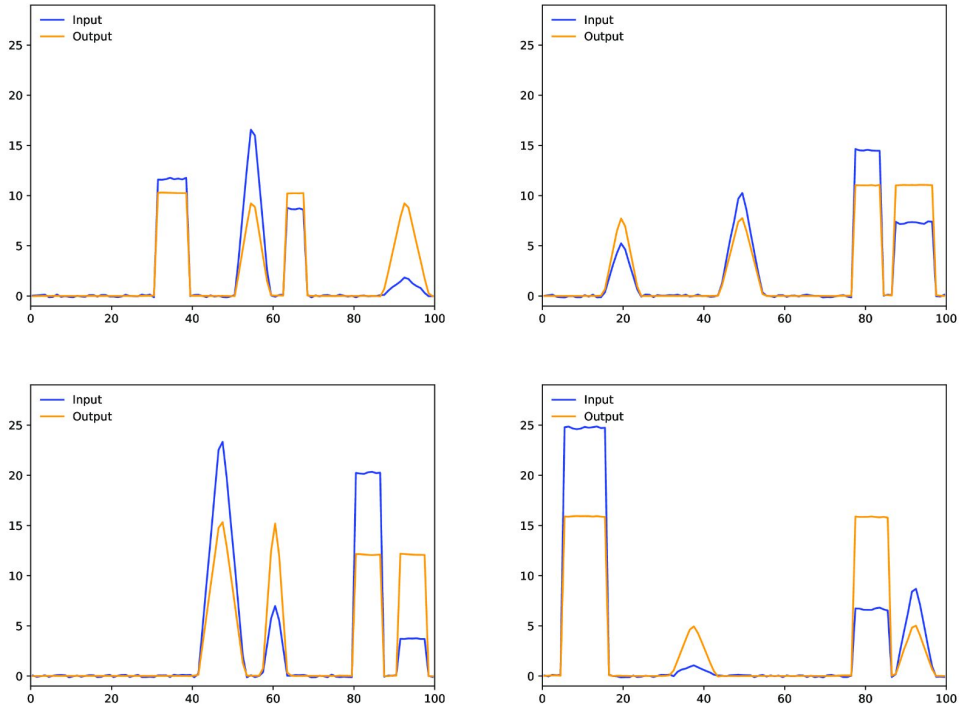
Good results with a 1D CNN network with **attention**:

```
class AttentionLayer(nn.Module):
    def __init__(self, in_channels, out_channels, key_channels):
        super(AttentionLayer, self).__init__()
        self.conv_Q = nn.Conv1d(in_channels, key_channels, kernel_size = 1, bias = False)
        self.conv_K = nn.Conv1d(in_channels, key_channels, kernel_size = 1, bias = False)
        self.conv_V = nn.Conv1d(in_channels, out_channels, kernel_size = 1, bias = False)

    def forward(self, x):
        Q = self.conv_Q(x)
        K = self.conv_K(x)
        V = self.conv_V(x)
        A = Q.permute(0, 2, 1).matmul(K).softmax(2)
        x = A.matmul(V.permute(0, 2, 1)).permute(0, 2, 1)
        return x

    def __repr__(self):
        return self._get_name() + \
            '(in_channels={}, out_channels={}, key_channels={})'.format(
                self.conv_Q.in_channels,
                self.conv_V.out_channels,
                self.conv_K.out_channels
            )
```

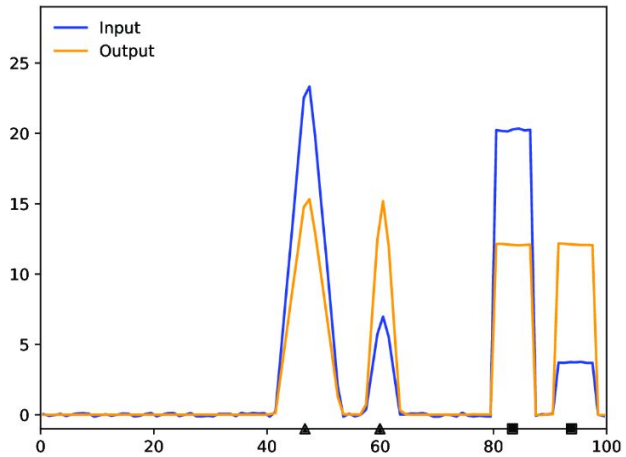
```
Sequential(
  (0): Conv1d(1, 64, kernel_size=(5,), stride=(1,), padding=(2,))
  (1): ReLU()
  (2): Conv1d(64, 64, kernel_size=(5,), stride=(1,), padding=(2,))
  (3): ReLU()
  (4): AttentionLayer(in_channels=64, out_channels=64, key_channels=64)
  (5): Conv1d(64, 64, kernel_size=(5,), stride=(1,), padding=(2,))
  (6): ReLU()
  (7): Conv1d(64, 1, kernel_size=(5,), stride=(1,), padding=(2,))
)
```

