An Introduction to Large Language Models (LLMs)

Large Language Models (LLMs) are foundational machine learning models that use deep learning algorithms to process and understand natural language. These models are trained on massive amounts of text data to learn patterns and entity relationships in the language. LLMs can perform many types of language tasks, such as translating languages, analyzing sentiments, chatbot conversations, and more. They can understand complex textual data, identify entities and relationships between them, and generate new text that is coherent and grammatically accurate.

Learning Objectives:

- Understand the concept of Large Language Models (LLMs) and their importance in natural language processing.
- Know about different types of popular LLMs, such as BERT, GPT-3, and T5.
- Discuss the applications and use cases of Open Source LLMs.
- Hugging Face APIs for LLMs.
- Explore the future implications of LLMs, including their potential impact on job markets, communication, and society as a whole.

This article was published as a part of the Data Science Blogathon.

Table of Contents

General Architecture

The architecture of Large Language Models primarily consists of multiple layers of neural networks, like recurrent layers, feedforward layers, embedding layers, and attention layers. These layers work together to process the input text and generate output predictions.

- The embedding layer converts each word in the input text into a high-dimensional vector representation. These embeddings capture semantic and syntactic information about the words and help the model to understand the context.
- The feedforward layers of Large Language Models have multiple fully connected layers that apply nonlinear transformations to the input embeddings. These layers help the model learn higher-level abstractions from the input text.
- The recurrent layers of LLMs are designed to interpret information from the input text in sequence. These layers maintain a hidden state that is updated at each time step, allowing the model to capture the dependencies between words in a sentence.
- The attention mechanism is another important part of LLMs, which allows the model to focus selectively on different parts of the input text. This mechanism helps the model attend to the input text's most relevant parts and generate more accurate predictions.

Examples of LLMs

Let's take a look at some popular large language models:

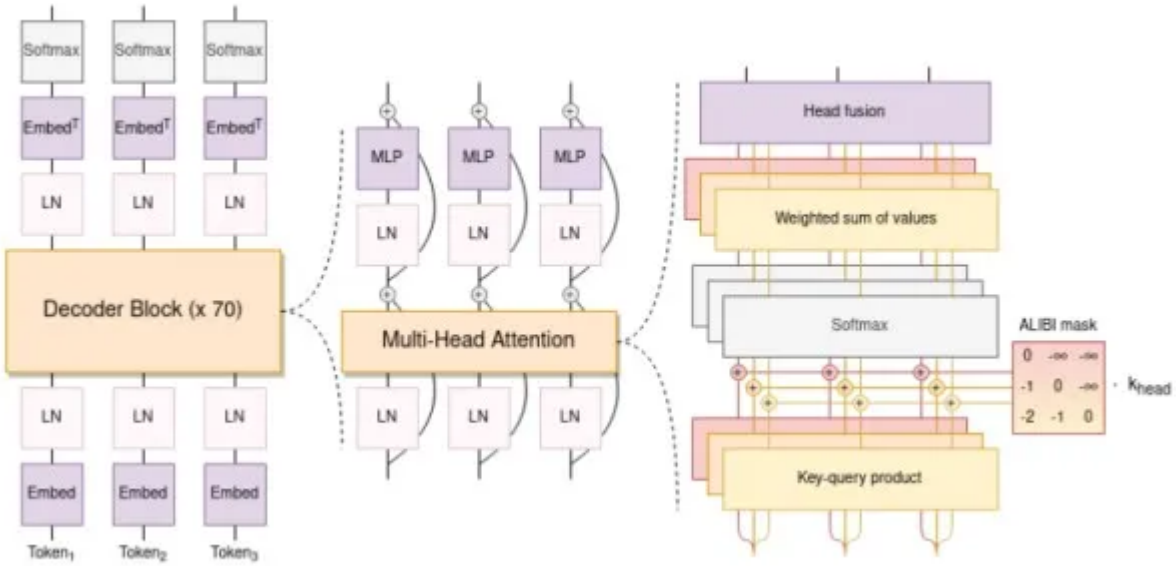
1. GPT-3 **(Generative Pre-trained Transformer 3) –** This is one of the largest Large Language Models developed by OpenAI. It has 175 billion parameters and can perform many tasks, including text generation, translation, and summarization.
2. BERT **(Bidirectional Encoder Representations from Transformers) –** Developed by Google, BERT is another popular LLM that has been trained on a massive corpus of text data. It can understand the context of a sentence and generate meaningful responses to questions.
3. **XLNet –** This LLM developed by Carnegie Mellon University and Google uses a novel approach to language modeling called "permutation language modeling." It has achieved state-of-the-art performance on language tasks, including language generation and question answering.
4. **T5 (Text-to-Text Transfer Transformer) –** T5, developed by Google, is trained on a variety of language tasks and can perform text-to-text transformations, like translating text to another language, creating a summary, and question answering.
5. RoBERTa **(Robustly Optimized BERT Pretraining Approach) –** Developed by Facebook AI Research, RoBERTa is an improved BERT version that performs better on several language tasks.

