MLP to GPT: A Deep Learning Retrospective

A BRIEF HISTORY OF BREAK-THROUGHS IN NLP

June 12, 2024

Table of contents

Pre-Transformer Developments

Multi-layered Perceptron (1960s)

A **Multi-Layer Perceptron (MLP)** is a name for a modern feedforward artificial neural network *consisting of fully connected neurons with a nonlinear activation function, organized in at least three layers*. Modern feedforward networks are trained using the **backpropagation method** and are colloquially referred to as the **"vanilla"** neural networks.

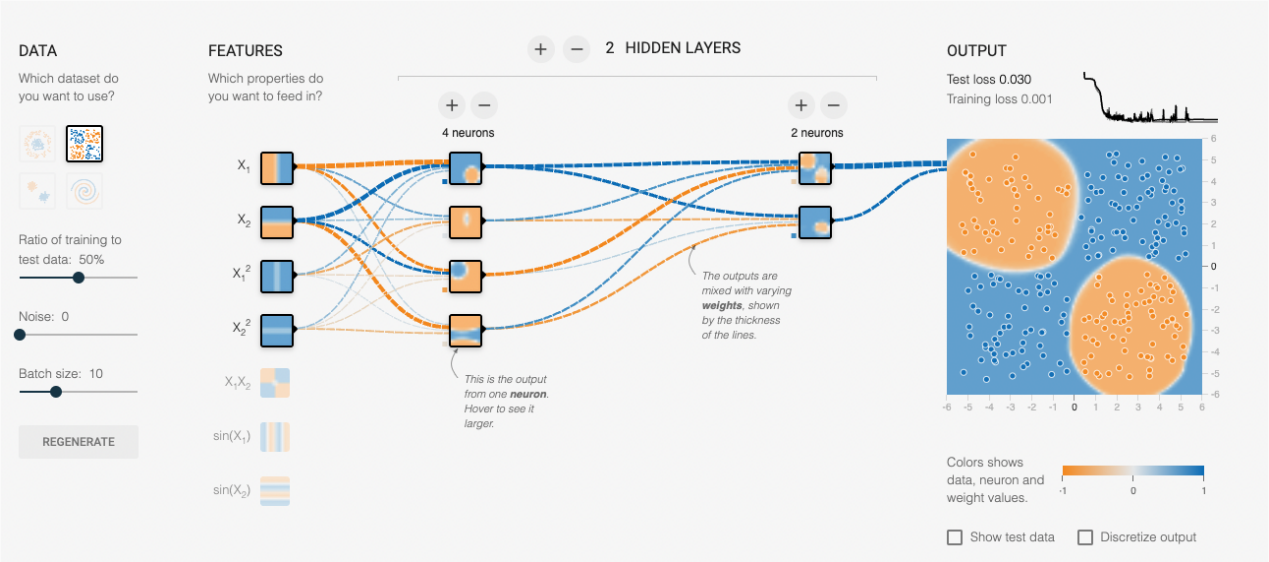
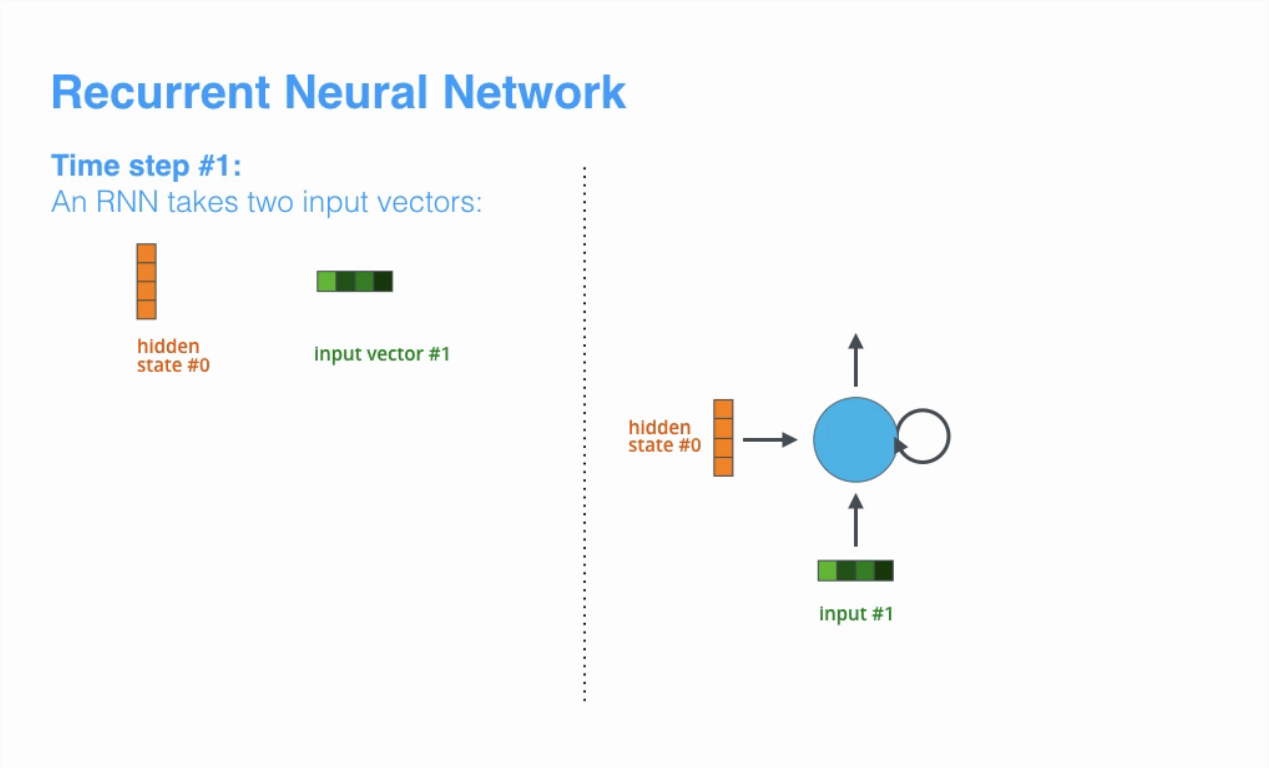


