Neural Machine Translation (NMT)

(a brief introduction)

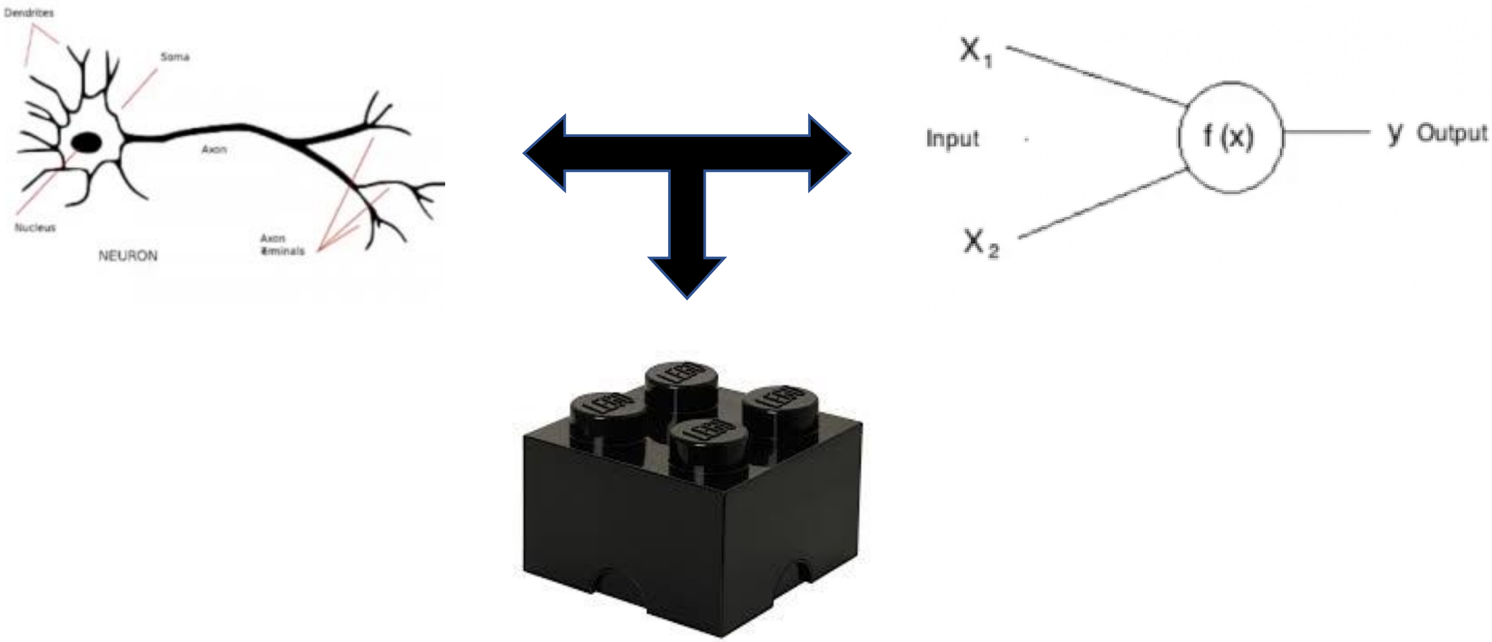
Dr Diptesh Kanojia

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

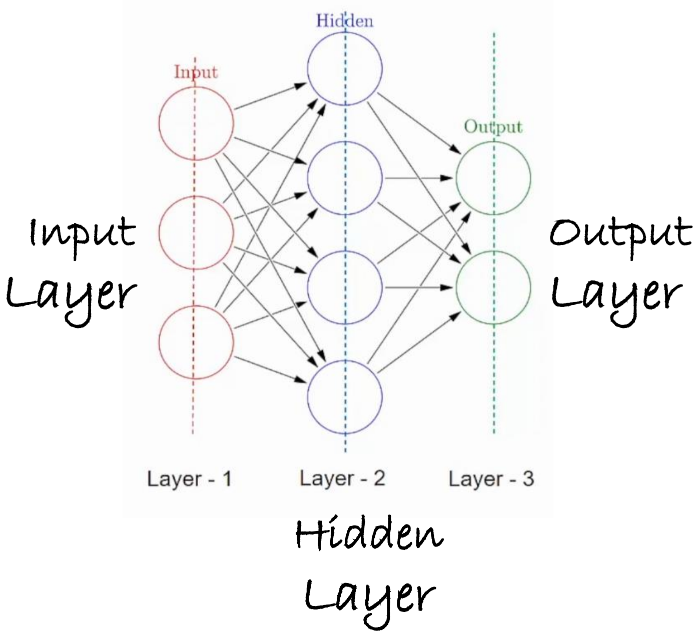
Artificial Neurons

- The idea behind an **artificial neuron** is akin to a biological neuron.
  - It can **receive input** from another neuron.
  - It can **process the input** signal.
  - It can **provide output** to another neuron.



Neural Networks

- Neural Networks consists of neurons stacked in **Layers**.
- Imagine the artificial neurons vertically stacked in the form shown here.
- A neural network built for a specific computational task can be thought of as a Lego toy (perhaps the Lego Death Star ☺)



Recurrent NN

CNN

Transformer

- **Input layer** accepts the data.
- **Hidden layers** process the data and does number crunching.
