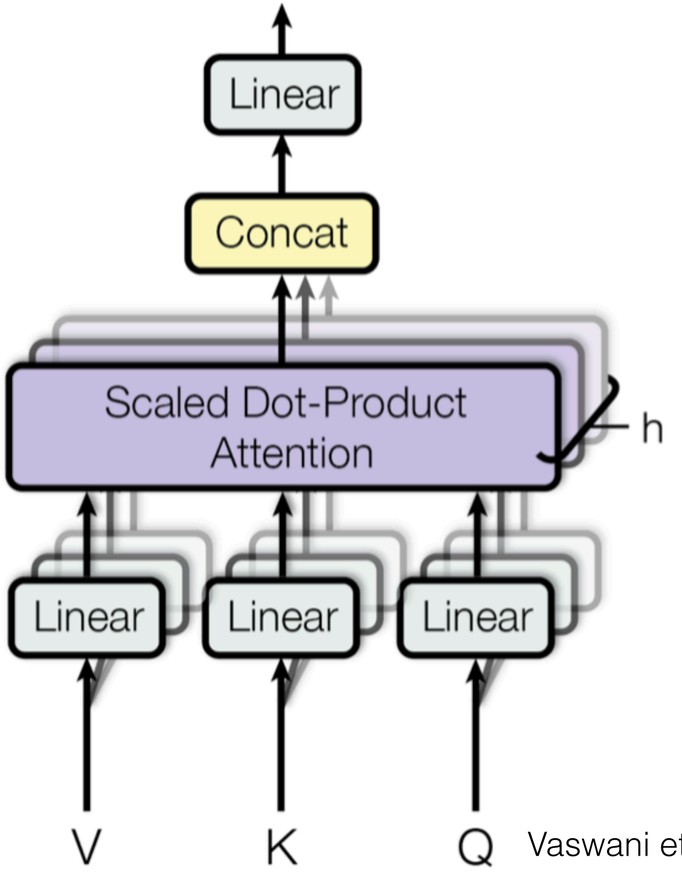
A Quick and Gentle Introduction to Transformers



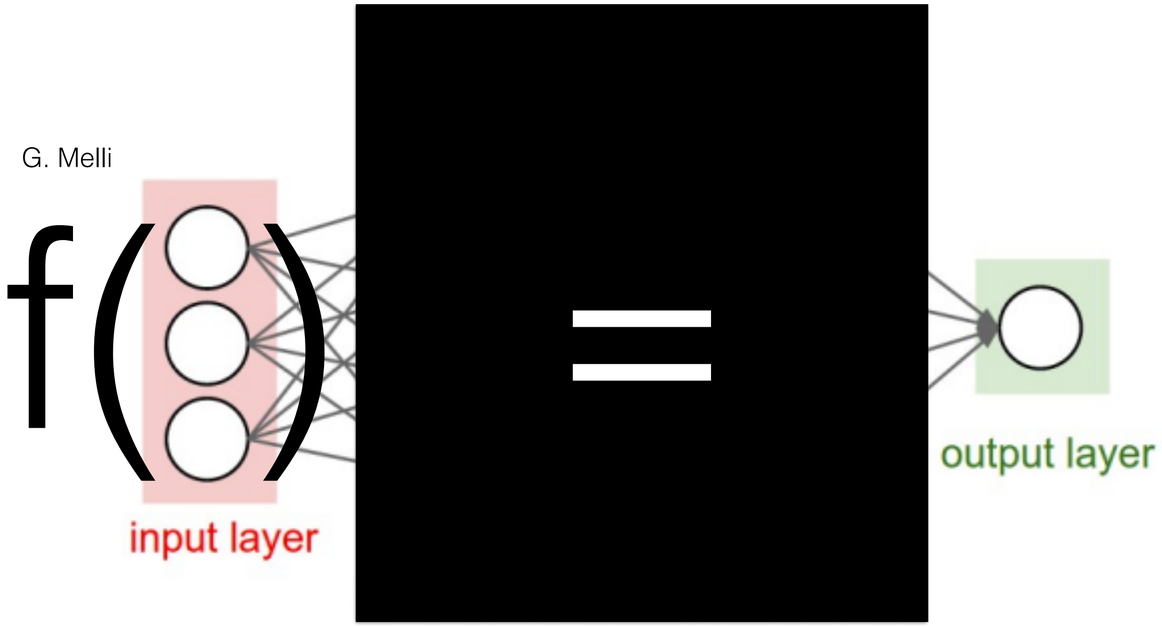
Vaswani et al. 2017

Rafael Martinez-Galarza

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

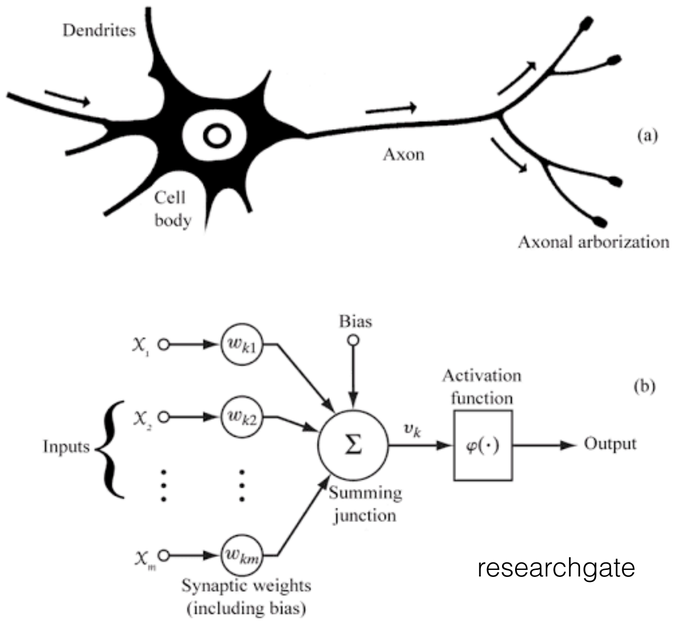
IVOA Interoperability Meeting. Bologna, May 2023

What is a neural network?