Open Source Large Language Models

The availability of open-source LLMs has revolutionized the field of natural language processing, making it easier for researchers, developers, and businesses to build applications that leverage the power of these models to build products at scale for free. One such example is Bloom. It is the first multilingual Large Language Model (LLM) trained in complete transparency by the largest collaboration of AI researchers ever involved in a single research project.

With its 176 billion parameters (larger than OpenAI's GPT-3), BLOOM can generate text in 46 natural languages and 13 programming languages. It is trained on 1.6TB of text data, 320 times the complete works of Shakespeare.

Bloom Architecture



The architecture of BLOOM shares similarities with GPT3 (auto-regressive model for next token prediction), but has been trained in 46 different languages and 13 programming languages. It consists of a decoder-only architecture with several embedding layers and multi-headed attention layers.

Bloom's architecture is suited for training in multiple languages and allows the user to translate and talk about a topic in a different language. We will look at these examples below in the code.

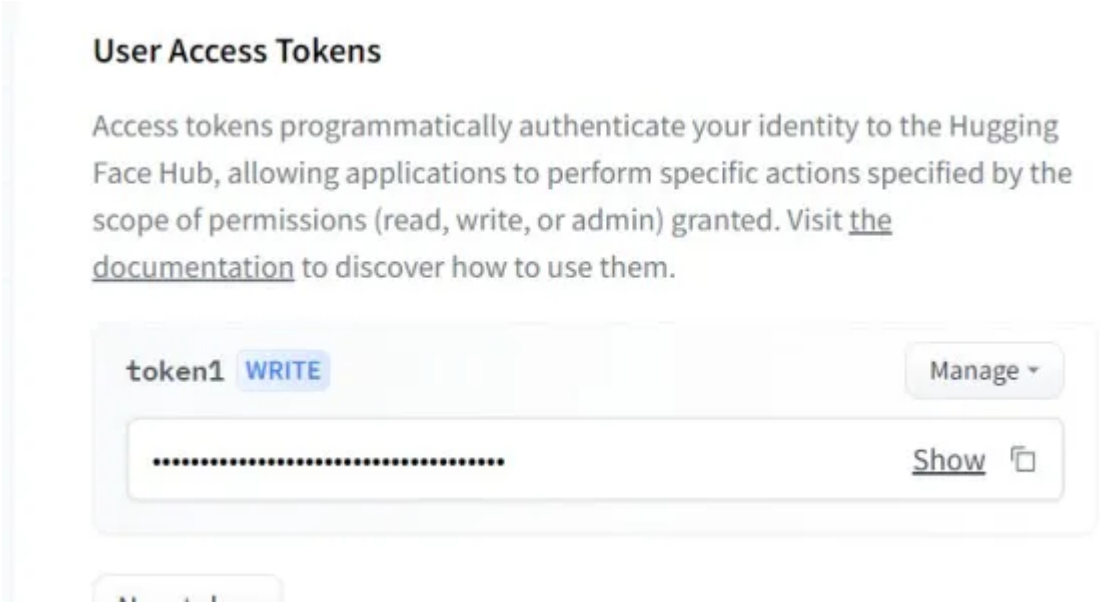
Other LLMs

We can utilize the APIs connected to pre-trained models of many of the widely available LLMs through Hugging Face.

Hugging Face APIs

Let's look into how Hugging Face APIs can help generate text using LLMs like Bloom, Roberta-base, etc. First, we need to sign up for Hugging Face and copy the token for API access. After signup, hover over to the profile

icon on the top right, click on settings, and then Access Tokens.



Example 1: Sentence Completion

Let's look at how we can use Bloom for sentence completion. The code below uses the hugging face token for API to send an API call with the input text and appropriate parameters for getting the best response.

```
import requests
from pprint import pprint

API_URL = 'https://api-inference.huggingface.co/models/bigscience/bloomz'
headers = {'Authorization': 'Entertheaccesskeyhere'}
# The Entertheaccesskeyhere is just a placeholder, which can be changed according to
the user's access key


def query(payload):
    response = requests.post(API_URL, headers=headers, json=payload)
    return response.json()

params = {'max_length': 200, 'top_k': 10, 'temperature': 2.5}
output = query({
    'inputs': 'Sherlock Holmes is a',
    'parameters': params,
})

pprint(output)
```

Temperature and top_k values can be modified to get a larger or smaller paragraph while maintaining the relevance of the generated text to the original input text. We get the following output from the code:

```
[{'generated_text': 'Sherlock Holmes is a private investigator whose cases '
                    'have inspired several film productions'}]
```

Let's look at some more examples using other LLMs.

