



AI (LLMs) in IoT

Mystica
Software Developer

Find me online
mystica.me

Links to slides:



mystica.me/assets/files/llm.pdf

Agenda

- 1. why do we even need AI in IoT?**
- 2. LLMs in IoT?!**
- 3. Bit of History of NLP**
- 4. NLP Tasks**
- 5. How NLP can be used in IoT?!**
- 6. Transformers - Architecture**
- 7. Demo**

Why do we even need ML/AI/Data Analytics in IoT?

Important values of IoT?

- **collect data from tons of sensors or other microcontrollers.**
- **business insights we can gather from those data (performance of machinery equipment, reducing downtime, predictive maintenance)**

- **Real-world data is messy and will be large too.**
- **The more IoT devices added to the networks, the more the data will be. It can be overwhelming for a system to handle such an amount of data in a small period.**
- **so we go for "Big Data Analytics" when our product/system scales.**
- **gain insights from data and use ML/AI to automate**



LLMs in IoT?!

What is that good for?!

LLM - Large Language Model

- NLP (Natural Language Processing) in AI**
- Making the AI to generate human-like outputs and understand text-like inputs**
- Recent trends: ChatGPT, Bard, GPT-4, Llama 2 etc**
- most models are based on a architecture called “transformers”.**

NLP Tasks

- **Text Summarization**
- **Question answering**
- **Text classification**
- **Causal language modeling**
- **Masked language modeling**
- **Translation**

Text Summarization

Paragraph

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Summary

Lorem ipsum.....

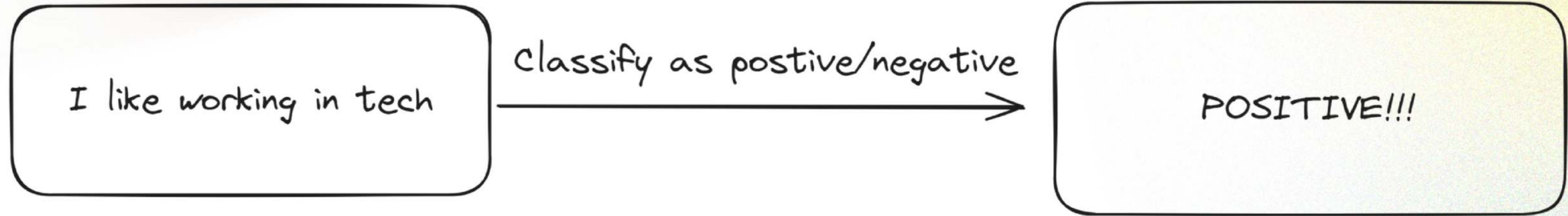
Question and Answering

The Amazon rainforest, covering much of northwestern Brazil and extending into Colombia, Peru and other South American countries, is the world's largest tropical rainforest, famed for its biodiversity. It's crisscrossed by thousands of rivers, including the powerful Amazon.

What are the countries mentioned in the passage?

Brazil, Colombia, Peru and others


Text Classification



Causal language modeling

Autocompletion

The Google logo is displayed in its characteristic multi-colored font (blue, red, yellow, green, red).

- 
- A screenshot of a Google search bar showing the text "what is NLP" and a list of suggestions. The suggestions are: "what is NLP - Google Search", "what is nlp in ai", "what is nlp training", "what is nlp in machine learning", "what is nlp therapy", "what is nlp coaching", "what is nlp in data science", "what is nlp used for", and "what is nlp in python".
- what is NLP
 - what is NLP - Google Search
 - what is nlp in ai
 - what is nlp training
 - what is nlp in machine learning
 - what is nlp therapy
 - what is nlp coaching
 - what is nlp in data science
 - what is nlp used for
 - what is nlp in python