Figure 1: Tensorflow Playground

Recurrent Neural Networks (1980s)

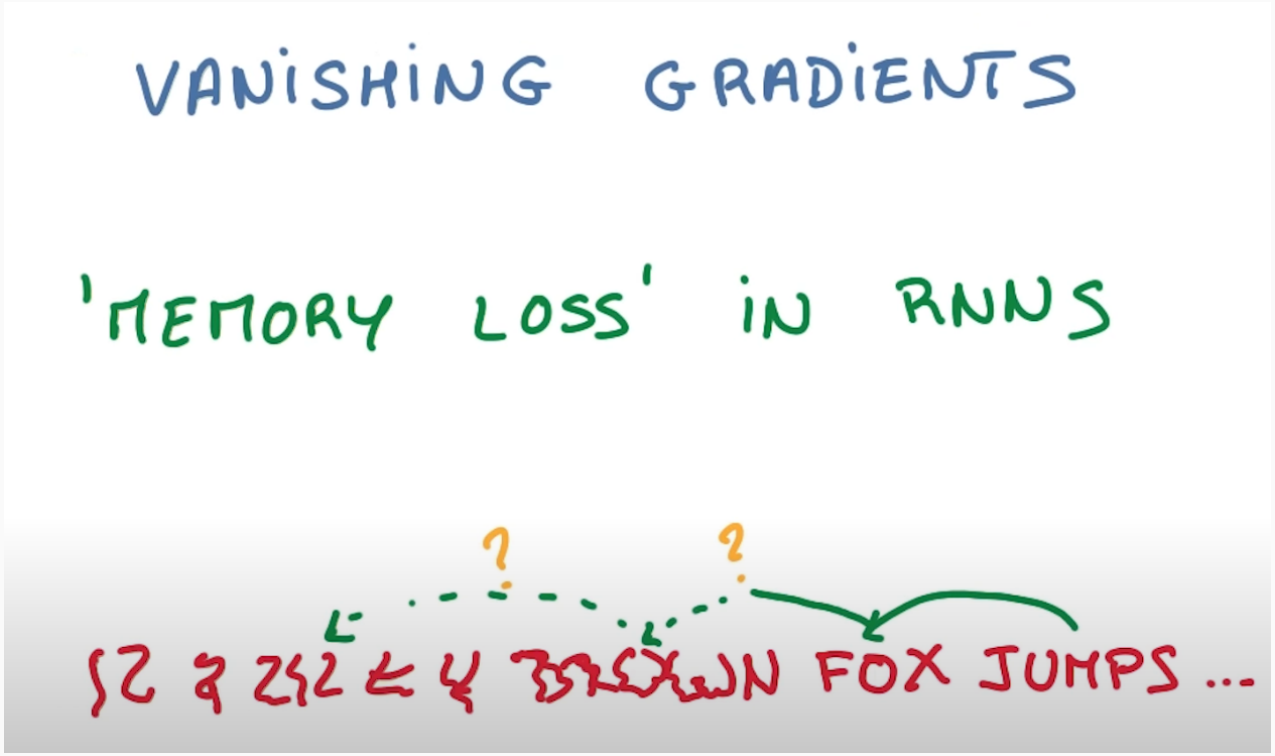
Recurrent Neural Networks (RNNs) are a class of neural networks designed for processing sequential data. The key differentiator with RNNs is the *hidden state* which capture information (context) about the sequence so far, enabling them to "remember" previous inputs.