This is a neural network.
It is just a set of inter-connected units acting on the input, to predict an output.

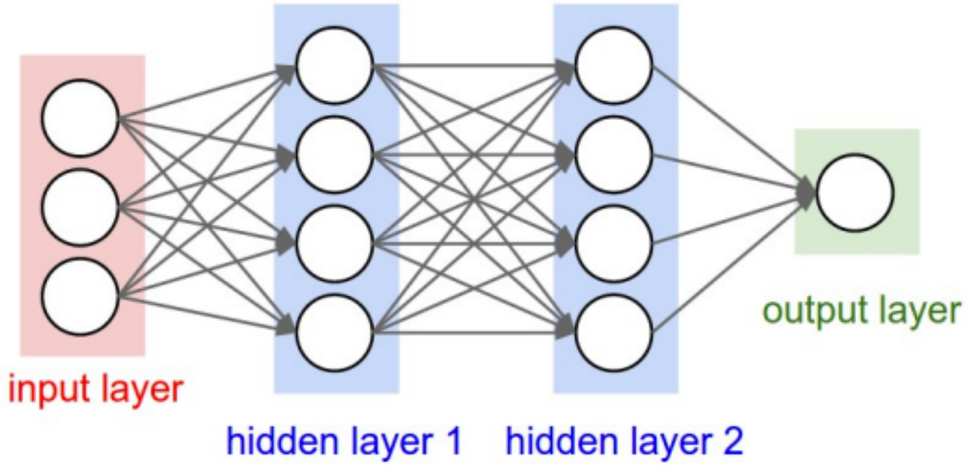
Usually trained by reducing the error between that prediction and the “ground truth” in the training set (backpropagation).



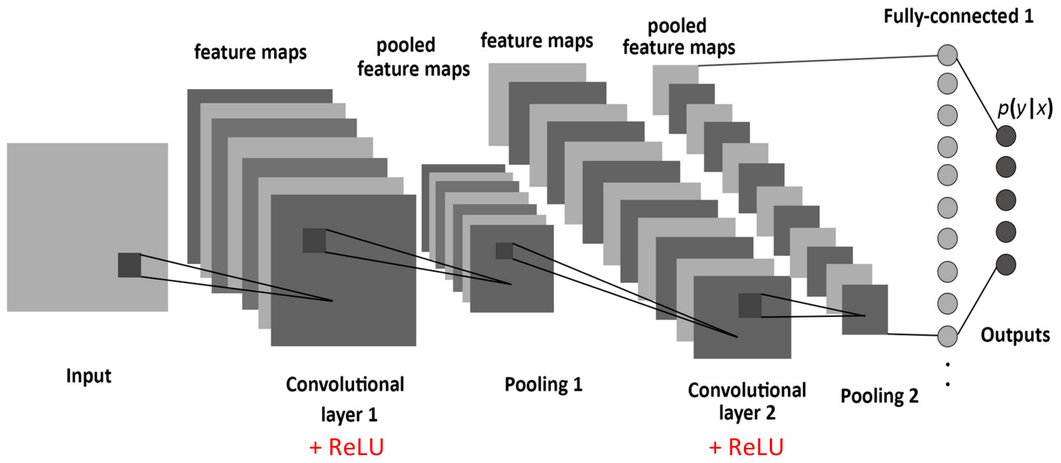
Inside the network, each unit (neuron) performs a series of linear transformations of the input, modified at each instance by a non linear activation function