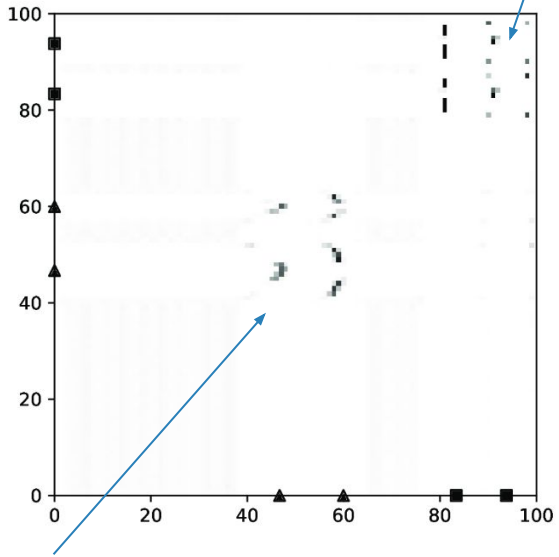


Source: François Floret, "Deep Learning" EPFL EE-559 2020 [tweet] [code]

Toy Example: Pulse equalization by shape group

The attention matrix shows how triangles attend to triangles, and squares to squares:





Squares attend to squares

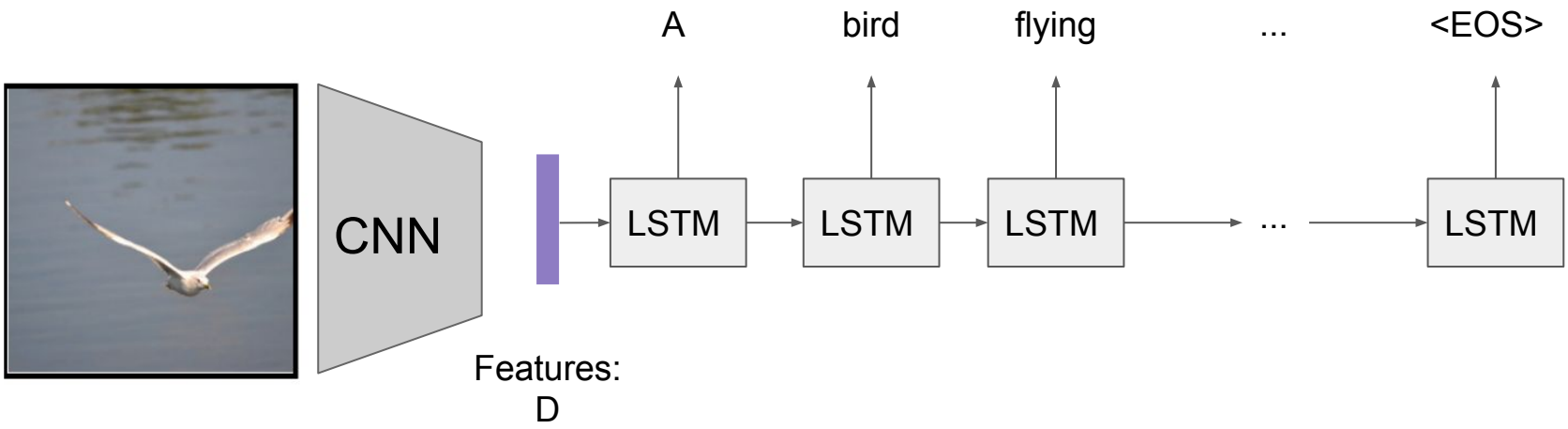
Triangles attend to triangles

Source: François Floret, "Deep Learning" EPFL EE-559 2020 [tweet] [code]