$$h_t = \tanh(W_h * h_{t-1} + W_x * x_t + b) \tag{1}$$



Vanishing and Exploding Gradients

The vanishing and exploding gradient problems are common issues encountered during the training of recurrent neural networks (RNNs) due to large number of steps/layers in the backpropagation process wherein the gradients either decrease or increase exponentially in which case they become unusable (vanishing gradients carry too little information; exploding gradients are too unstable).



Long Short-Term Memory (LSTM)

It's difficult for vanilla RNNs to learn to preserve information over many timesteps. How about an RNN with memory?

Long Short-Term Memory (LSTM) is a type of RNN architecture designed to address the traditional RNN's difficulty in learning long-term dependencies due to the vanishing gradient problem.

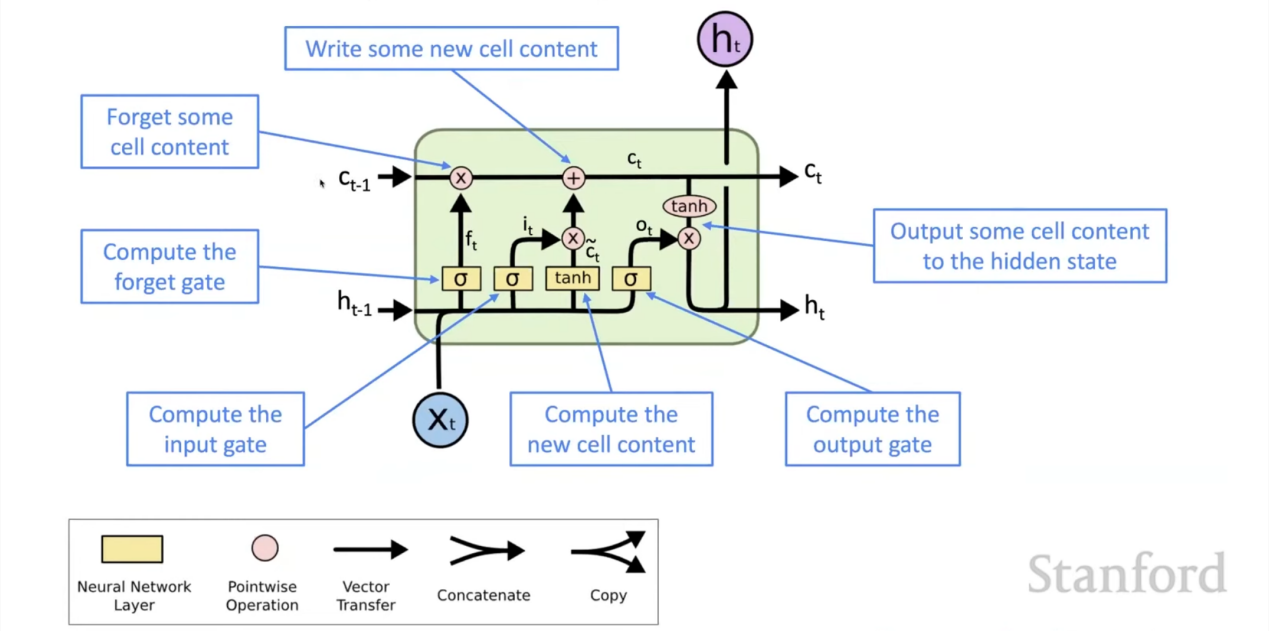
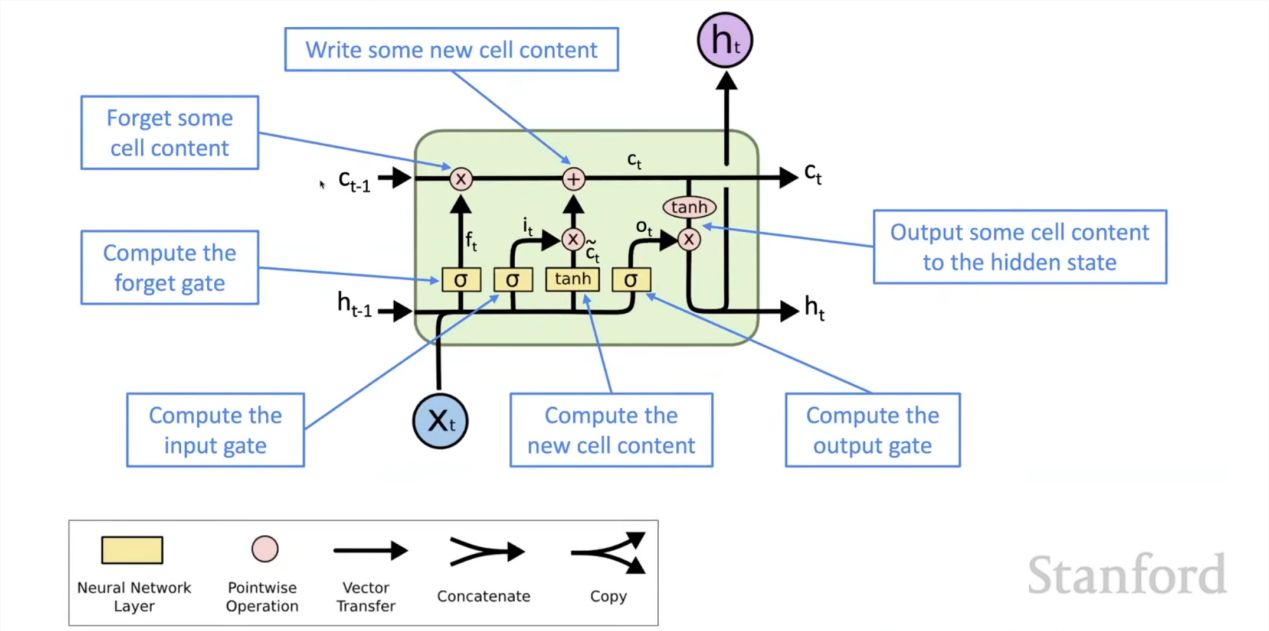


Figure 3: CS224N : NLP with Deep Learning

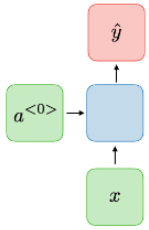
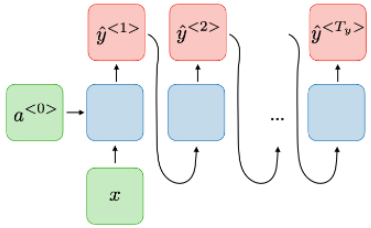
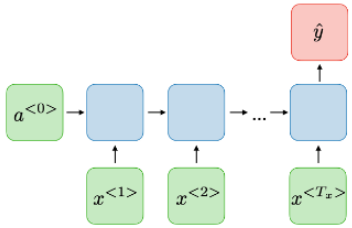
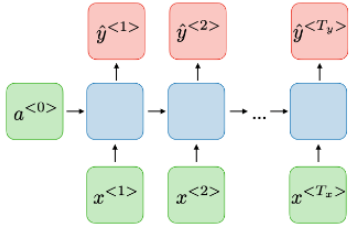
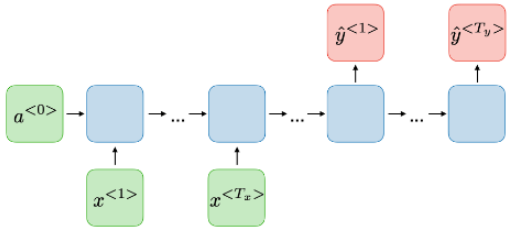
Long Short-Term Memory (LSTM)

LSTM has following key components:

- Memory Cell: Maintains information over time.
- Forget Gate: Decides what information to throw away from cell-state.
- Input Gate: Decides which new information to add to the cell-state.
- Output Gate: Controls how much of the cell state is exposed to the output.