Types of (deep) networks

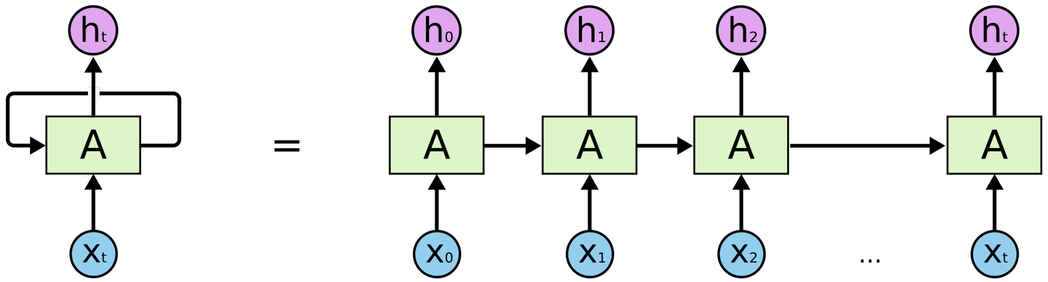
Fully connected



Convolutional



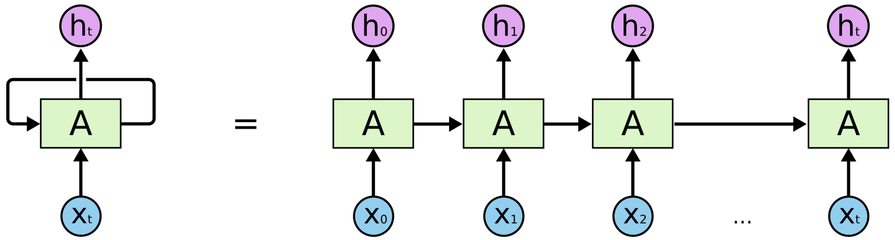
Recurrent



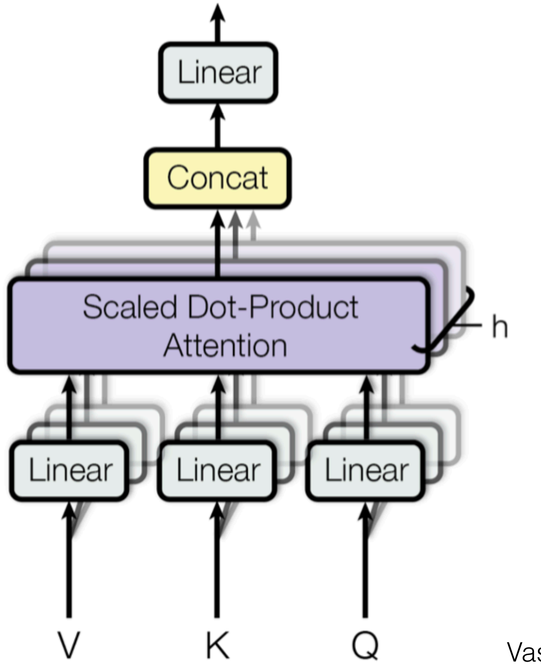
Credits: Google, G. Melli, C. Olah

Persistence vs. Attention

I grew up in France, I speak fluent…French



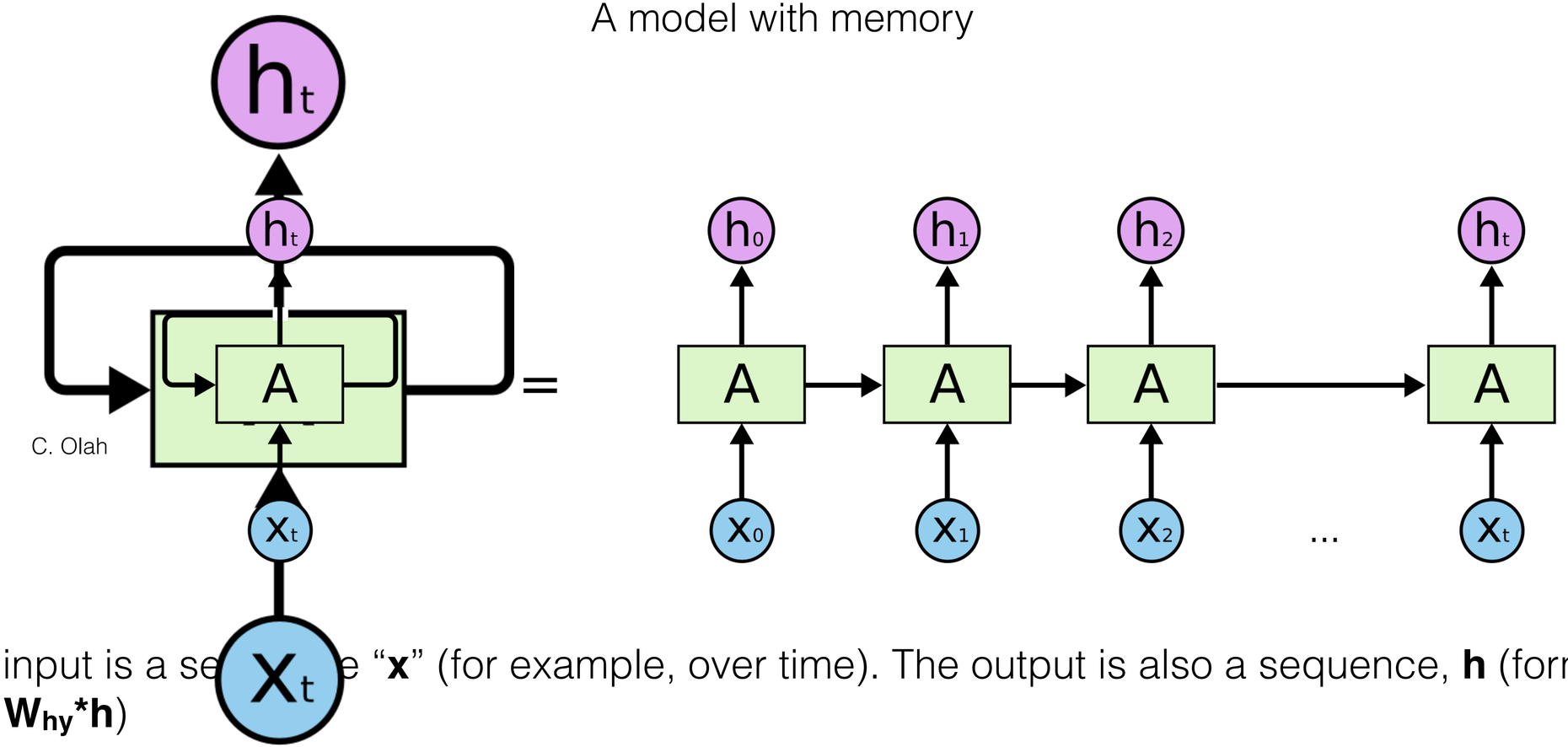
She is eating a green apple



Vaswani et al. 2017

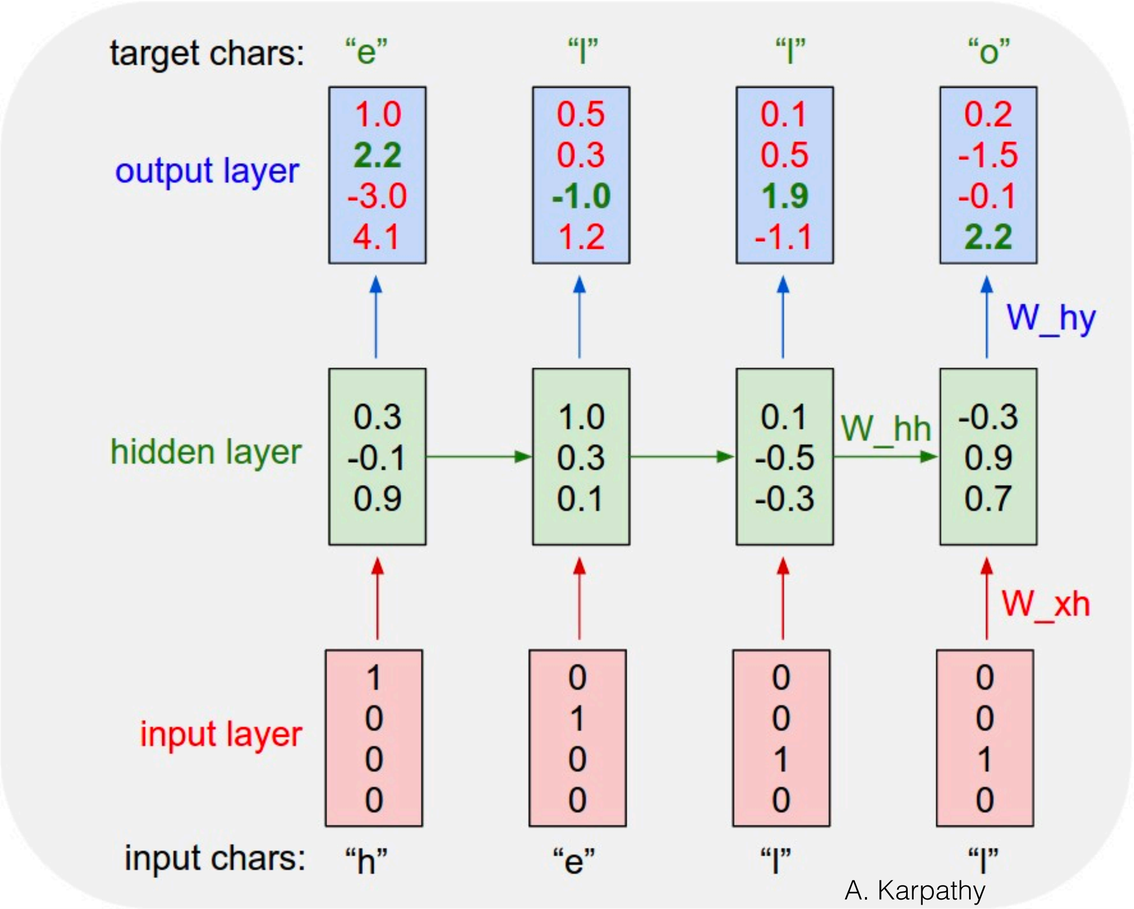
How do RNNs work?

A model with memory



- The input is a se e "**x**" (for example, over time). The output is also a sequence, **h** (formally **y = W_{hy}*h**)