Outline

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

Image Captioning without Attention

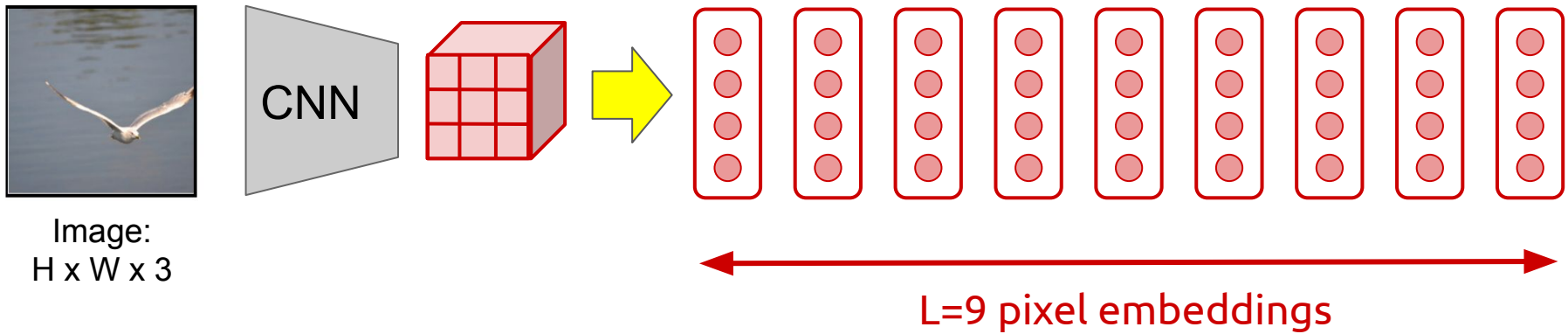


Limitation: All output predictions are based on the **final and static** output of the encoder

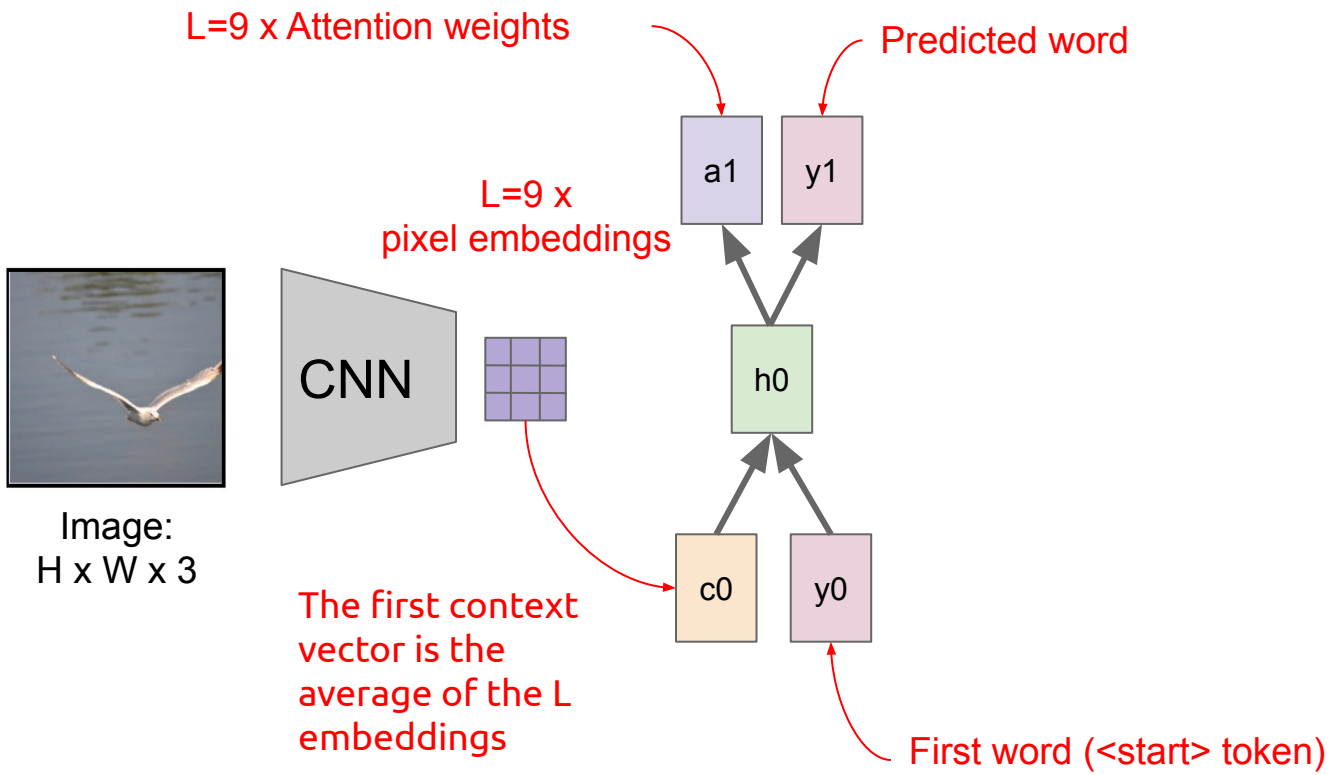
Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. CVPR 2015.