Masked language modeling

Sky is _____

Mask

Fill in the blanks

- Sky is grey
- Sky is the limit
- Sky is blue

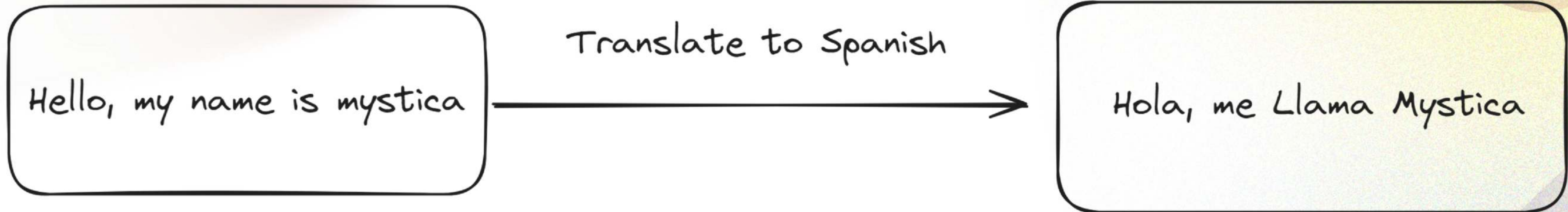
Prompt Engineering

If we are creating “chat with pdf” app, then the developer might have had the prompt as below

You are AI Assitant and have to answer the questions according to the <MASK>,

Where <MASK> is the content of the PDF

Translation



Bit of a history...

Before transformers,

- **NLP tasks (seq2seq) mainly uses RNN (Recurrent Neural Network), LSTMs**

RNN (Recurrent Neural Network)

- **Sequential data more effectively (time-series, audio, nlp tasks)**
- **Address the need to remember previous elements in a sequence, which is essential for tasks like predicting the next word in a sentence or analyzing time series data**

Disadv

- **Short term memory**
- **lacks parallel computing and doesn't take advantage of modern hardware**

Example:

Apple falls from the tree.

....

....

....

....

And Newton discovered the theory of gravity. It gave him insights about the nature of gravitation. Newton observed the fall of ???

Example:

Apple falls from the tree.

....

....

....

....

And Newton discovered the theory of gravity. It gave him insights about the nature of gravitation. Newton observed the fall of ???

Think of the bordered line as sliding window, our model can only remember that part, but we need to refer to the first sentence in the paragraph...

Example:

Apple falls from the tree.

....

....

....

....

And Newton discovered the theory of gravity. It gave him insights about the nature of gravitation. Newton observed the fall of ???

Think of the bordered line as sliding window, our model can only remember that part, but we need to refer to the first sentence in the paragraph...

“Short term dependencies”

LSTM

- **type of RNN**
- **Sequential data more effectively (time-series, audio, nlp tasks)**
- **Address the need to remember previous elements in a sequence, which is essential for tasks like predicting the next word in a sentence or analyzing time series data**
- **their ability to selectively remember or forget information over time, allowing them to capture long-term dependencies more effectively (uses something called cells, it will be containing input, forget, output part in each iteration) to overcome the short term dependencies**