- Hidden neural units "A" have an internal state h_t that gets updated based on the input "x_t", and on the **previous internal state.**
- Therefore, the content of the output h_t depends not only on the current input x_t, but on **the entire history of previous inputs**.

Training RNNs

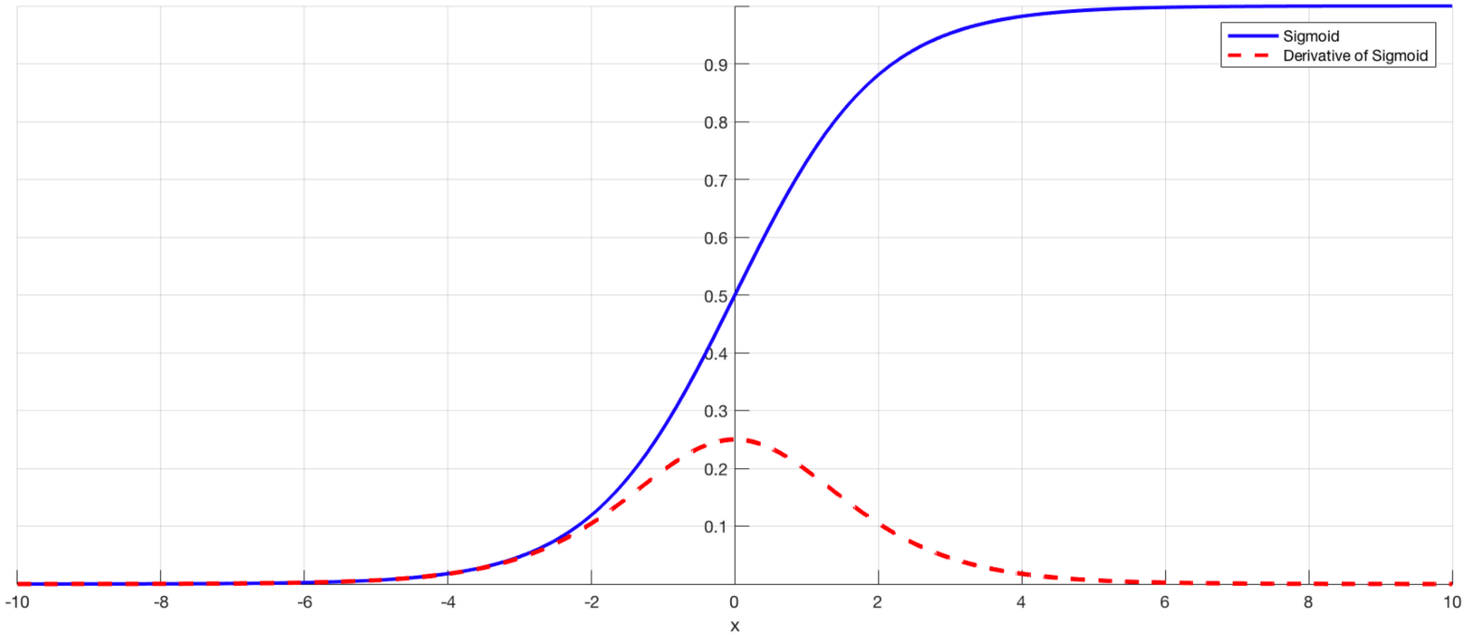


A. Karpathy

Muchos años después, frente al pelotón de fusilamiento, el coronel Aureliano Buendía había de recordar aquella tarde remota en que su padre lo llevó a conocer el hielo.