Image Captioning with Attention

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



Attention for Image Captioning



Slide: Amaia Salvador

Attention for Image Captioning

Image:
H x W x 3

CNN

a1 y1 a2 y2

h0 h1

c0 y0 c1 y1

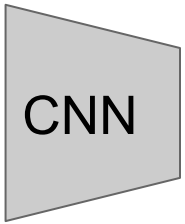
Visual features weighted with attention give the next context

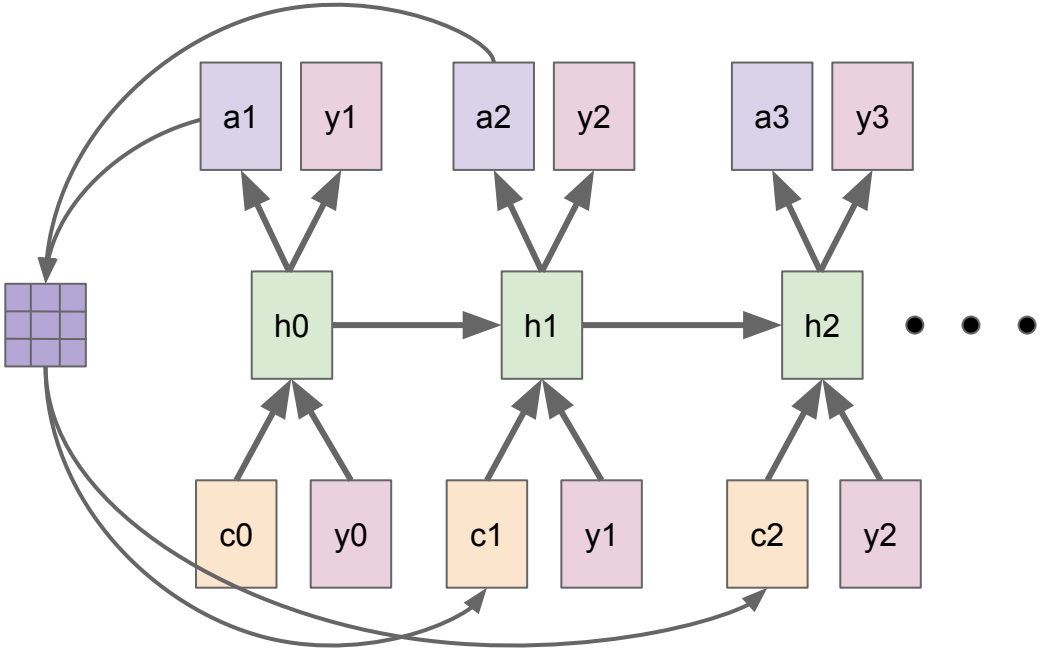
Predicted word in previous timestep

Slide: Amaia Salvador

Attention for Image Captioning

Image:
H x W x 3





Slide: Amaia Salvador

Attention for Image Captioning

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "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.