Disadv

- **Long Training Time**
- **Computational Complexity (requires more memory)**

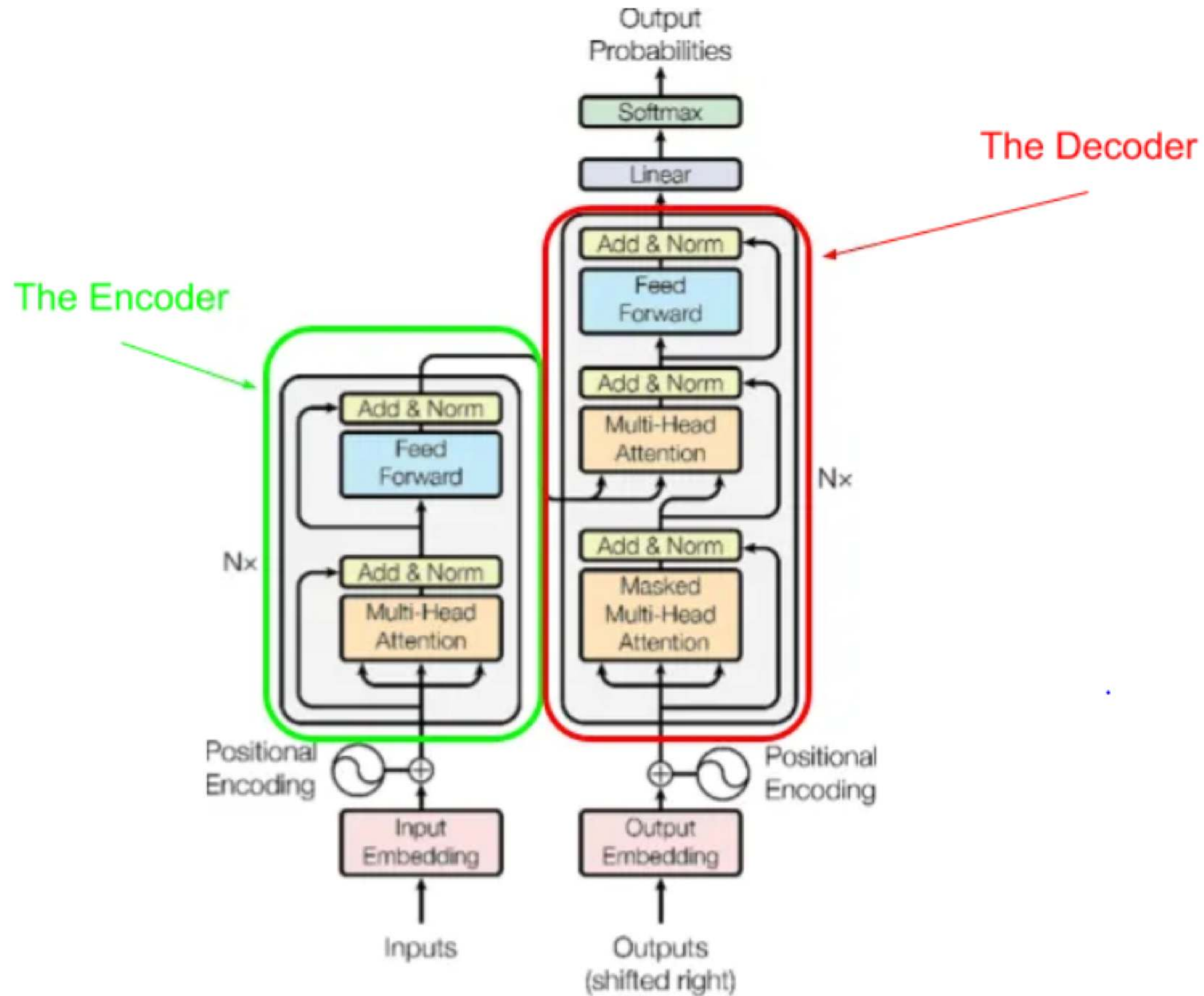
**Transformer is first introduced by the people @GoogleResearch -
“Attention is all you need”**

Transformers

- **encoder - decoder**
- **relies on a mechanism called “Self Attention”**
- **process texts in parallel**
- **uses tokenizers, embedding layers, positional encoding and multi-head attention, which contribute to their ability to process sequential data effectively.**
- **Google’s “Attention is all you need”**



Transformers



Tokenizers

- **Converts Text to Numerical representation**
- **BPE (Byte-Pair Encoding), Word Piece, sub-word, Sentence-Piece tokenizers**

Embedding layer

- **Converts tokenized input into a dense vector representation**

Positional Encoding

- **model with information about the order of tokens in the sequence**

Self Attention

- capture complex relationships within the input sequence by comparing all input sequence members with each other and modifying the corresponding output sequence positions
- Transformers relies on self attention because unlike rnn, the texts are given to the black box parallelly. And self attention is the mechanism which calculates the how much current word is relatable to self and others

Decoder

- the decoder, also known as an autoregressive model, is a crucial component in transformers, trained on the traditional language modeling problem of guessing the next token after reading the preceding ones

Transformers

- **since it does parallel processing, it takes advantage of modern GPUs**
- **used by GPT-3, DALL-E, BERT, RoBERTa, and ChatGPT etc**

Getting back to lot now,

How will LLM be useful in the Industrial lot

Getting back to lot now,

How will LLM be useful in the Industrial lot

Automate a few things, some use cases that we can think of would be:

1. Report Generation

- **Process sensor data, production logs, or maintenance records to generate detailed reports summarizing the performance, issues, or trends in an industrial setting.**

2. Q/A model

- **What was the last known incident?, what are the protocols to follow during natural disasters?**

3. Incident Announcement and Alerting

- **Announce like Voice Assistant: Attention: An incident is currently occurring in building 3. Please follow safety protocols and evacuate the area**

Creating AI Apps using LLM Models

Do you have to know in and out of NLP Practises?!