- **Output layer** shows you the desired output based on probability.

Vector Representations

Embeddings

- Vectors are **mathematical representations of words and phrases** used in NMT
- Vectors **capture the meaning and context** of words in a *continuous space*
- NMT models use vectors to represent the source and target language text
- During **training**, the **model learns to map** the source language vectors to the target language vectors
- During **inference**, the **model translates** a source sentence by finding the target language vectors closest to the source language vectors
- The final translation is generated by mapping the **target language vectors back to the target language words**.

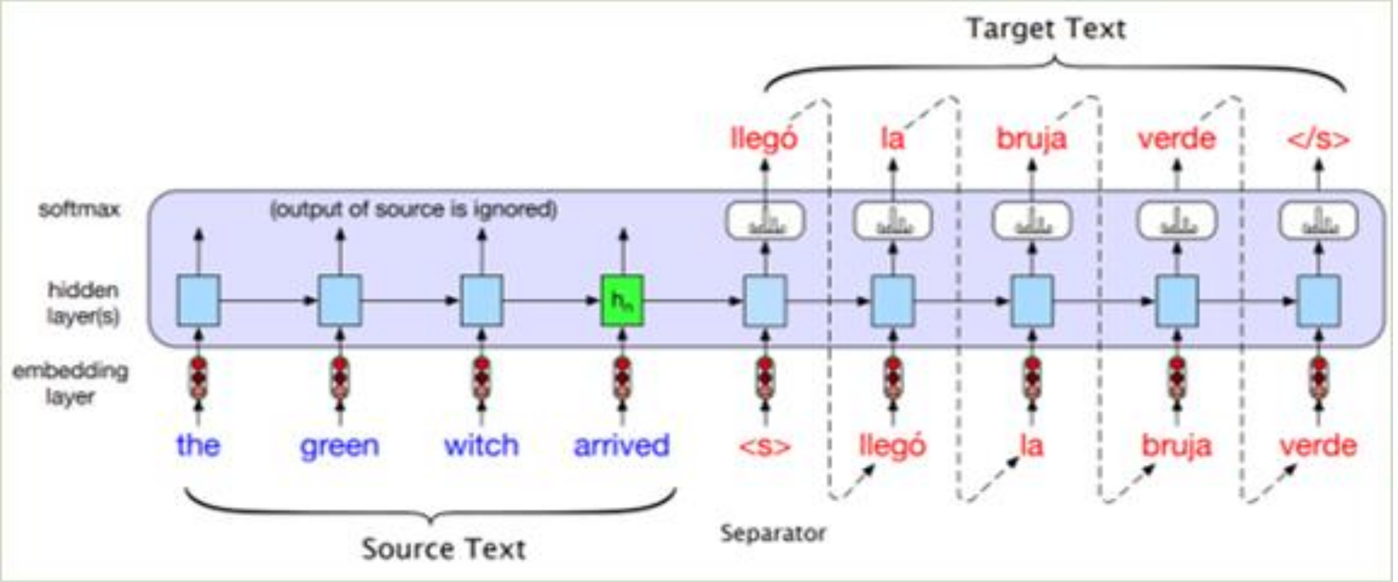
$$v_{\text{cat}} = \begin{pmatrix} -0.224 \\ 0.130 \\ -0.290 \\ 0.276 \end{pmatrix} \qquad v_{\text{dog}} = \begin{pmatrix} -0.124 \\ 0.430 \\ -0.200 \\ 0.329 \end{pmatrix}$$