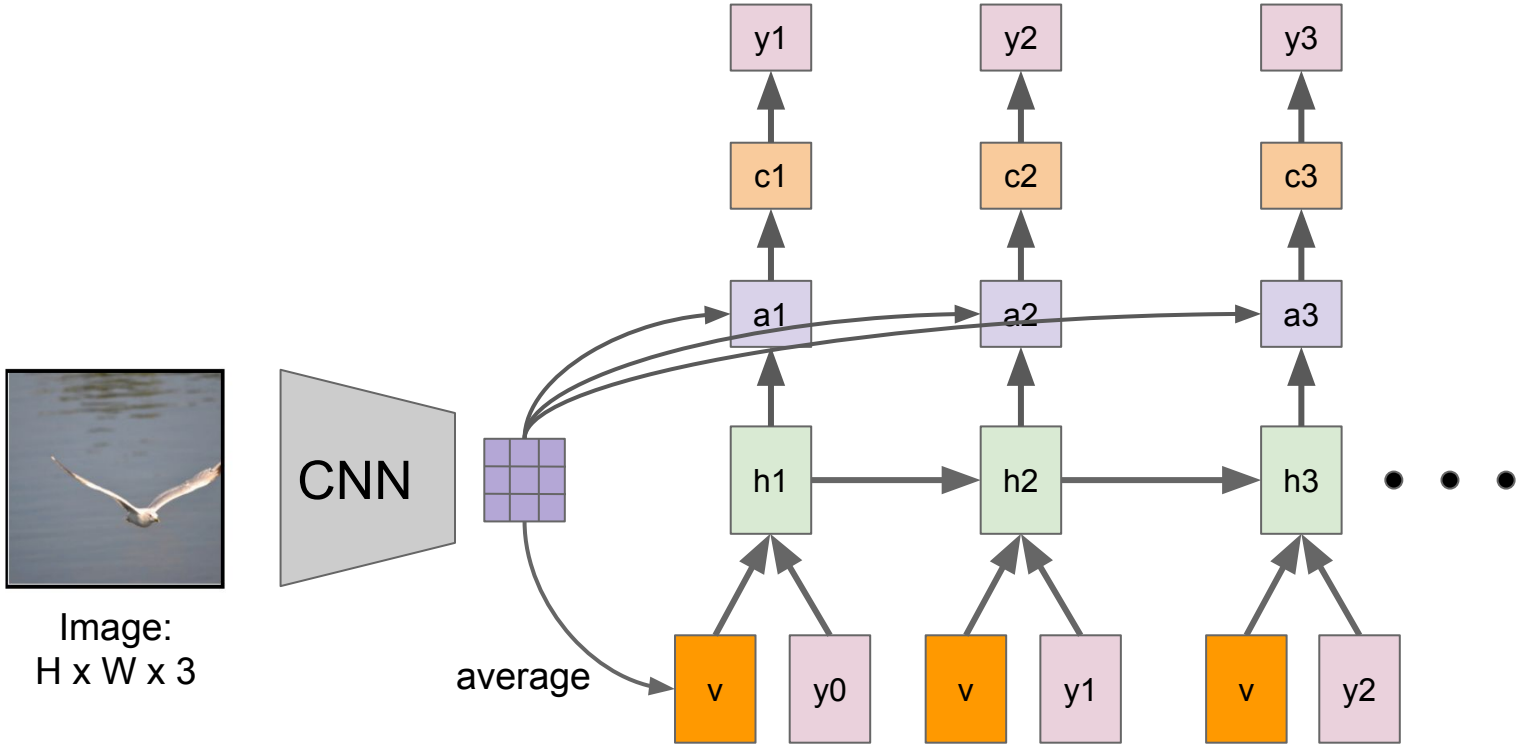
A group of people sitting on a boat in the water.

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

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." ICML 2015

Attention for Image Captioning

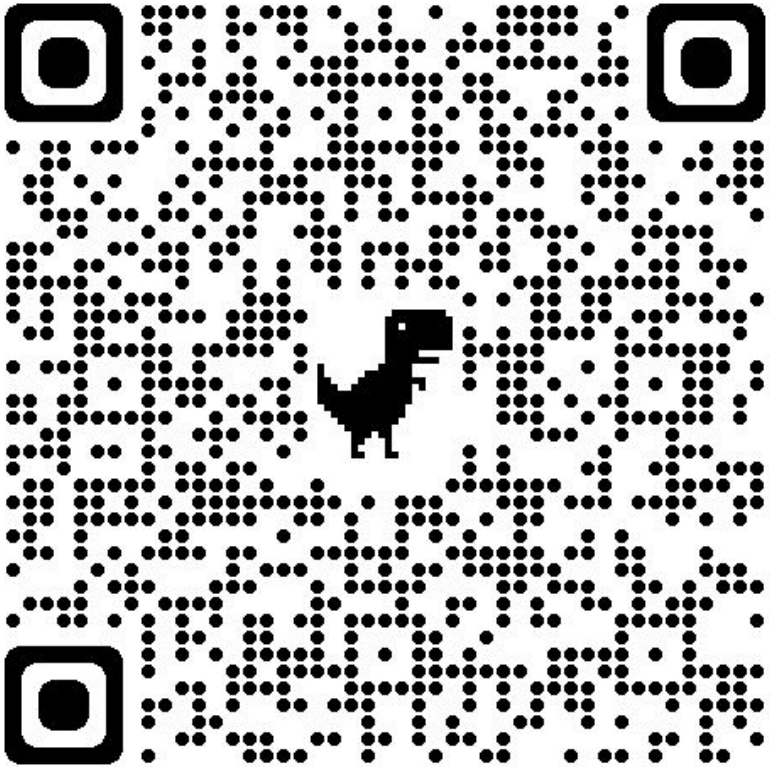
Side-note: attention can be computed with previous or **current** hidden state