Creating AI Apps using LLM Models

Do you have to know in and out of NLP Practises?!

Answer is BIG NO!!



Then how?!

Open Source Models

Then how?!

Open Source Models

There are enormous amount of LLM Open Source/Licensed Models that we can use that abstracts away lot of complexities for us to develop and frameworks that simplifies the workflow too...

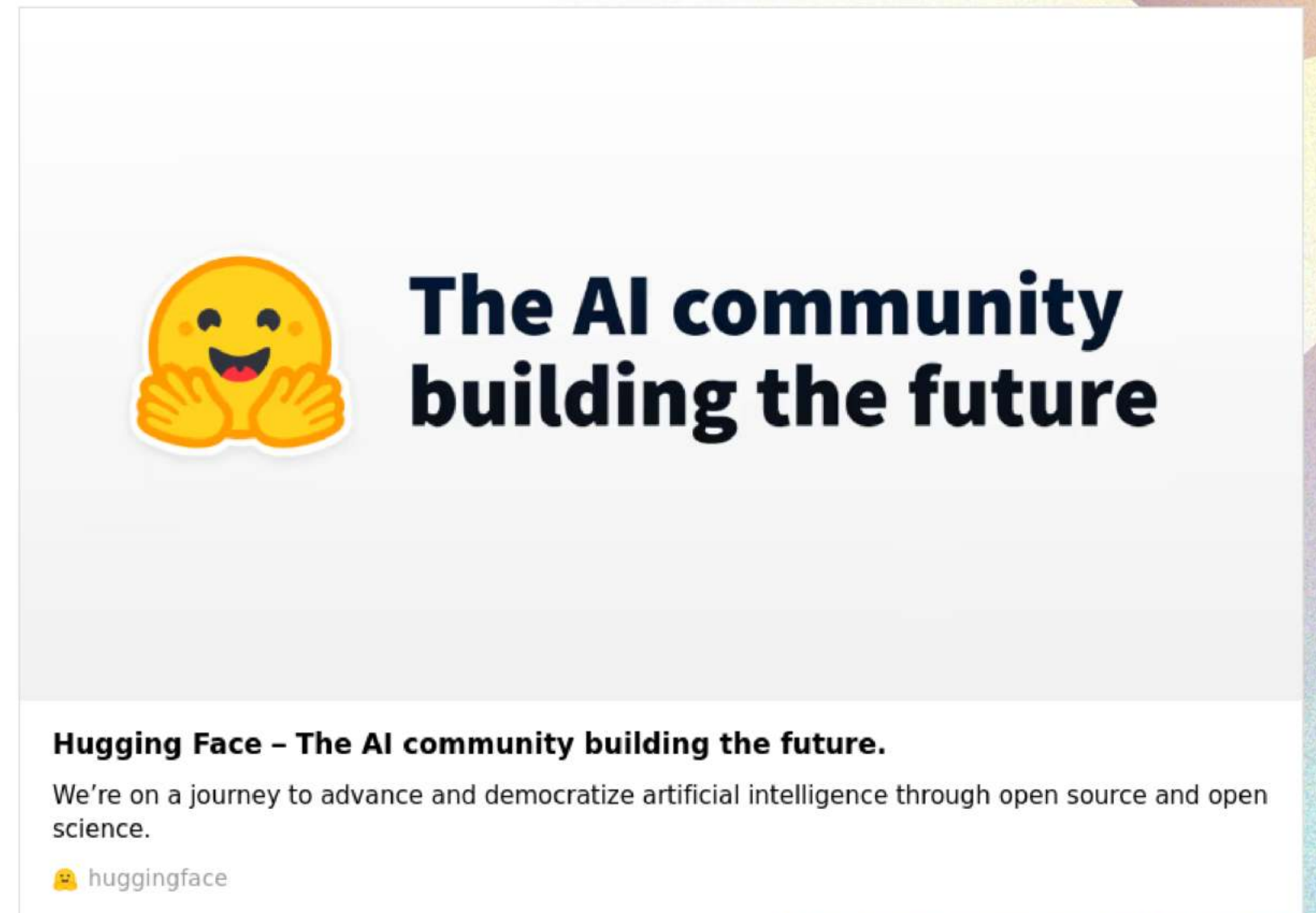
Then how?!