$$v_{\text{the}} = \begin{pmatrix} 0.234 \\ 0.266 \\ 0.239 \\ -0.199 \end{pmatrix} \qquad v_{\text{language}} = \begin{pmatrix} 0.290 \\ -0.441 \\ 0.762 \\ 0.982 \end{pmatrix}$$

Encoders & Decoders

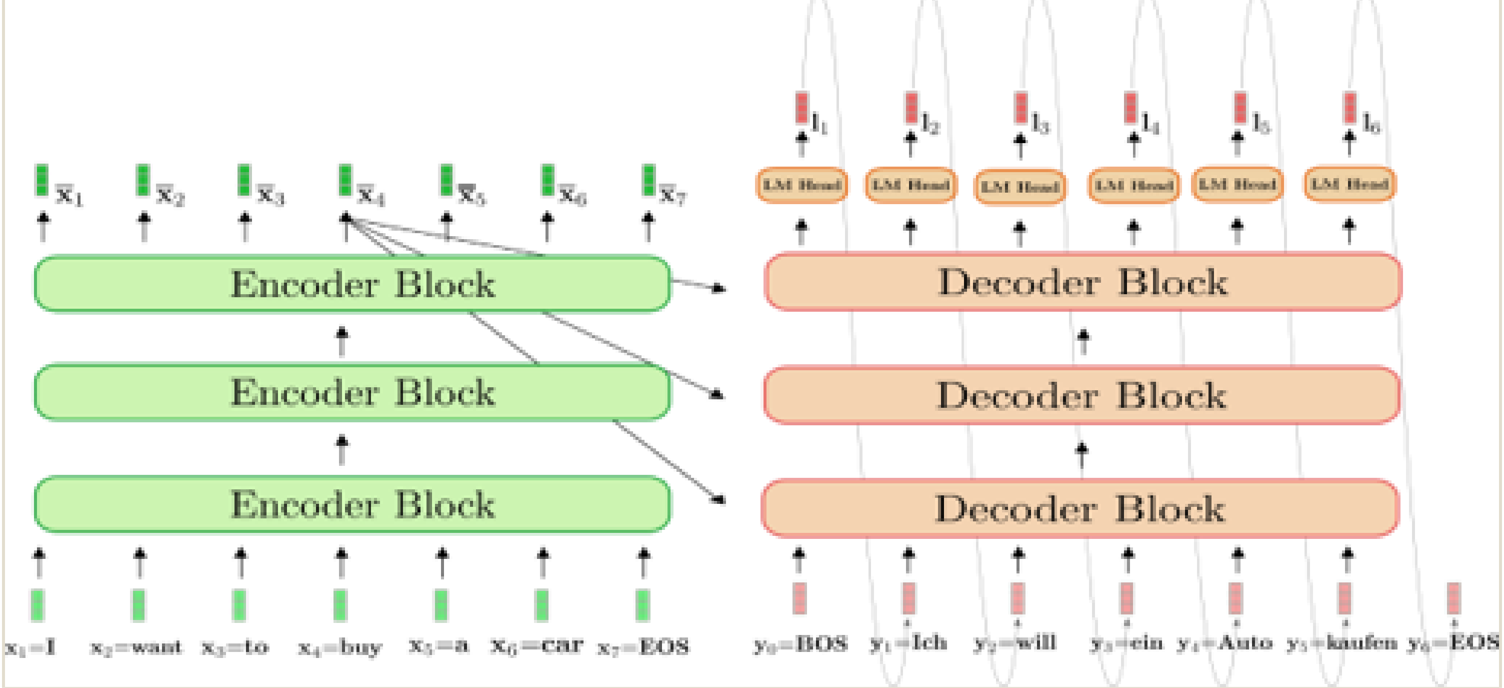
- Transform input source data to the target mathematical representation or vector (*encode*).
- Find an approximated sequence of words (*decode*), based on this target language representation.
- Each set of numbers is obtained from the hidden layers and **a single context vector** is formed.