Outline

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

Quizz



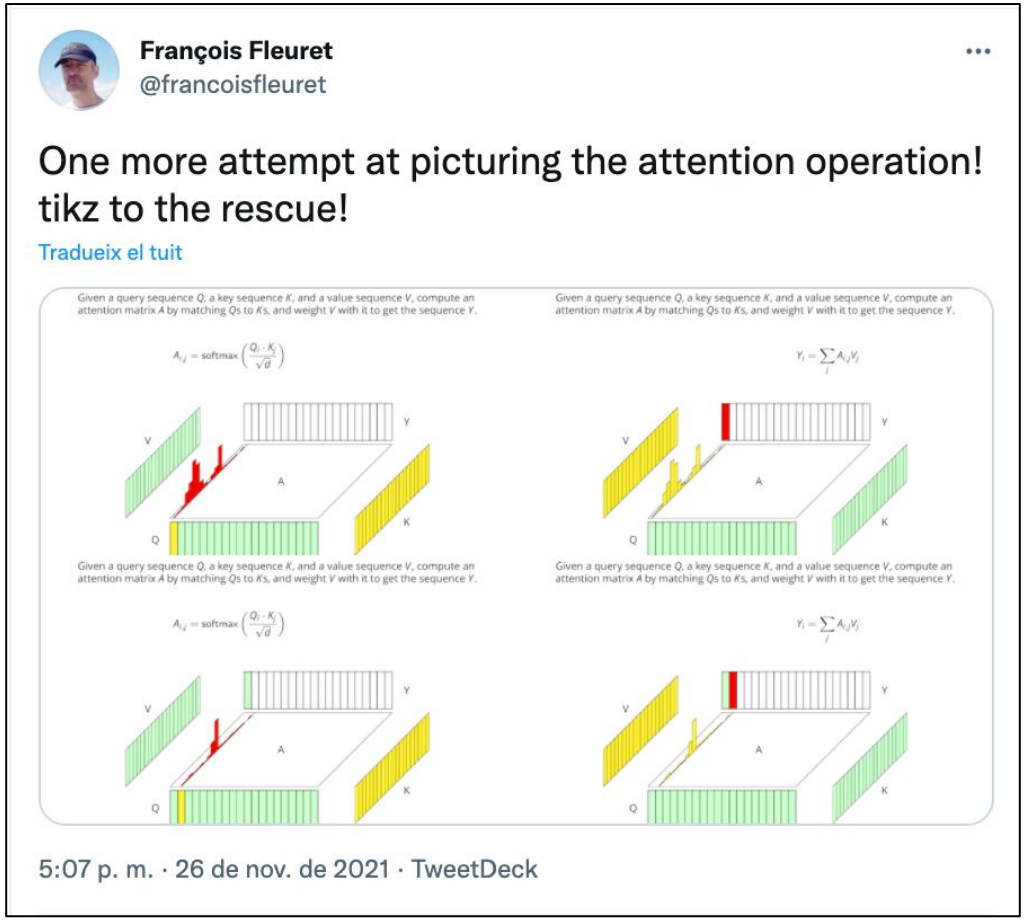
Learn more

Alex Graves, UCL x Deepmind 2020 [slides]

