RÉPUBLIQUE
FRANÇAISE

Liberté
Égalité
Fraternité

LaSTI

IGN
INSTITUT NATIONAL
DE L'INFORMATION
GÉOGRAPHIQUE
ET FORESTIÈRE

CHANGER
D'ÉCHELLE

Université
Gustave Eiffel

Institut national de l'information
géographique et forestière

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

LaSTI

IGN INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE
CHANGER D'ÉCHELLE

Université Gustave Eiffel

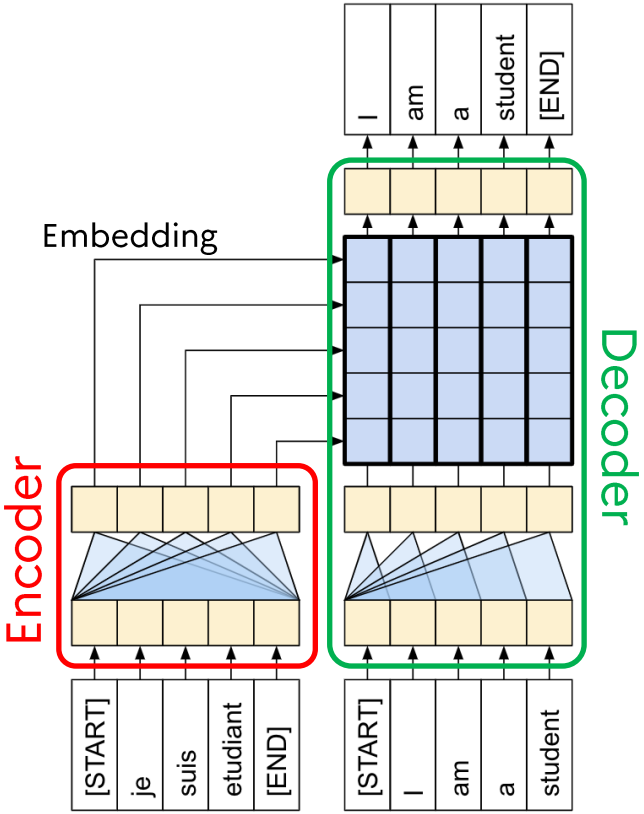
NATURAL LANGUAGE PROCESSING

Grégoire Grzeczkowicz

1. Why?


NLP Tasks
Text Preprocessing
Coreference Resolution
Part Of Speech Tagging
Word Sense Disambiguation
Grammatical Error Correction
Text-to-Data and viceversa
Speech-to-Text
Text-to-Image
Text-to-Speech
Data-to-Text
Fake News and Hate Speech Detection
Fake News Detection
Stance Detection
Hate Speech Detection
Text Reasoning
Common Sense Reasoning
Natural Language Inference
Classification
Text Classification
Feature Extraction
Sentiment Analysis
Information Retrieval and Document Ranking
Sentence/Document Similarity
Question Answering
Chatbots
Slot Filling
Intent Detection
Keyword Extraction
Topics and Keywords
Keyword Extraction
Topic Modeling
Knowledge Bases, Entities and Relations
Relation Extraction
Entity Linking
Relation Prediciton
Named Entity Recognition
Text-to-Text Generation
Machine Translation
Paraphrase Generation
Text Generation
Lexical Normalization
Text Summarization
Text Simplification

2. How?



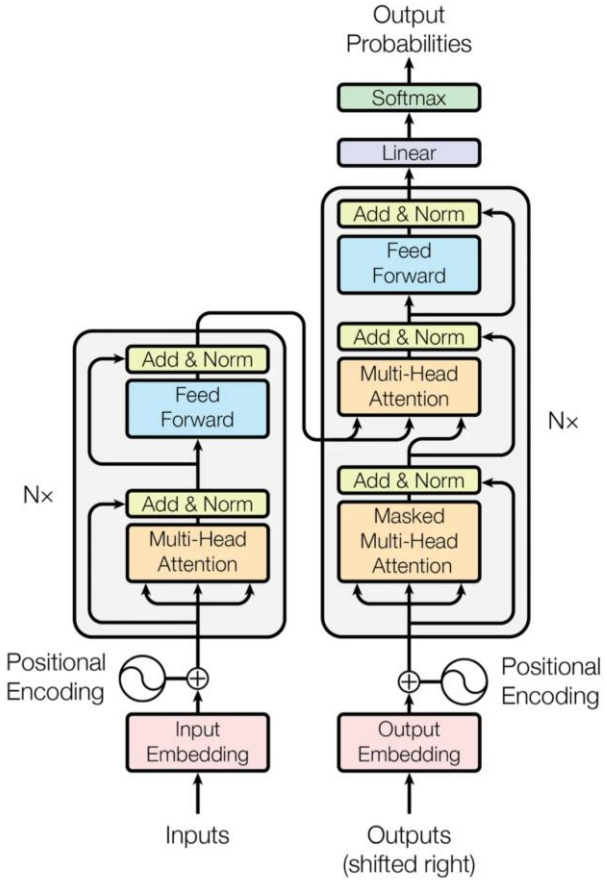


Figure 1: The Transformer - model architecture. [1]