Recurrent Neural Networks

Type of RNN	Illustration	Example
One-to-one $T_x = T_y = 1$		Traditional neural network
One-to-many $T_x = 1, T_y > 1$		Music generation
Many-to-one $T_x > 1, T_y = 1$		Sentiment classification
Many-to-many $T_x = T_y$		Name entity recognition
Many-to-many $T_x \neq T_y$		Machine translation

Recurrent Neural Networks : Sequence-To-Sequence

RNNs are often used in applications where input and output are in sequence format e.g. machine translation, text prediction, speech recognition, stock price forecasting etc.

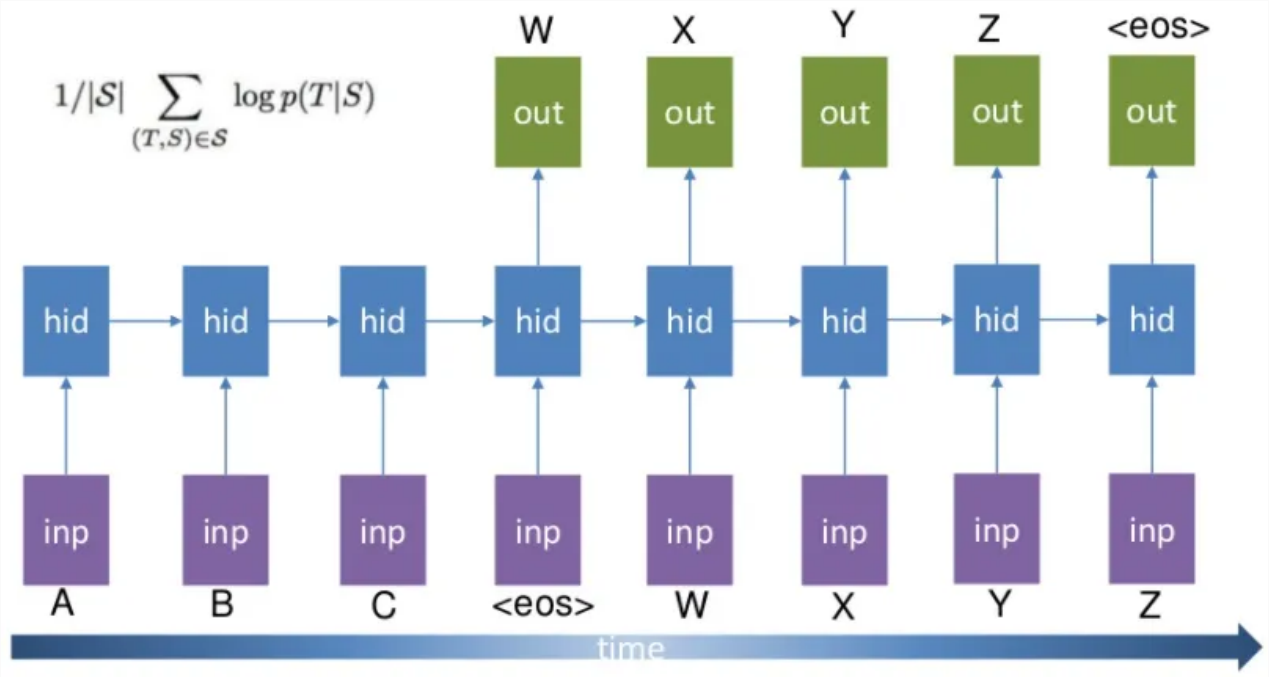


Figure 4: Sequence to Sequence Learning with Neural Networks by Ilya Sutskever, et al

Recurrent Neural Networks : Encoders & Decoders

The encoder processes the input sequence and distills it into a fixed-length representation, often referred to as the **context vector**. This vector serves as a condensed summary of the input, capturing its essence for the decoder to interpret.

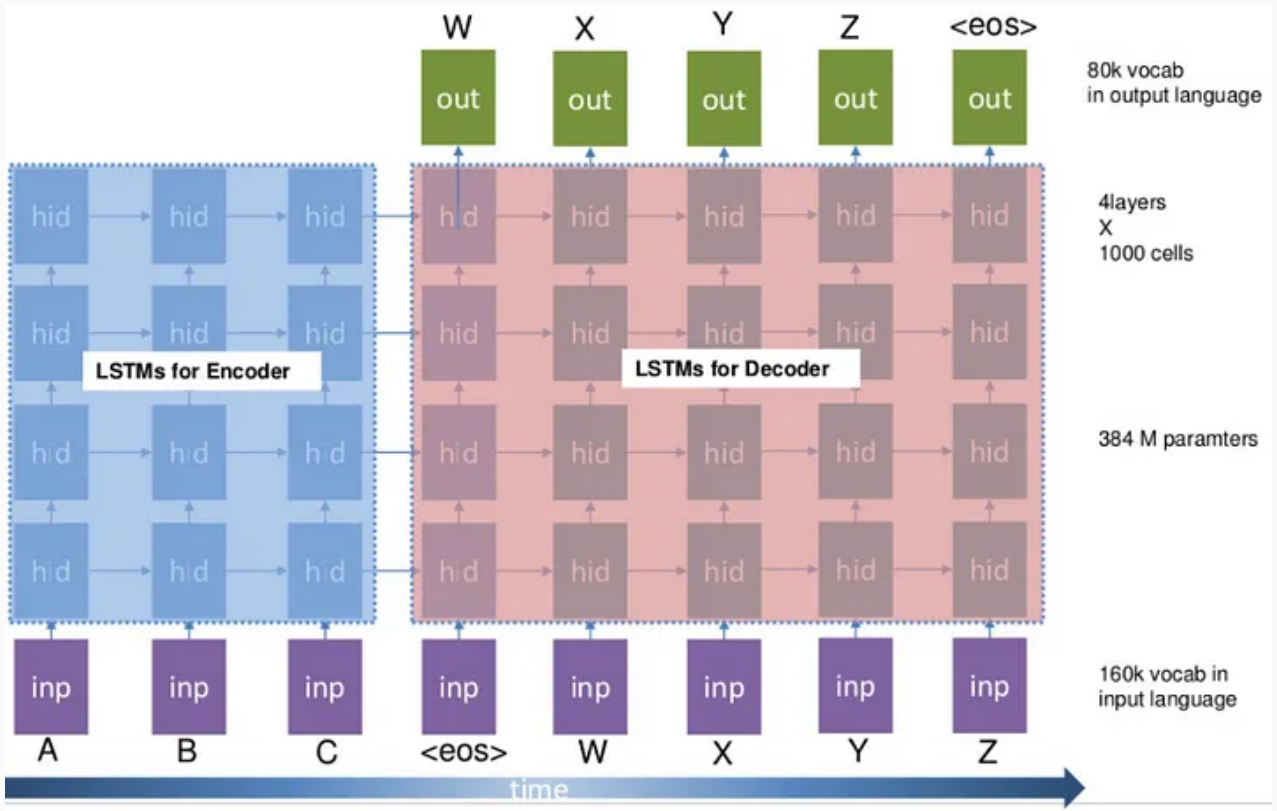


Figure 5: Seq2Seq Learning with Neural Networks by Ilya Sutskever, et al

Recurrent Neural Network : Encoders & Decoders

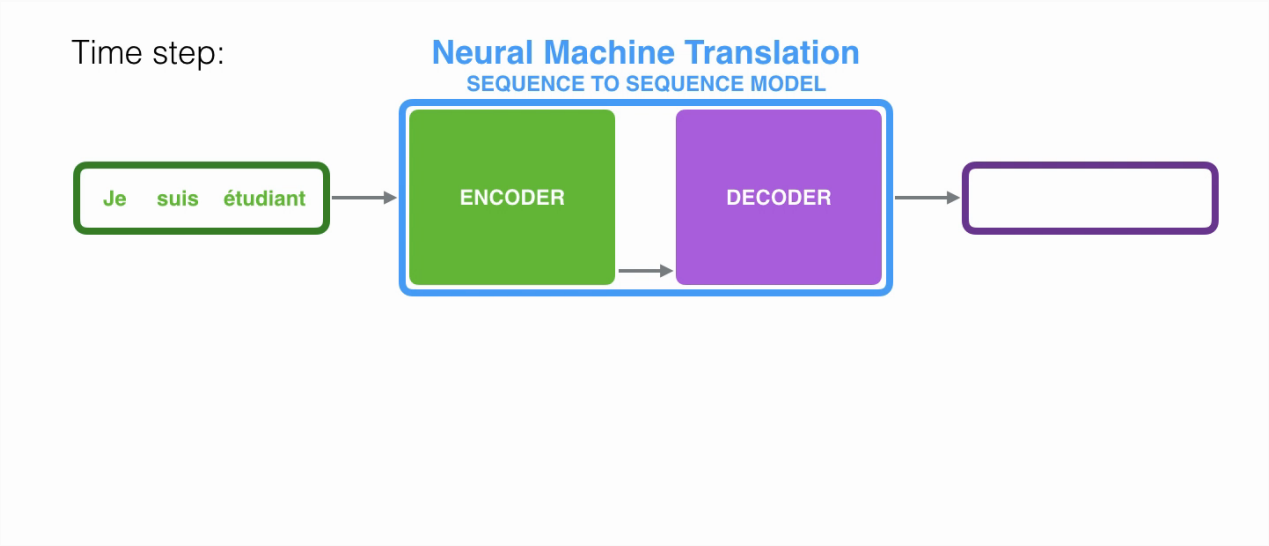


Figure 6: jalammar.github.io

Transformers

Attention

Enable the theme by loading

Self-Attention

Enable the theme by loading

multi-head attention

Enable the theme by loading

Transformers

Transformers and attention mechanisms have revolutionized the field of deep learning, offering a powerful way to process sequential data and capture long-range dependencies.

In this article, we will put our serious hat on and truly explore the basics of transformers and the importance of attention mechanisms in enhancing model performance and coherence.

Key Takeaways - Attention mechanisms are crucial in transformers, allowing different tokens to be weighted based on their importance, enhancing model context and output quality. - Transformers operate on self-attention, enabling the capture of long-range dependencies without sequential processing. - Multi-head attention in transformers enhances model performance by allowing the model to focus on different aspects of the input data simultaneously. - Transformers outperform RNNs and LSTMs in handling sequential data due to their parallel processing capabilities. - Applications of transformers span across NLP, computer vision, and state-of-the-art model development. Evolution of Transformers in Deep Learning

Transformer-Based Models

BERT

Enable the theme by loading

GPT

Enable the theme by loading

Llama

Enable the theme by loading

Mistral

Enable the theme by loading

Quickstart