Open Source Models

There are enormous amount of LLM Open Source Models that we can use that abstracts away lot of complexities for us to develop and frameworks that simplifies the workflow too...

Transformers - Hugging Face 🤗

- Framework to load pretrained model reducing compute time
 - Models
 - Datasets
 - Libraries
 - Tokenizers, etc..



Not to be confused with transformer architecture!!

- **transformer architecture - modern AI is based upon its architecture**
- **transformer library - hugging face's product that helps to make AI apps faster**

DEMO!!!

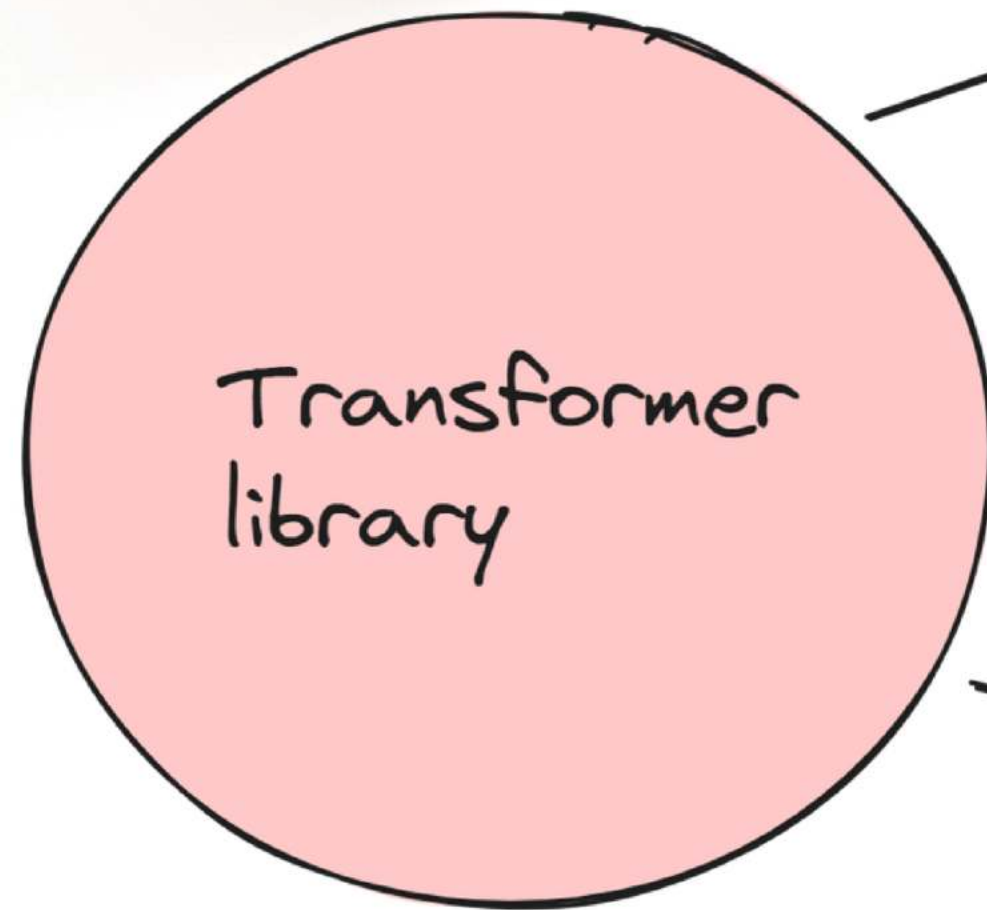
Create a text summarization, QA App

Tools:

Google colab + Pytorch + Transformers

!pip install transformers datasets

Two ways of creating AI apps



1. use as it is

- use pretrained model which is already trained using a dataset
- give input to the model
- get results!

2. finetune the model

- use a pretrained model and load our dataset into it, adjust parameters, train the domain specific model
- get results!

1. Use as It is

- use ``pipeline`` to load model and tokenizers
- give input and `tokenize.encode`
- load the tokenized input to the model
- decode the output

2. FineTuning the model

- **Load the dataset**
- **train/test split**
- **preprocess the data**
- **load the tokenizer**
- **tokenize the preprocessed data**
- **Load the model**
- **train the model with tokenized preprocessed data**
- **evaluate, test the model, save & load**

Building Demo!