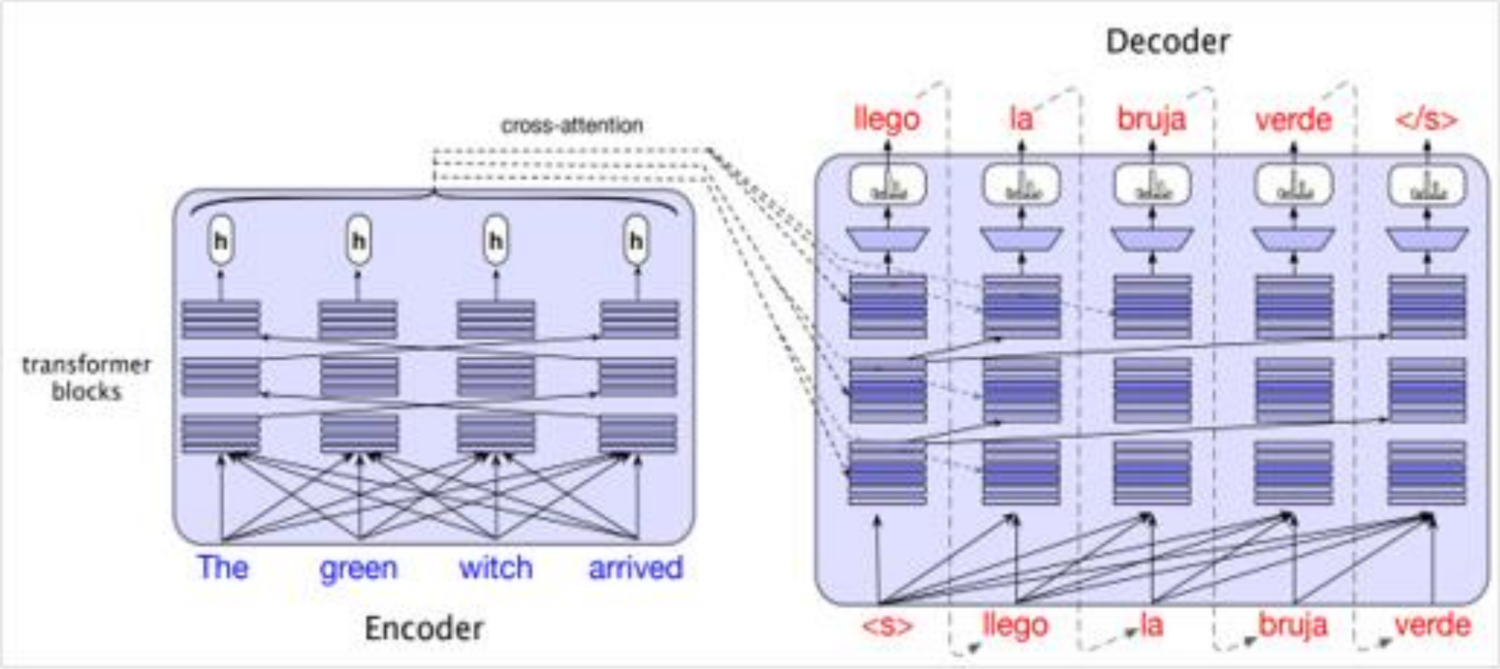
- This **context vector SHOULD inform each hidden layer** at the decoding stage.