starting a new nlp project

Starting Point for a new project -

- Task based split
- General Papers

starting a new DL project

Github Reference Guide

resources

PapersWithCode

MadeWithML

resources

Awesome NLP Papers

Brief Recent History of NLP Deep-Learning Models

- **1965:** *First feedforward network* was published by Alexey Grigorevich Ivakhnenko and Valentin Lapa
- **1967:** *Stochastic gradient descent* was used for the first time for training neural network
- **1982:** *BackPropogation* method was applied in the way that has become standard, for the first time by Paul Werbos.
- **1986:** David Rumelhart, Geoffrey Hinton, and Ronald Williams developed the idea of *backpropagation through time (BPTT)* to train RNNs.
- **1997:** Sepp Hochreiter and Jürgen Schmidhuber introduced *Long Short-Term Memory (LSTM)* networks.

Brief Recent History of NLP Deep-Learning Models

- **2013:** Mikolov et al. developed *Word2Vec* to produce word embeddings (dense vector representations of words in a continuous vector space).
- **2014:** Ilya Sutskever et al. introduced the *Sequence-to-Sequence (Seq2Seq)* model, a type of RNN architecture which used *encoder-decoder framework* for tasks such as machine translation.
- **2015-2016:** *Attention Mechanisms* were introduced to improve Seq2Seq models.
- **2017:** Vaswani et al. published the *"Attention Is All You Need"* paper, introducing the Transformer model.

References

Some references to showcase [allowframebreaks] [1]

Questions?

References i


J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu.
Harnessing the power of llms in practice: A survey on chatgpt and beyond.
2023.