The three matrices (W_xh, W_hh, and W_hy) are initialized randomly. Then weights are adjusted (using gradient descent) to obtain desired output.

The vanishing gardens format

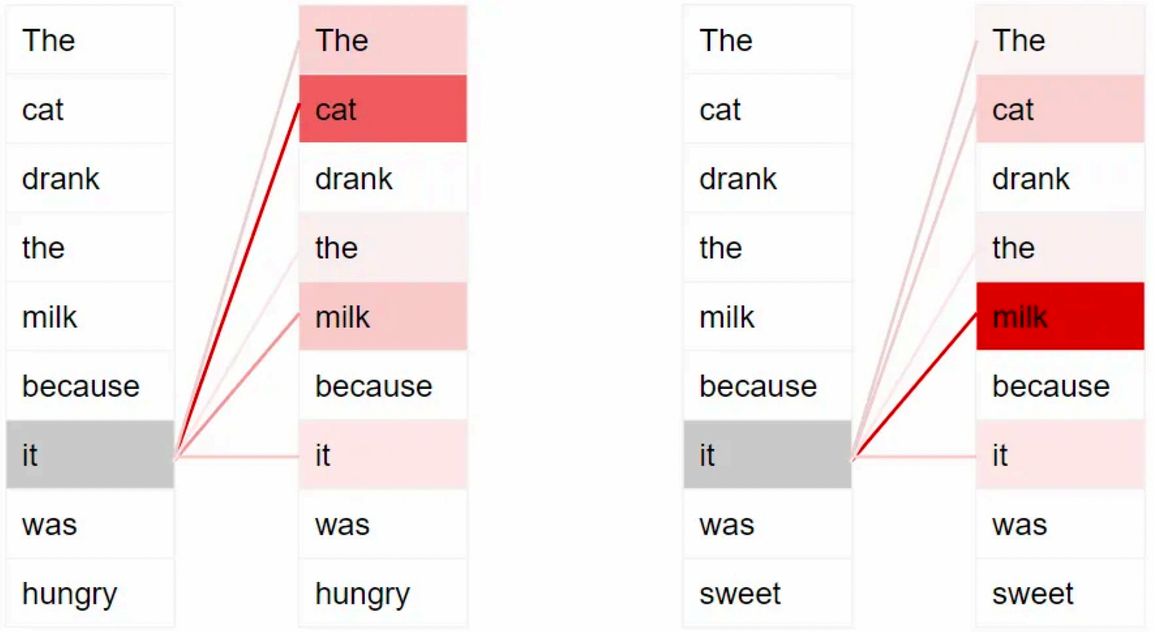


- At large values of the input, the derivative of the activation function is very small.
- Not a problem if it's only a few layers, but as you add laters, you multiply many small numbers together. This leads to disaster in training.
- Solutions: different activation functions, batch normalization, or different architectures

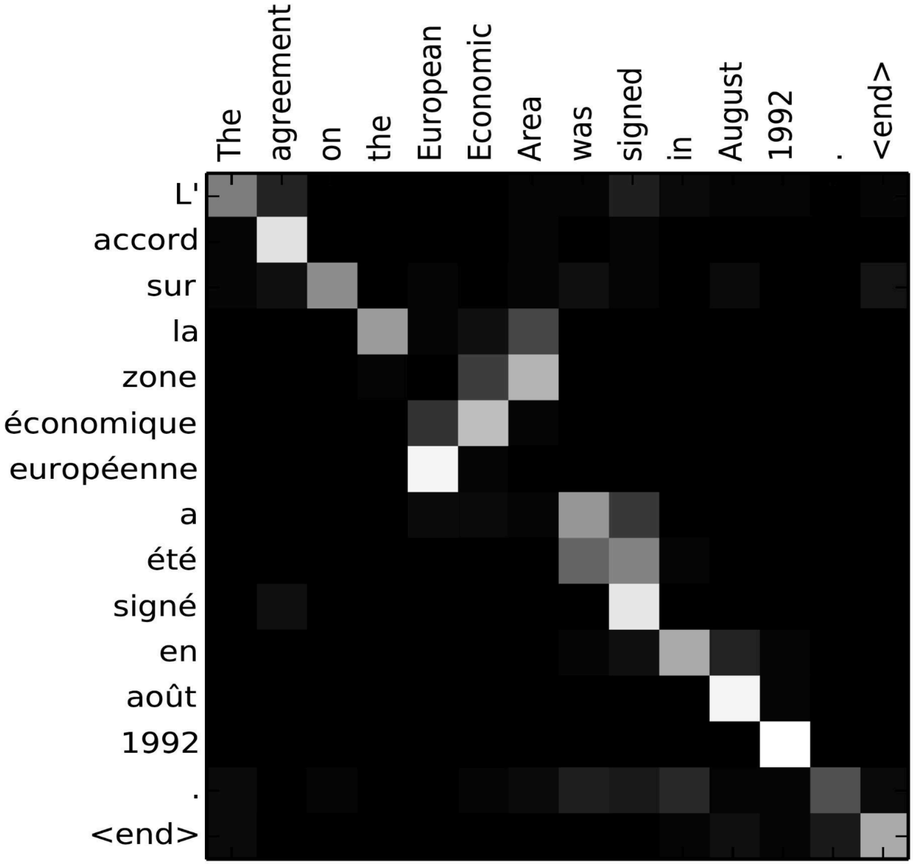
Attention Is All You Need

“The cat drank the milk because it was hungry”

“The cat drank the milk because it was sweet”