Example 2: Question Answers

We can use the API for the Roberta-base model which can be a source to refer to and reply to. Let's change the payload to provide some information about myself and ask the model to answer questions based on that.

```
API_URL = 'https://api-inference.huggingface.co/models/deepset/roberta-base-squad2'
headers = {'Authorization': 'Entertheaccesskeyhere'}


def query(payload):
    response = requests.post(API_URL, headers=headers, json=payload)
    return response.json()

params = {'max_length': 200, 'top_k': 10, 'temperature': 2.5}
output = query({
    'inputs': {
            "question": "What's my profession?",
            "context": "My name is Suvojit and I am a Senior Data Scientist"
        },
    'parameters': params
})

pprint(output)
```

The code prints the below output correctly to the question – What is my profession?:

```
{'answer': 'Senior Data Scientist',
 'end': 51,
 'score': 0.7751647233963013,
 'start': 30}
```

Example 3: Summarization

We can summarize using Large Language Models. Let's summarize a long text describing large language models using the Bart Large CNN model. We modify the API URL and added the input text below:

```
API_URL = "https://api-inference.huggingface.co/models/facebook/bart-large-cnn"
headers = {'Authorization': 'Entertheaccesskeyhere'}


def query(payload):
    response = requests.post(API_URL, headers=headers, json=payload)
    return response.json()

params = {'do_sample': False}

full_text = '''AI applications are summarizing articles, writing stories and
engaging in long conversations — and large language models are doing
the heavy lifting.

A large language model, or LLM, is a deep learning model that can
understand, learn, summarize, translate, predict, and generate text and other
content based on knowledge gained from massive datasets.

Large language models - successful applications of
transformer models. They aren’t just for teaching AIs human languages,
but for understanding proteins, writing software code, and much, much more.

In addition to accelerating natural language processing applications —
like translation, chatbots, and AI assistants — large language models are
used in healthcare, software development, and use cases in many other fields.'''

output = query({
    'inputs': full_text,
    'parameters': params
})

pprint(output)
```

The output will print the summarized text about LLMs:

```
[{'summary_text': 'Large language models - most successful '
                  'applications of transformer models. They aren’t just for '
                  'teaching AIs human languages, but for understanding '
                  'proteins, writing software code, and much, much more. They '
                  'are used in healthcare, software development and use cases '
                  'in many other fields.'}]
```

These were some of the examples of using Hugging Face API for common large language models.

Future Implications of LLMs

In recent years, there has been specific interest in large language models (LLMs) like GPT-3, and chatbots like ChatGPT, which can generate natural language text that has very little difference from that written by humans. While LLMs have seen a breakthrough in the field of artificial intelligence (AI), there are concerns about their impact on job markets, communication, and society.

One major concern about LLMs is their potential to disrupt job markets. Large Language Models, with time, will be able to perform tasks by replacing humans like legal documents and drafts, customer support chatbots, writing news blogs, etc. This could lead to job losses for those whose work can be easily automated.

However, it is important to note that LLMs are not a replacement for human workers. They are simply a tool that can help people to be more productive and efficient in their work. While some jobs may be automated, new jobs will also be created as a result of the increased efficiency and productivity enabled by LLMs. For example, businesses may be able to create new products or services that were previously too time-consuming or expensive to develop.

LLMs have the potential to impact society in several ways. For example, LLMs could be used to create personalized education or healthcare plans, leading to better patient and student outcomes. LLMs can be used to help businesses and governments make better decisions by analyzing large amounts of data and generating insights.

Conclusion

Large Language Models (LLMs) have revolutionized the field of natural language processing, allowing for new advancements in text generation and understanding. LLMs can learn from big data, understand its context and entities, and answer user queries. This makes them a great alternative for regular usage in various tasks in several industries. However, there are concerns about the ethical implications and potential biases associated with these models. It is important to approach LLMs with a critical eye and evaluate their impact on society. With careful use and continued development, LLMs have the potential to bring about positive changes in many domains, but we should be aware of their limitations and ethical implications.

Key Takeaways:

- Large Language Models (LLMs) can understand complex sentences, understand relationships between entities and user intent, and generate new text that is coherent and grammatically correct
- The article explores the architecture of some LLMs, including embedding, feedforward, recurrent, and attention layers.
- The article discusses some of the popular LLMs like BERT, BERT, Bloom, and GPT3 and the availability of open-source LLMs.
- Hugging Face APIs can be helpful for users to generate text using LLMs like Bart-large-CNN, Roberta, Bloom, and Bart-large-CNN.
- LLMs are expected to revolutionize certain domains in the job market, communication, and society in the future.