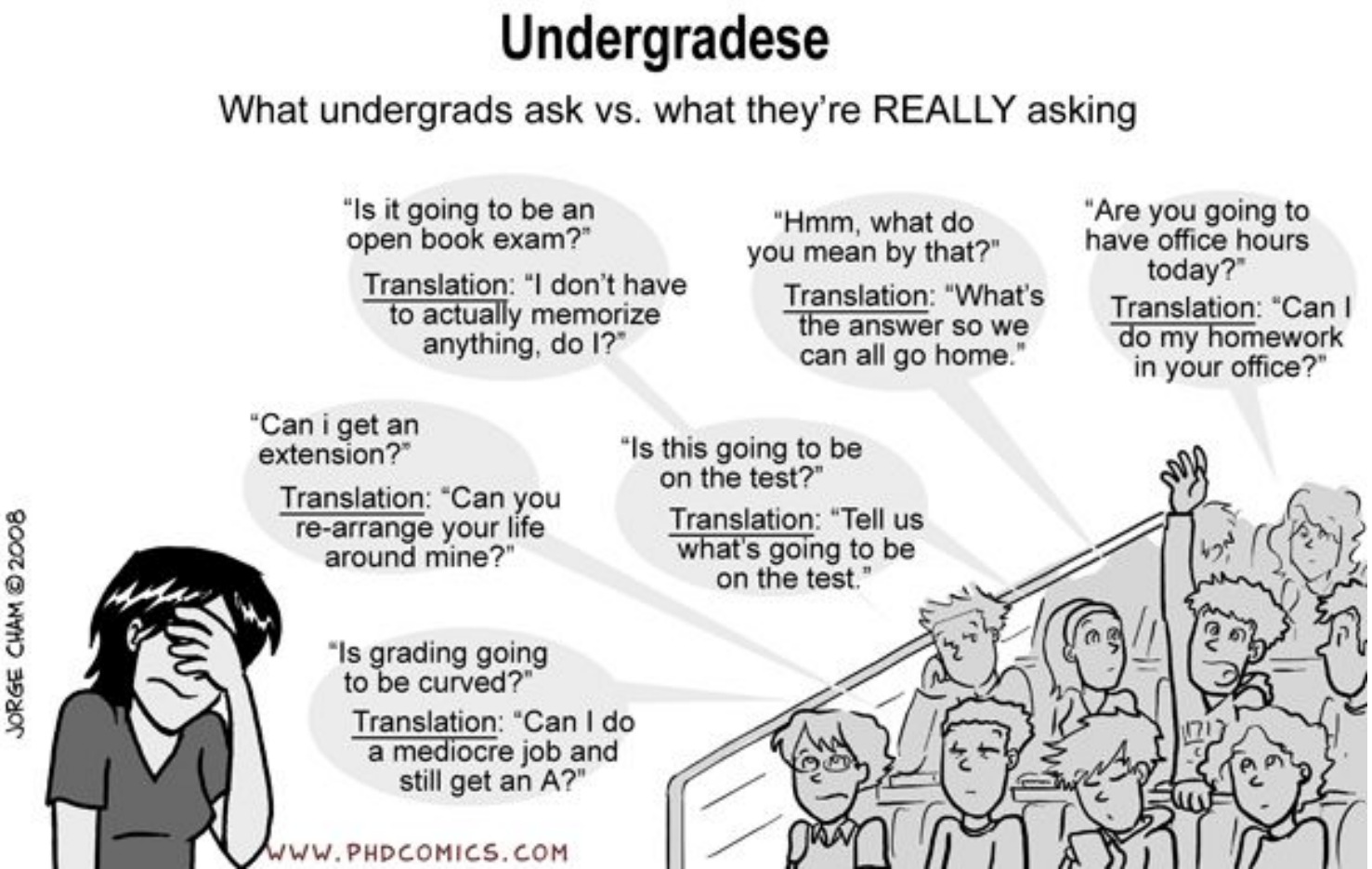
Learn more

- 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.

Tweetverse


François Fleuret
@francoisfleuret
One more attempt at picturing the attention operation! tikz to the rescue!
Tradueix el tuit
Given a query sequence Q, a key sequence K, and a value sequence V, compute an attention matrix A by matching Qs to Ks, and weight V with it to get the sequence Y.
$A_{i,j} = \text{softmax}\left(\frac{Q_i \cdot K_j}{\sqrt{d}}\right)$
$Y_i = \sum_j A_{i,j} V_j$
V
Y
A
Q
K
5:07 p. m. · 26 de nov. de 2021 · TweetDeck

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?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM