Modern NLP

Based on Deep Learning and Language models.
Day 3 Afternoon

Alexandre Gazagnes

2nd Day

1. Afternoon (~ 4h)
    - Very small remainder of Day 2
    - Using modern transformers
    - Transfert learning
    - Fine Tuning
    - Final Project insights

First … Let’s Talk !

NLP Tasks with pretrained models

Alexandre Gazagnes

What is a pretrained model ?

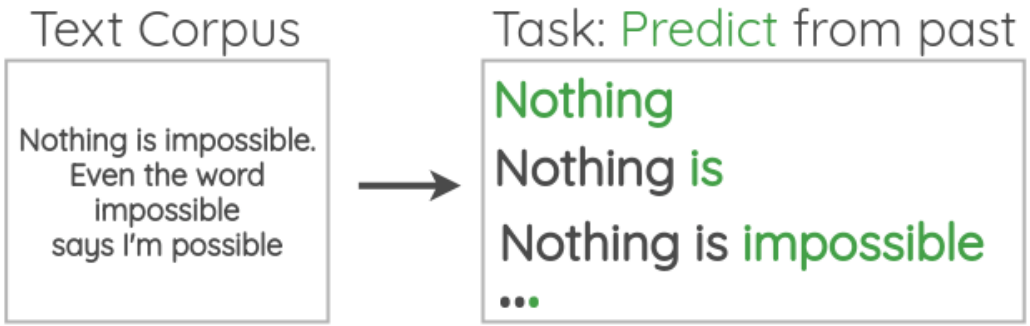
A **pretrained** model is a neural network that has been trained on a large dataset and can be used directly on your data or as a starting point for other tasks.

These models can s**ave a significant amount of time and computational resources,** as it is not necessary to train a model from scratch on a new task.

Finetuning is the process of further training a pretrained model on a specific task using task-specific data (usually small dataset). The degree of finetuning varies depending on the size and similarity of the new task to the pretraining task.

Large language models case

LLM model is presented with a large dataset of text and is then trained **to predict the next word** in a sequence of words (**Self-supervised learning)**.



Large language models have achieved **state-of-the-art performance** on a wide range of natural language processing tasks, including language translation, text summarization, and question answering.

Most used pretrained models in NLP

Bard: **Google's** Bard, initially powered by the LaMDA model with **137 billion** parameters, has been enhanced with the PaLM model, which has a staggering **540 billion parameters**. The upgrade with PaLM significantly boosts Bard's functionality, including improved reasoning, multi-step prompt handling

LLaMA: Meta's family of LLaMA models, ranging from **7 to 65 billion parameters**, provides a robust and efficient alternative for various NLP tasks, outperforming larger models in many benchmarks. Used in facebook for Content Moderation.

GPT: Developed by **OpenAI**, it is a transformer-based model with over **175 billion parameters for GPT3**, trained on a dataset of petabytes of data collected over 8 years of web crawling. GPT4 has almost one trillion parameters (the specific number of parameters for GPT-4 hasn't been publicly disclosed).

Classification & Sentiment Analysis

The process of d**etermining the category or a given text**. Sentiment Analysis is one of the most used case.

Examples:

- Determining whether a customer review of a product is positive or negative.
- Classifying a news article as belonging to a specific category (politics, sports, entertainment, etc.).

Zero-Shot Classification

Zero-shot classification is an NLP algorithm that can classify text data into a set of predefined categories or labels **without the need for any labeled training data.** Instead, the algorithm relies on natural language understanding and general knowledge to make predictions based on the input text and a set of given parameters.

Examples:

- Classifying news articles into predefined topics such as politics, sports, and entertainment, without being trained on any specific labeled data for those categories. The algorithm relies on its ability to understand the language and the context of the articles to make accurate predictions.
- Classifying customer feedback into different categories such as positive, negative, or neutral sentiments, without any labeled training data for sentiment analysis. The algorithm can use its understanding of language and common patterns in text to identify the sentiment expressed in the feedback.
- Classifying documents or texts into different languages, without any labeled data for language classification. The algorithm can analyze the input text's linguistic features and identify the most likely language based on its knowledge of language structure and patterns.
- Classifying products or services into predefined categories based on their features and characteristics, without any labeled data for product classification. The algorithm can analyze the text describing the product's features and characteristics and classify it into the most appropriate category based on its understanding of the product domain.

Information Extraction & Questing Answering

The process of extracting **specific information from a text** and answering questions based on that information.

Examples:

- Extracting information about a person from a Wikipedia article and answering questions about their education and career.

- Extracting information about a company from a financial report and answering questions about their revenue and profits.

Translation

The process of converting text from one language to another.

Example:

- Translating a Spanish news article into English.
- These models can also be adopted to generate voice : Text-to-Speech

Summarization

The process of **condensing a text** to its main points or **key information**.

Examples:

- Summarizing a long research paper into a brief summary for a conference presentation.
- Summarizing a news article about a complex legal case into a summary for a non-legal audience.

Topic Modeling

The goal of topic modeling is to uncover the underlying structure and patterns in the data, and to group together documents that are related to each other based on their content.

Topic modeling algorithms typically use techniques such as latent Dirichlet allocation (LDA) to identify the underlying topics present in a set of documents. These topics may be represented as a set of keywords or phrases, and each document may be assigned a probability distribution over the topics, indicating the degree to which it relates to each topic

Examples:

- A news organization may use topic modeling to identify the most common themes and issues discussed in their articles, and to group similar articles together for easier organization and analysis.
- A social media monitoring tool may use topic modeling to identify the most frequently discussed topics and hashtags on a particular platform, and to track changes in these topics over time.

Text Generation & Prompting

The process of creating new text based on a given input or set of parameters.

Prompting refers to the process of providing input, context and/or starting point for a text generation.

Examples:

- Generating personalized email responses based on the recipient's name and previous interactions with the company.
- Generating product descriptions for an e-commerce website based on the product's features and target audience.
- Generating creative headlines for news articles based on the content of the article.
- Generating social media posts for a brand based on current events and trending topics.

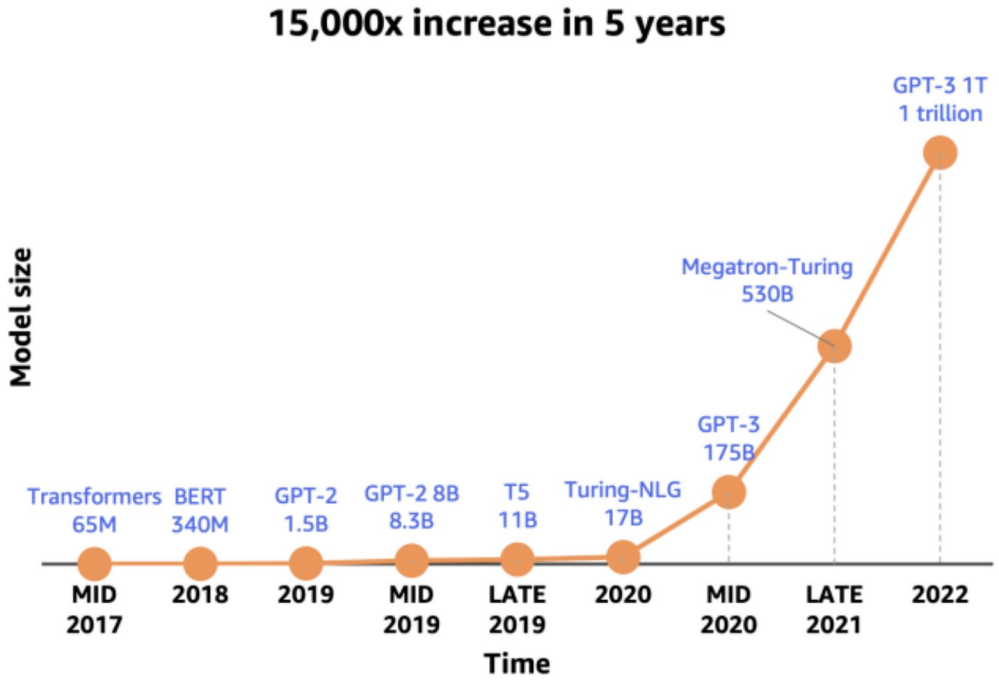
Evaluation

There are several common evaluation metrics used in natural language processing (NLP) tasks, depending on the specific task and dataset. Some of the most common metrics include:

- **Accuracy**: This is the most basic and commonly used metric, which simply calculates the percentage of correctly predicted labels or outputs.
- **Precision, Recall and F1-score**: These are used for classification tasks and take into account both true positives and false positives. Precision is the number of true positives divided by the number of true positives and false positives, recall is the number of true positives divided by the number of true positives and false negatives, and F1-score is the harmonic mean of precision and recall.
- **BLEU, ROUGE and METEOR:** These are used for text generation tasks such as machine translation, summarization, and text completion. **BLEU** compares the generated text to one or more reference texts, **ROUGE** compares the generated text to a reference text, and **METEOR** compares the generated text to a reference text, taking into account synonyms and stemming.

It is important to note that these metrics should be chosen based on the specific NLP task, and that a single metric may not provide a complete picture of the model's performance.

Exponential complexity



Practice !

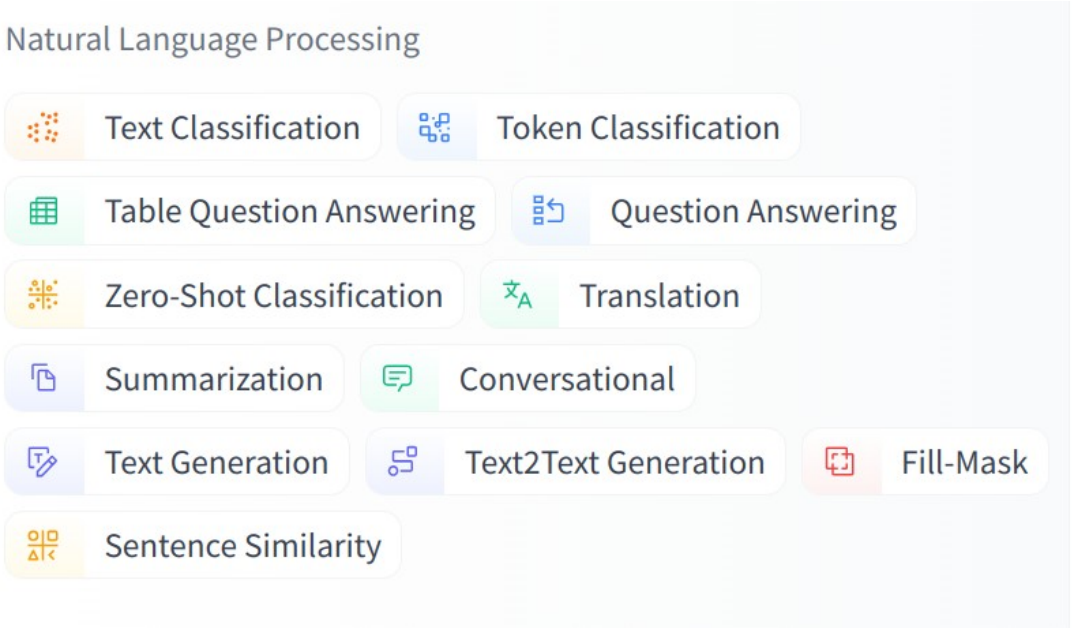
Alexandre Gazagnes

Hugging Face

About Hugging Face

- Hugging Face **democratized access to the latest NLP** models by providing an easy-to-use platform where **pre-trained models and fine-tuning capabilities are available to the public. This significantly lowers the barrier to entry for using advanced models like BERT, GPT-**2, T5, etc
- Released in **2018**, the library quickly became the go-to resource for anyone looking to implement transformer models.
- Hugging Face has a **strong community** around machine learning. They **frequently update** their libraries with the latest research outputs, new models, and tools, keeping the community at the cutting edge of AI.
- Core features :
  - **TRANSFORMERS** : The core library that provides access to pre-trained models for a wide range of NLP tasks such as text classification, information extraction, question answering, summarization, translation, and text generation.
  - **DATASETS** : An easy-to-use library that provides access to many NLP and machine learning datasets, along with tools for easily loading, processing, and transforming these datasets.
  - **HUB** : An online platform where community members can share and discover pre-trained models. This hub allows users to download models that have been trained by others, fostering a collaborative environment.

HuggingFace NLP Tasks classification



Core Function of transformers

```
from transformers import pipeline

classifier = pipeline("sentiment-analysis")
classifier("I've been waiting for a HuggingFace course my whole life.")
```

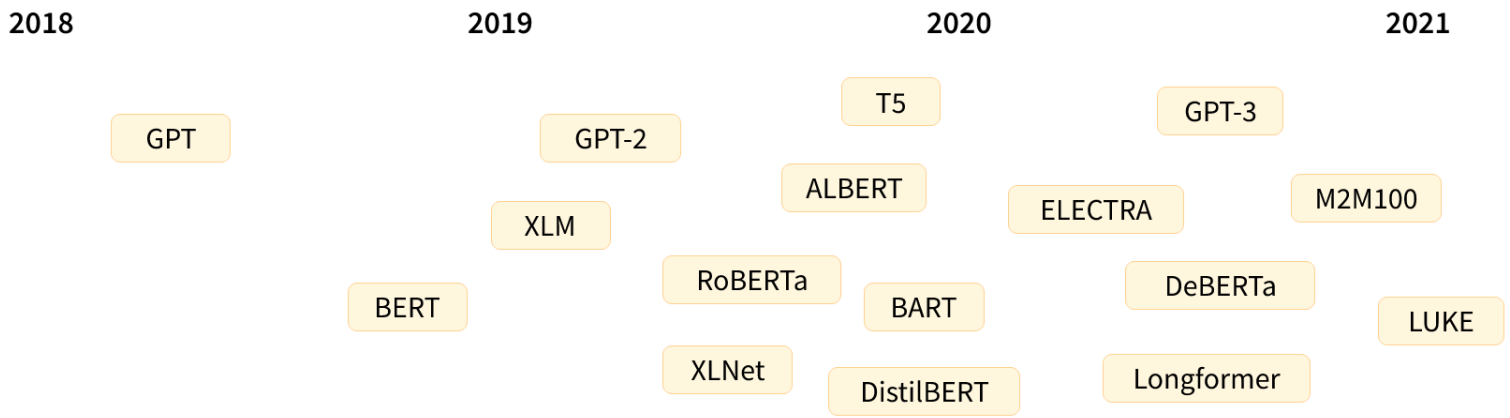
```
[{'label': 'POSITIVE', 'score': 0.9598047137260437}]
```

We can even pass several sentences!

```
classifier(
    ["I've been waiting for a HuggingFace course my whole life.", "I hate this so much!"]
)
```

```
[{'label': 'POSITIVE', 'score': 0.9598047137260437},
 {'label': 'NEGATIVE', 'score': 0.9994558095932007}]
```

Transformer's zoo



Practice !

Annexes

Annexes

Bla bla