[1] A. Vaswani *et al.*, 'Attention is All you Need', in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

Word embedding

One-hot encoding:

V words in the dictionary => each word is a vector of size V

The cat sat on the mat

The: [0 1 0 0 0 0 0]

cat: [0 0 1 0 0 0 0]

sat: [0 0 0 1 0 0 0]

on: [0 0 0 0 1 0 0]

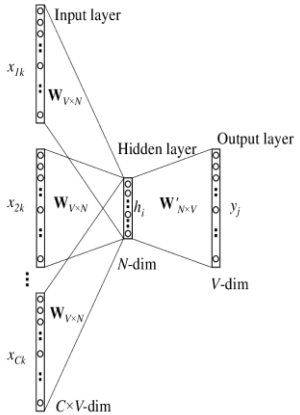
the: [0 0 0 0 0 1 0]

mat: [0 0 0 0 0 0 1]

Word2Vec [1]:

Unsupervised learning of a $V \times N$ matrix to encode each word of V size dictionary into a vector of size N

2 methods: CBOW and Skip-Ngram



[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, 'Efficient Estimation of Word Representations in Vector Space'. arXiv, 2013.

Byte-Pair Encoding (BPE) [2]:

Use a big corpus to split words in small meaningful tokens

The token set is built incrementally from the base characters and by adding among all the previous token pairs, the one most present in the corpus.

```
I am Grégoire Grzeczkowicz and

I am doing a presentation about NLP.
```

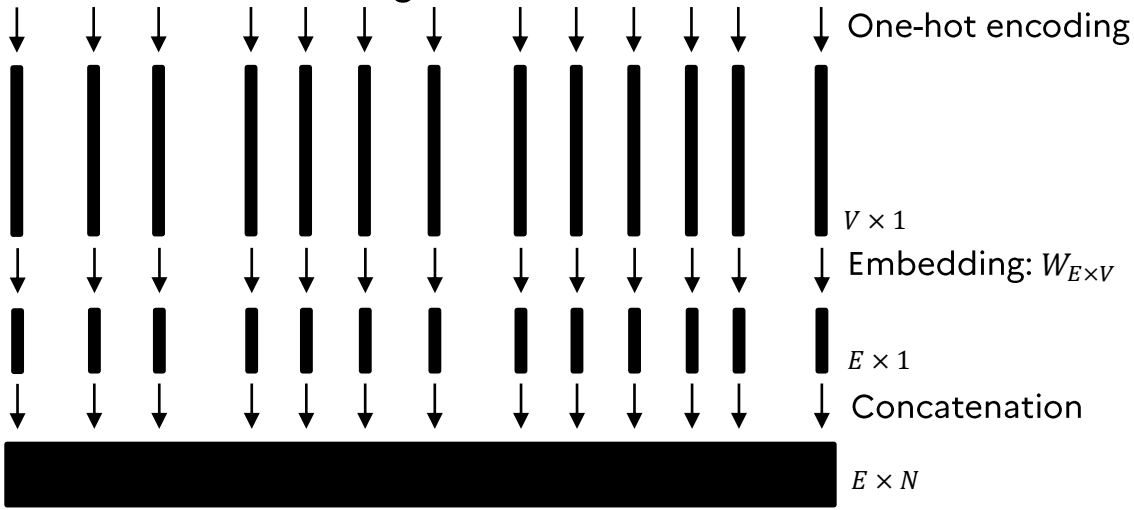
[2] R. Sennrich, B. Haddow, and A. Birch, 'Neural machine translation of rare words with subword units', *arXiv preprint arXiv:1508. 07909,* 2015.

Word embedding

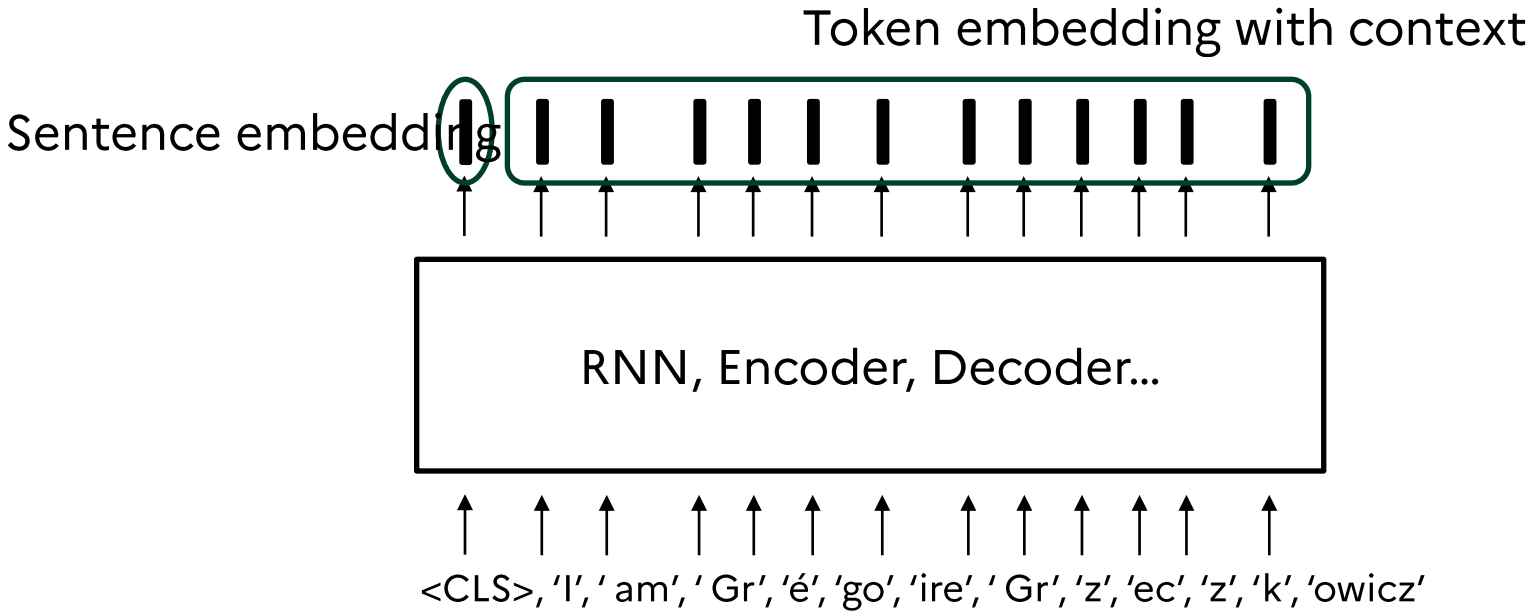
I am Grégoire Grzeczkowicz

↓ Tokenization: BPE

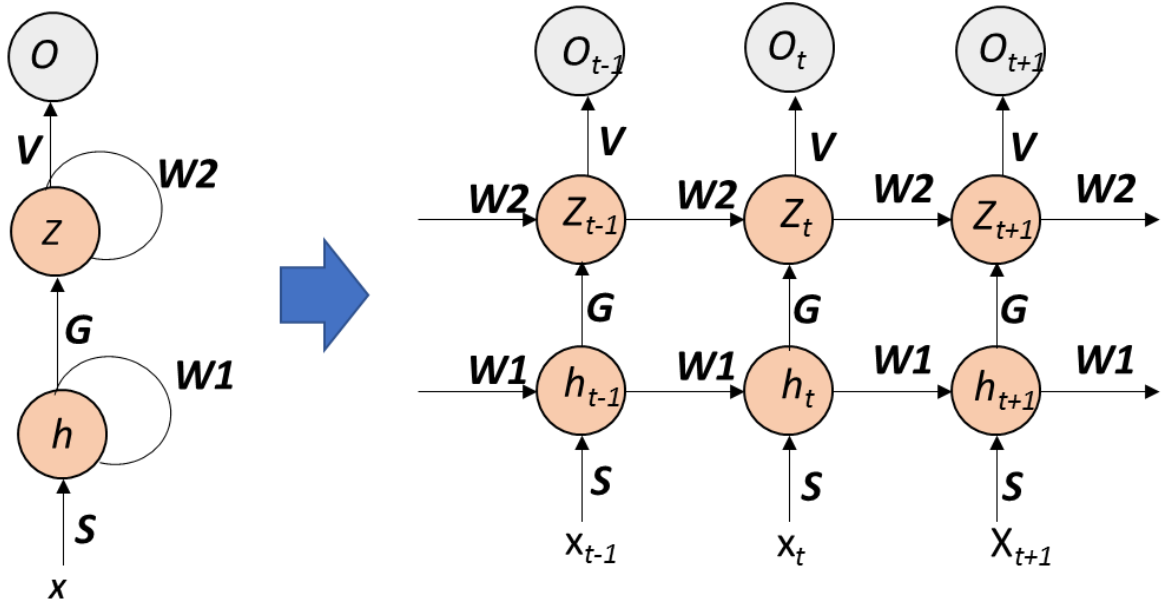
<CLS>, ‘I’, ‘ am’, ‘ Gr’, ‘é’, ‘go’, ‘ire’, ‘ Gr’, ‘z’, ‘ec’, ‘z’, ‘k’, ‘owicz’



Generic architecture



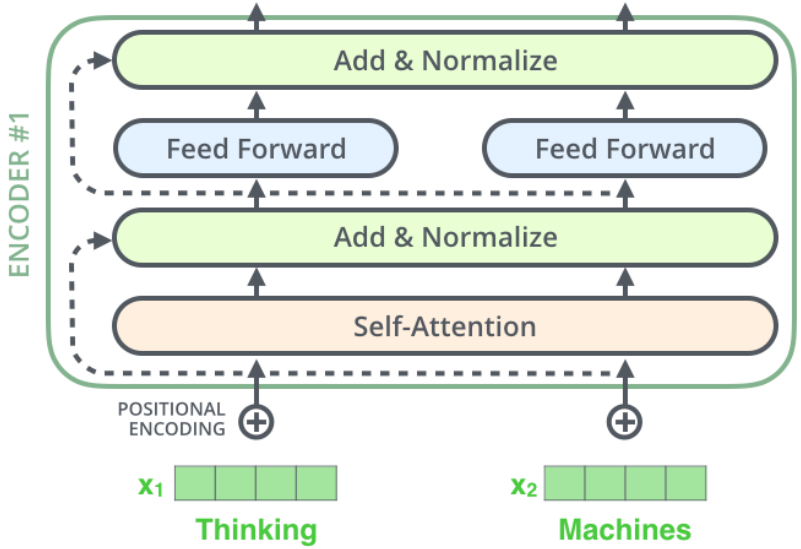
Deep recurrent neural network (RNN)

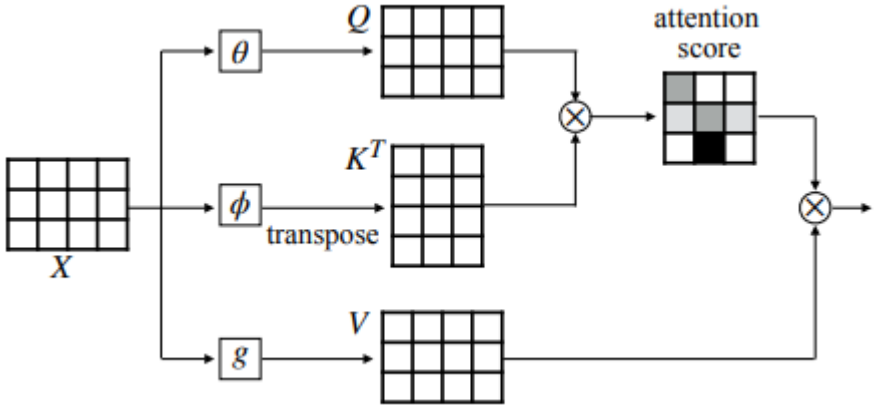


a) 2-layer Recurrent Neural Network (RNN)

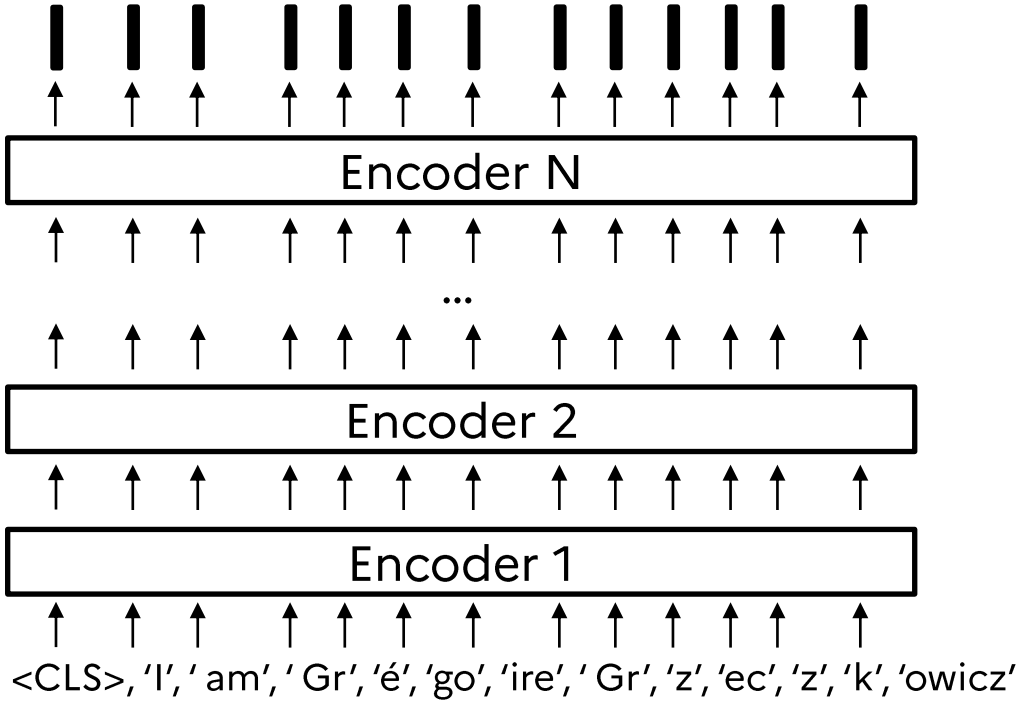
b) Unfolded 2-layer Recurrent Neural Network (RNN)

Attention model

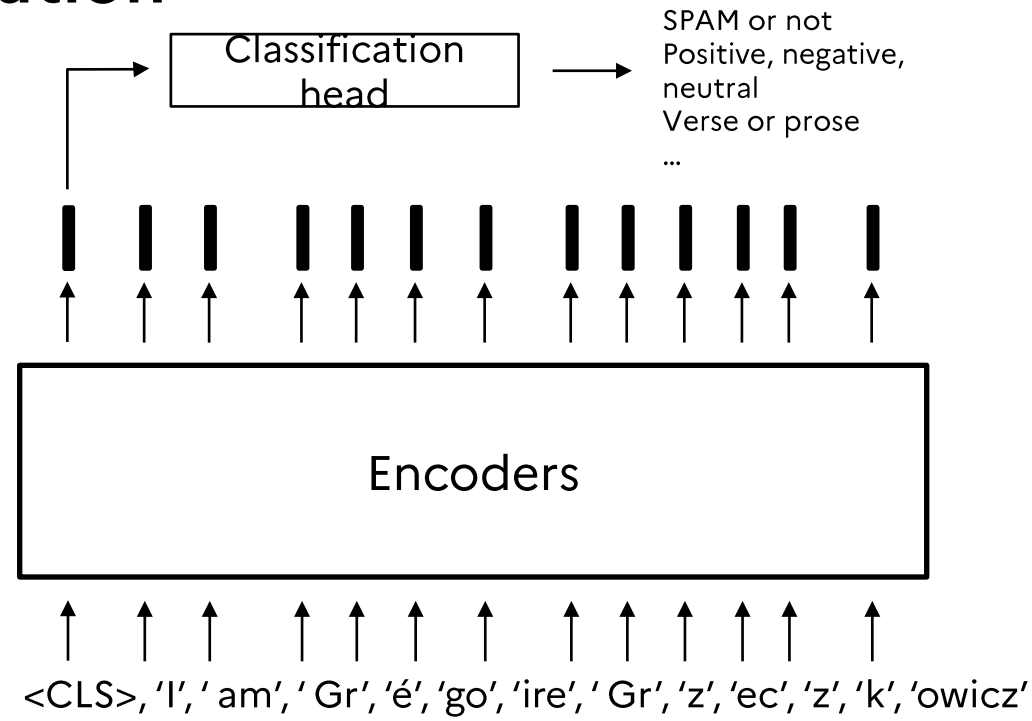




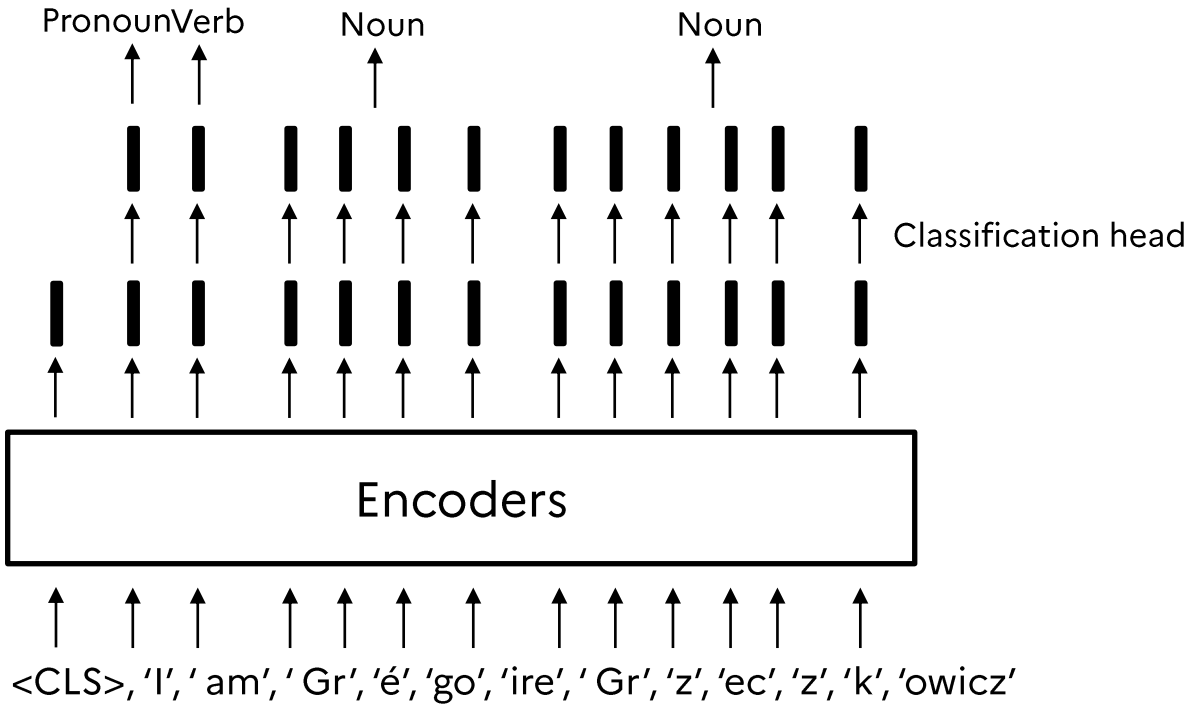
Attention model



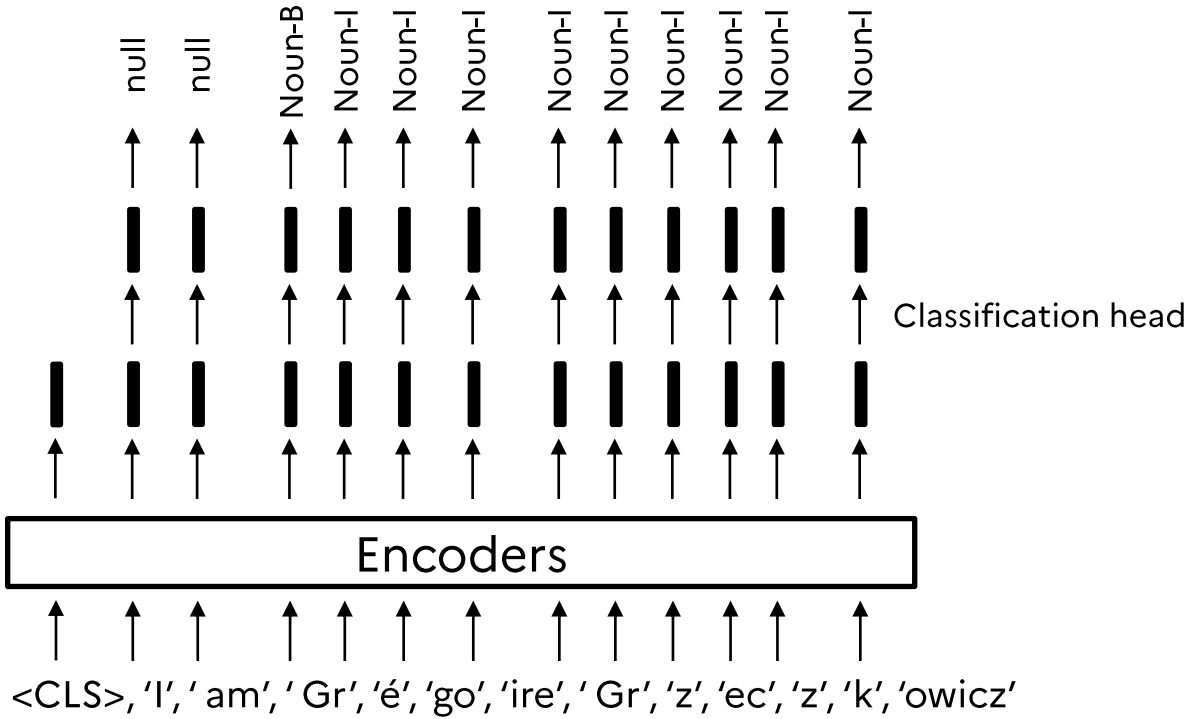
Text classification



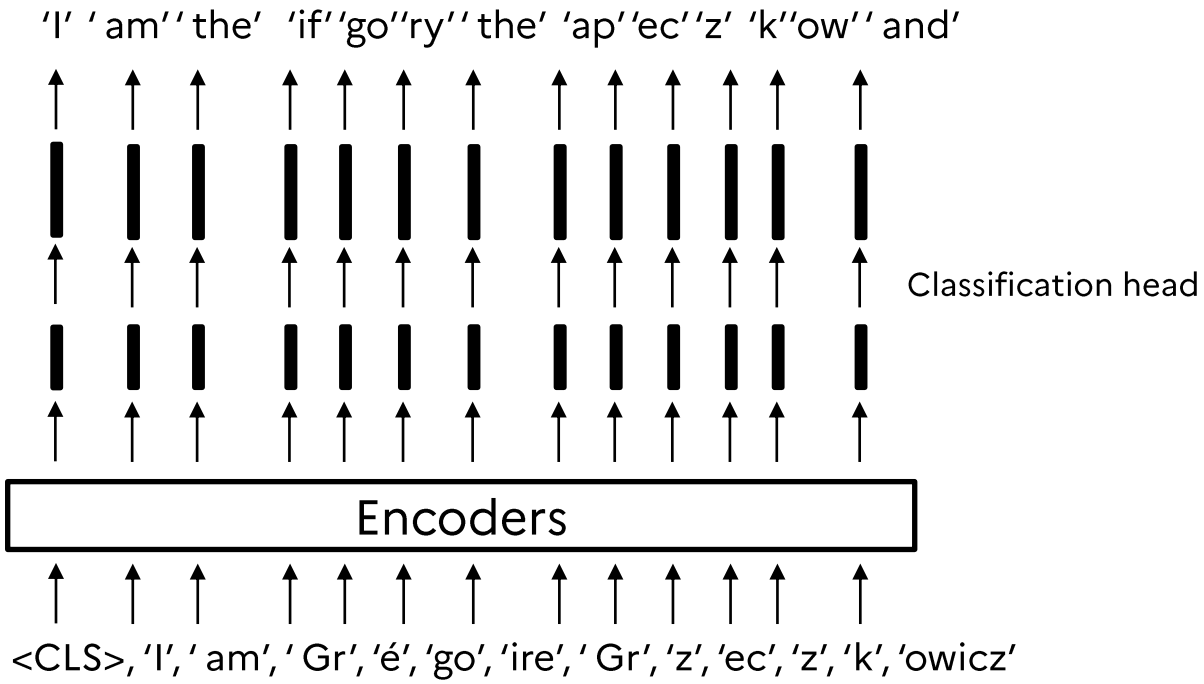
Token classification



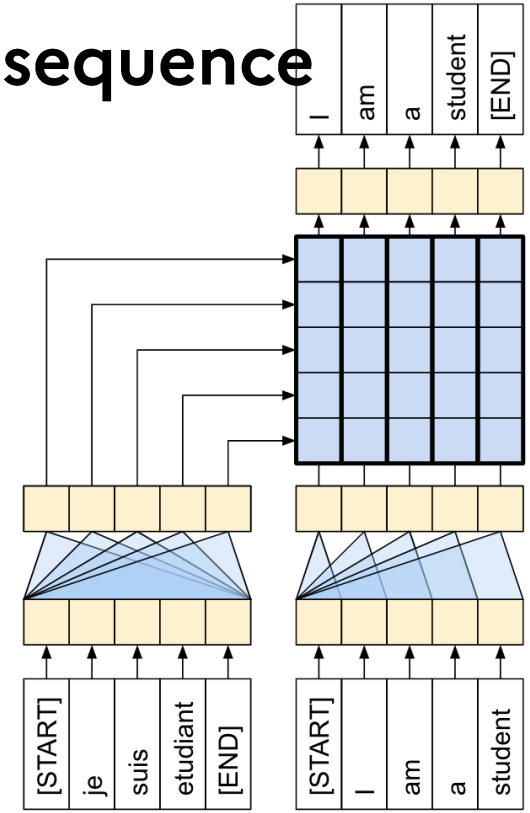
Named entity recognition



Language model



Sequence to sequence



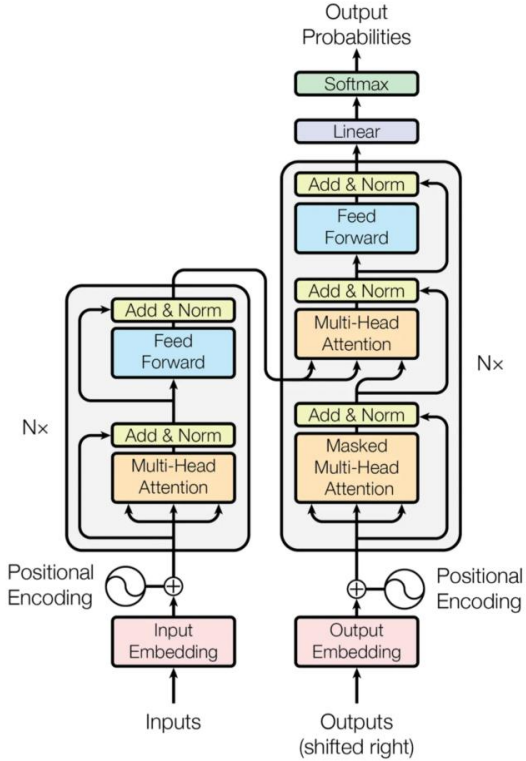
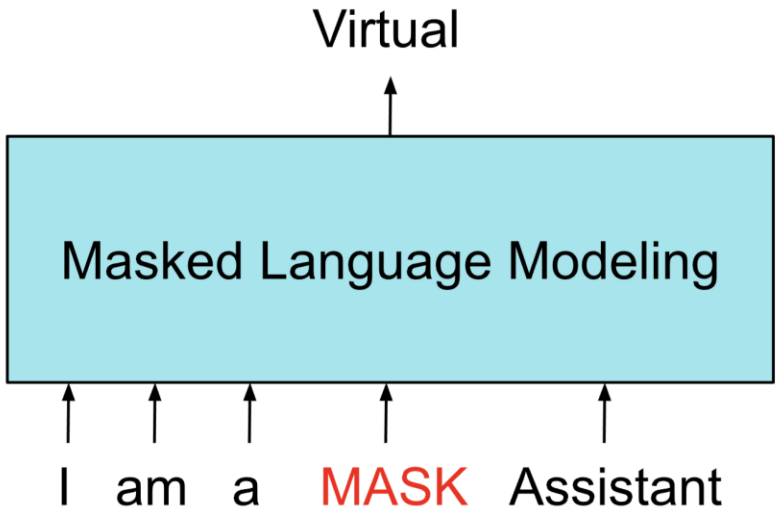
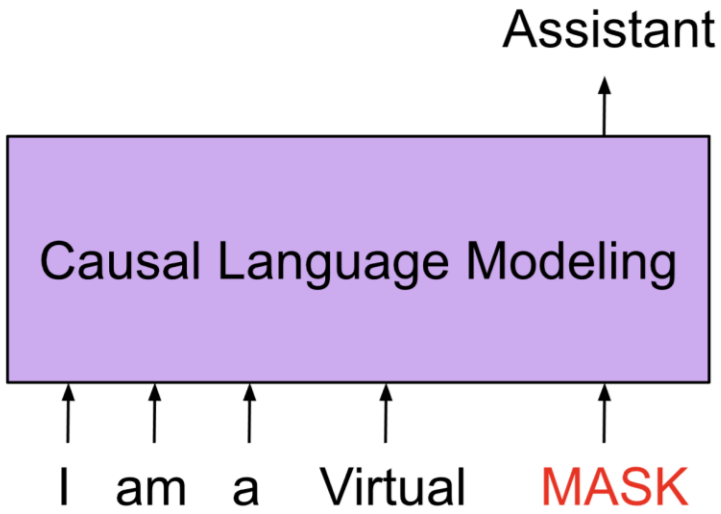


Figure 1: The Transformer - model architecture.

How to train : causal and masked language model





GPT-3

- Causal language model
- 175 billion parameters (800GB)
- Training data: 570 GB of plaintext (500 billion tokens).
  - Filtered CommonCrawl : extract from web pages in English (60 %)
  - WebText : web pages well noted on Reddit (22 %)
  - English Wikipedia (3 %)
  - Books1 and Books2 : internet-based books corpora (16 %)
- 2048-token-long context

No need for finetuning and specific head with GPT-3

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

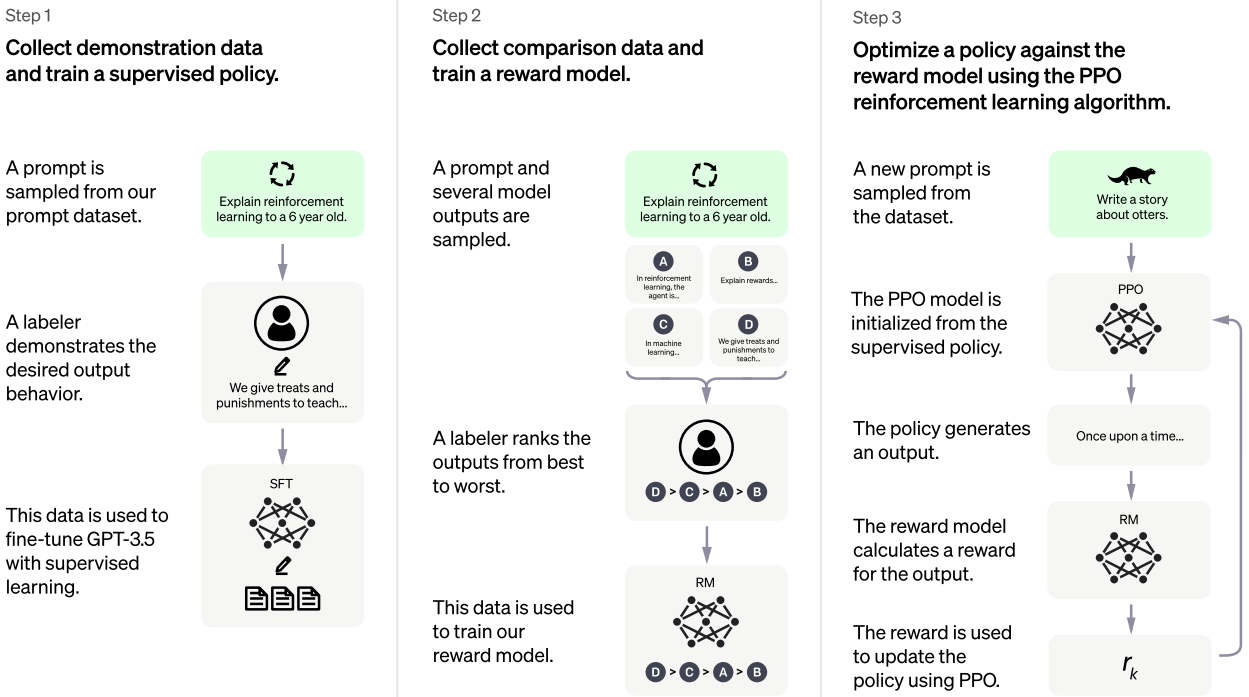
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

From GPT-3 to ChatGPT through InstructGPT

- InstructGPT: Use reinforcement learning from human feedback (RLHF) on prompt submitted by previous user.
- ChatGPT: Same as InstructGPT but with additional prompt / answer from human in form of question / answer



A good prompt is all you need !

- ChatGPT can make up facts with great confidence as long as they are credible, so don't use it for bibliography !
- Very good at summarizing text, article (it has nothing to invent). However, be careful with the size of the prompt! To generate longer texts, you must first ask him for a detailed summary.
- It is very good for form, so can be very useful to rewrite paragraphs of articles, or even generate a good first draft for your abstract or conclusion.
- To make the generated text as relevant as possible, don't hesitate to give it a lot of context : "Ignore all instructions before this one. You are a photogrammetry researcher specialized in 3D reconstruction for more than 20 years. You write a paper for CVPR. This one is about ... and is entitled "...". Your task now is to write the abstract of this paper."
- You can ask him to ask you questions to further improve the result: "You must ALWAYS ask questions BEFORE you answer so you can better zone in on what the questioner is seeking. Is that understood ?"