An expanded view of when this architecture is applied to the task of NMT.

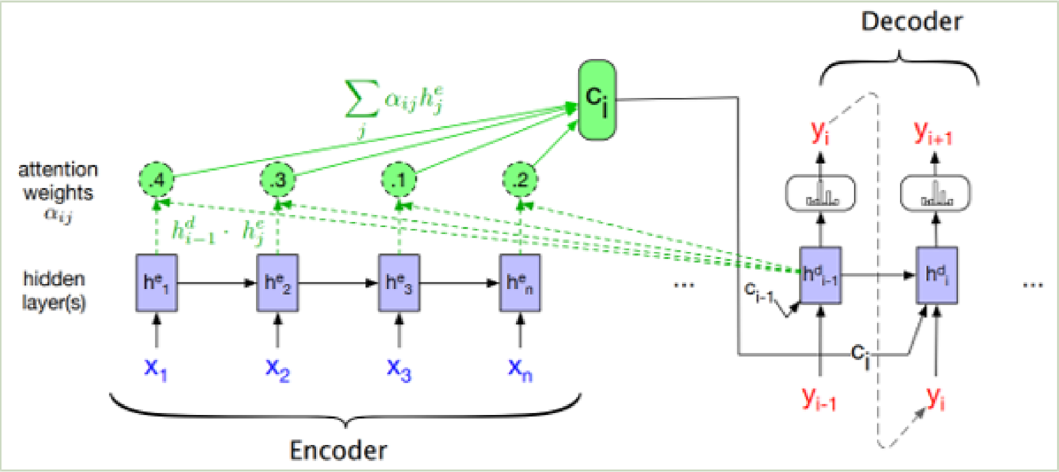
NMT Architectures





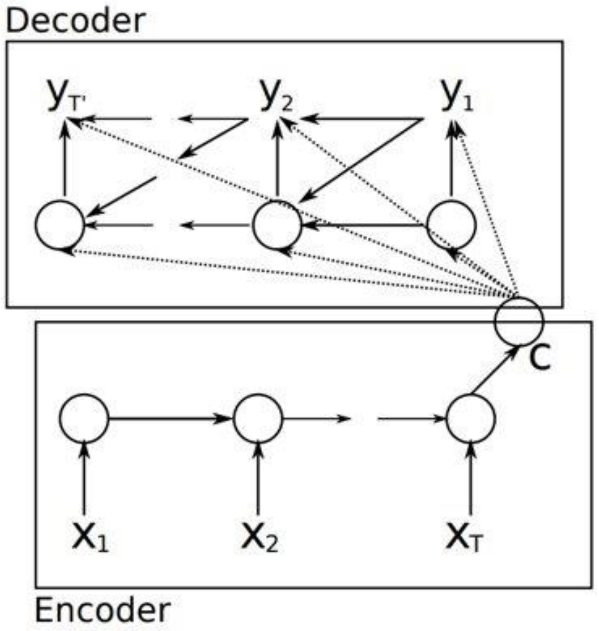
Attention Mechanism

- Allows the decoder to **focus on different parts of input sequence** at different points during the translation.
  - Helps generate a more accurate translation.
- NMT before attention was plagued with errors, especially for **long sentences**.
- Computes ‘attention weights’ for each word in the input text.
  - Weights the contribution of each word
- **Cross-attention** focuses on helping the alignment/mapping of source words to target words.



NMT Pipeline

- In terms of processing stages, an NMT pipeline consists of:
  - Data Pre-processing
  - Training
  - Model Output Evaluation
  - Model Deployment



Thank you!

QUESTIONS/DISCUSSION AT THE END OF THE NEXT PRESENTATION.

DR LEONARDO WILL NOW DISCUSS MORE PRACTICAL ASPECTS OF NEURAL MACHINE TRANSLATION

CENTRE FOR TRANSLATION STUDIES

UNIVERSITY OF SURREY