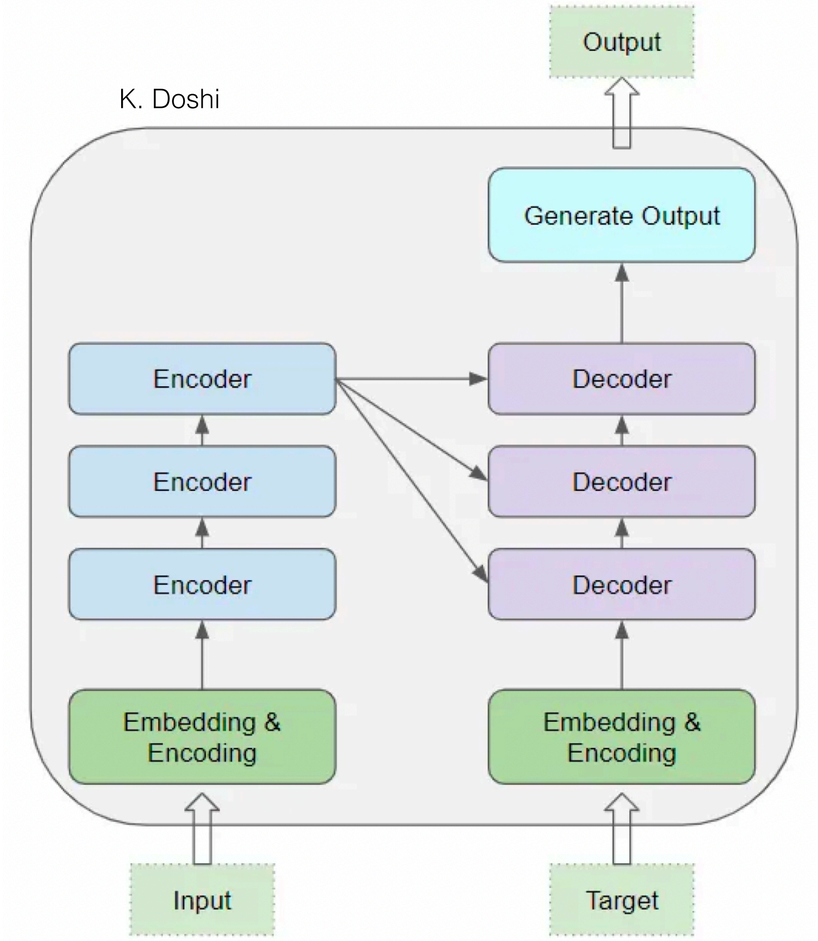


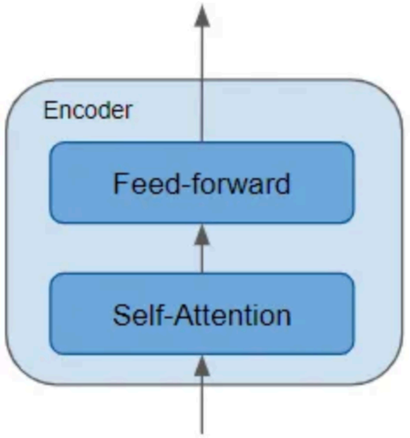
K. Doshi

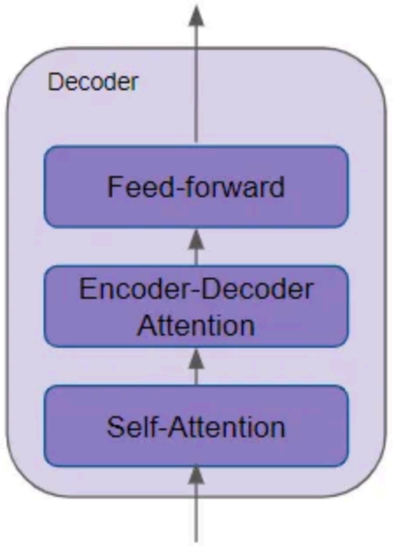


Bahdanau et al. 2015

General architecture - Attention layers







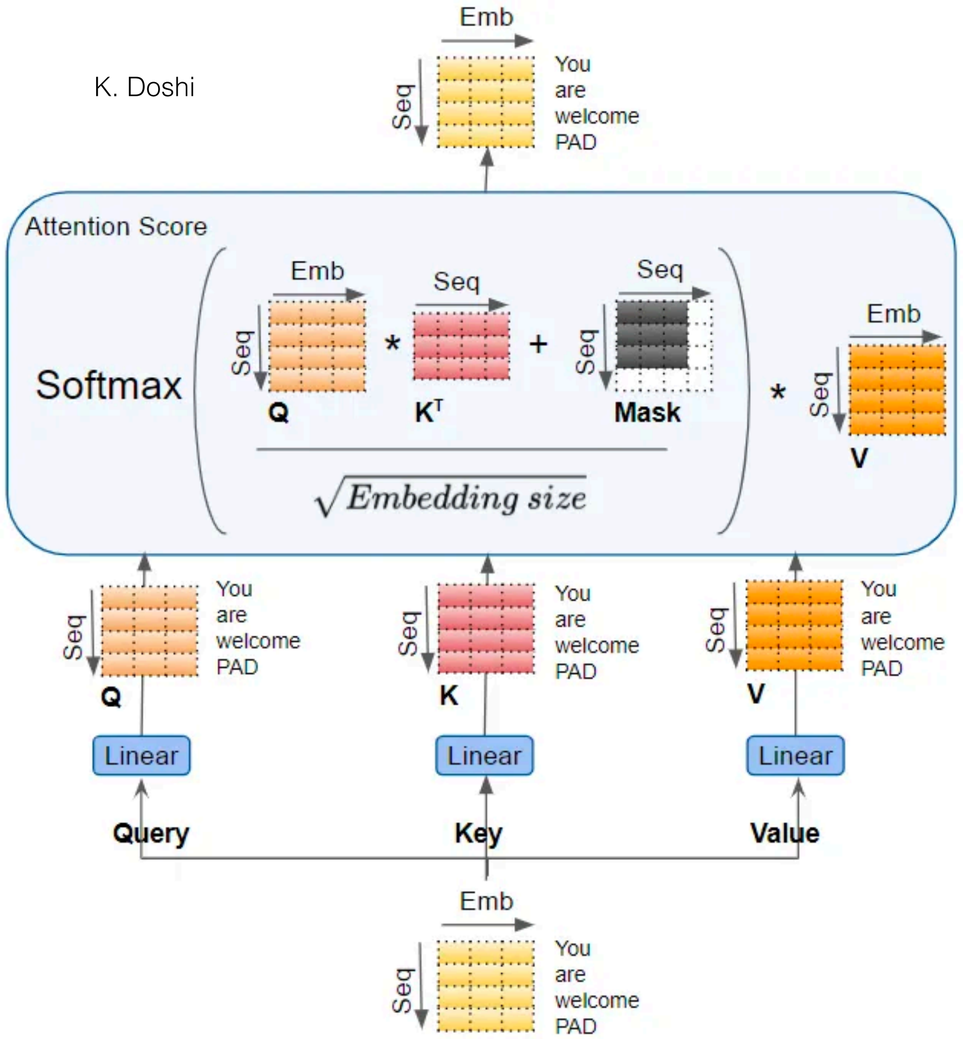
$$\mathbf{c}_t = \sum_{i=1}^{n} \alpha_{t,i} \boldsymbol{h}_i$$

Context vector for output

$$\alpha_{t,i} = \mathbf{align}(y_t, x_i)$$

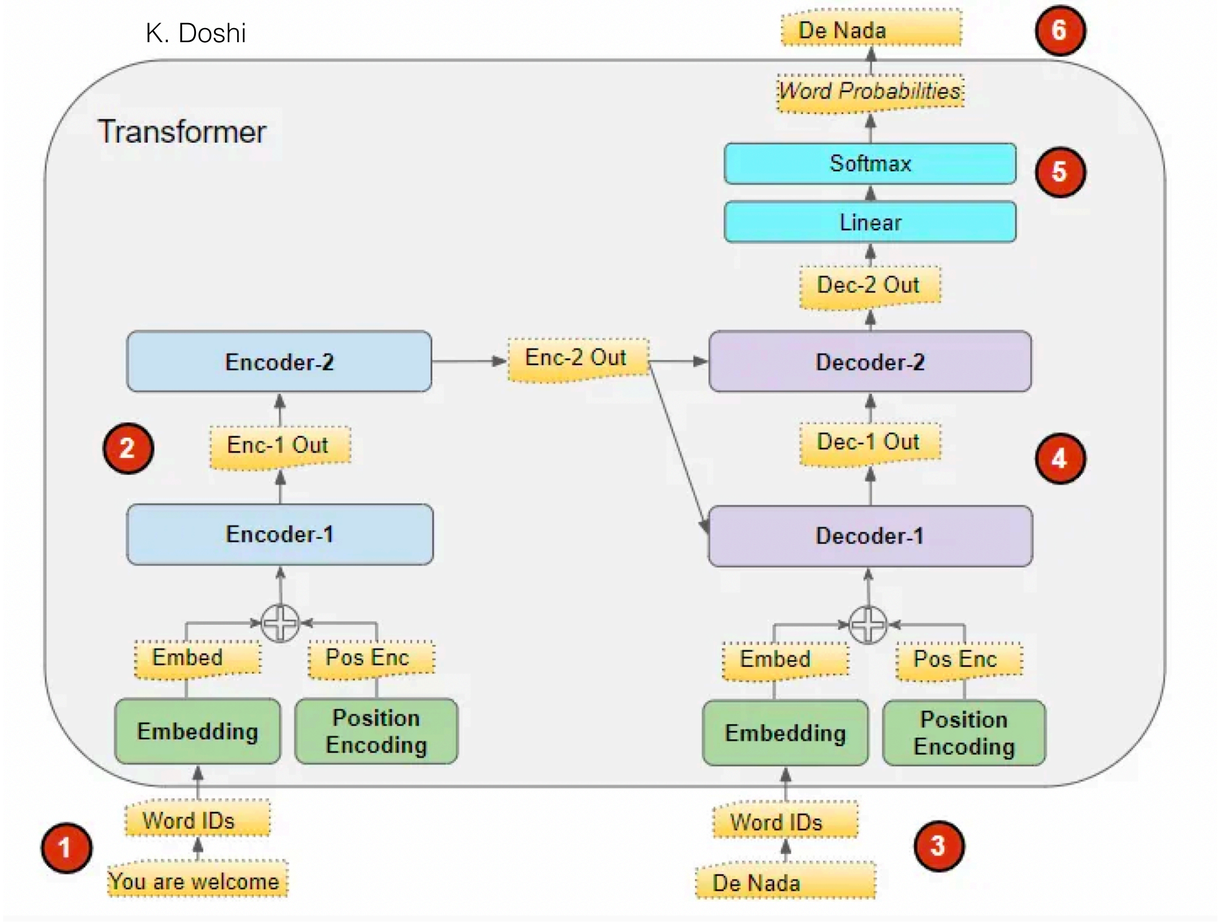
How well two words align

What happens inside an attention layer



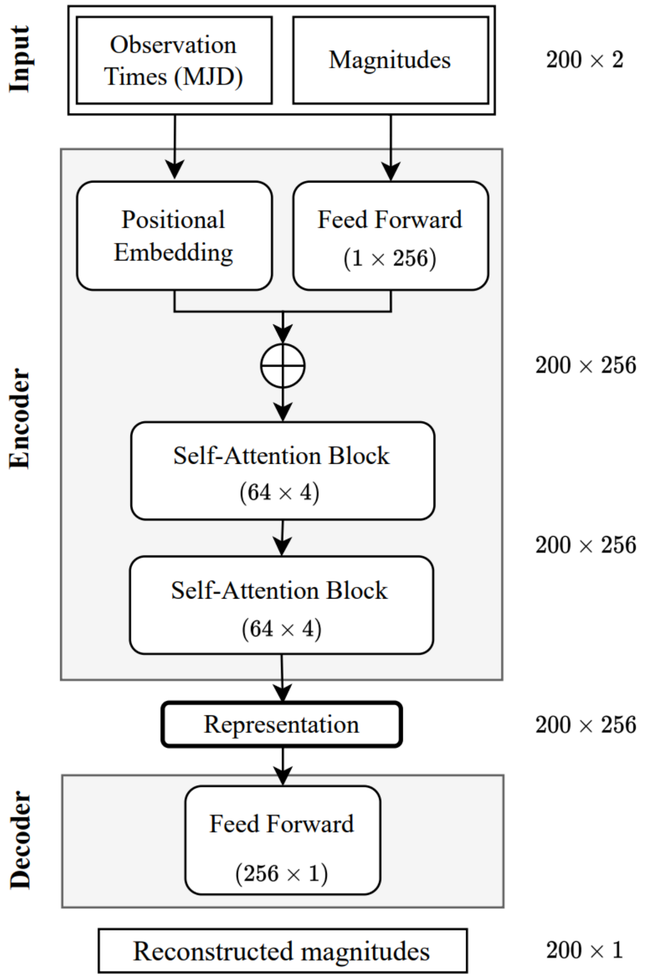
- Query (Q), Key (K) and Value (V) are matrices that carry an encoded representation of each word in the sequence.
- Q is the word for which we are calculating attention, whereas that K and V word is the one we are paying attention to.
- The attention score thus captures the interaction (alignment) between a word, and every other word in the sequence.
- We want the attention score to be high for two words that are relevant to each other in the sentence. And we want the score to be low for two words that are unrelated to one another —> Training

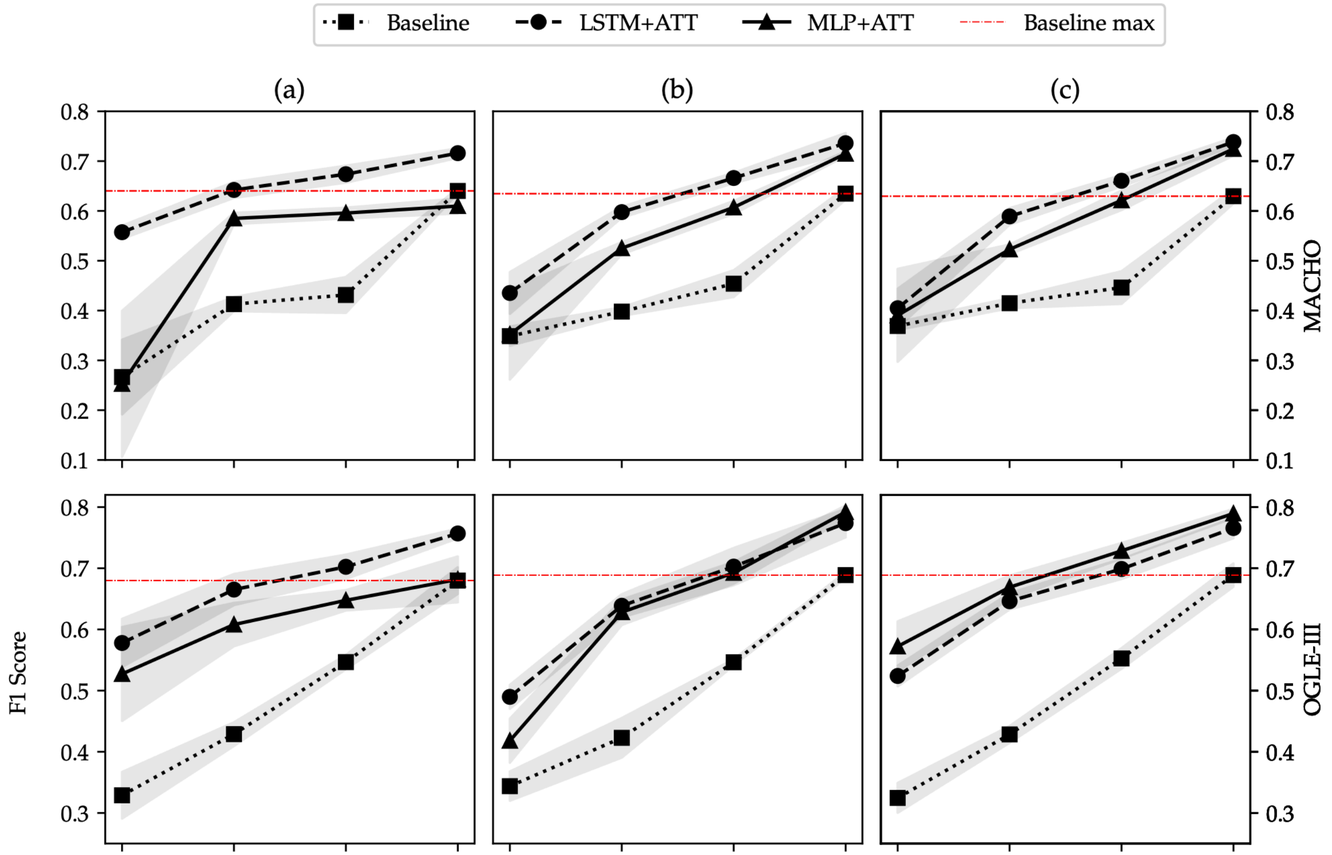
Training a Transformer



1. Embed the input sequence.
2. Stack of encoders produces an encoded representation of the sequence (with self-attention).
3. Target sequence is embedded and passed to the decoder.
4. Stack of decoders processes the target with self attention and encoder-decoded attention.
5. The output layer converts it to word probabilities and output sequence.
6. Compute error by comparing with target, and backpropagate.

Use in Astronomy: Light Curve Classification





Donoso-Oliva et al. 2022