Hands-On GenAI: LLMs, RAGs, and Agentic Systems for Beginners

Day 10

Adya Bhat and Tejas Venugopalan

Recap

LLMs

1. User input: is a (long) string
2. Preprocess text: convert to lowercase, remove special characters, etc.
3. Tokenize text: convert to separate tokens

4. Convert to embeddings (numerical representations)

1. `"What is the weather today?"`
2. `"what is the weather today"`

3. `["what", "is", "the", "weather", "today"]`

4.
```
[ [0.78 0.98 0.86],
  [0.76 0.54 0.38],
  [0.73 0.58 0.84],
  [0.28 0.23 0.88],
  [0.18 0.17 0.59]
]
```

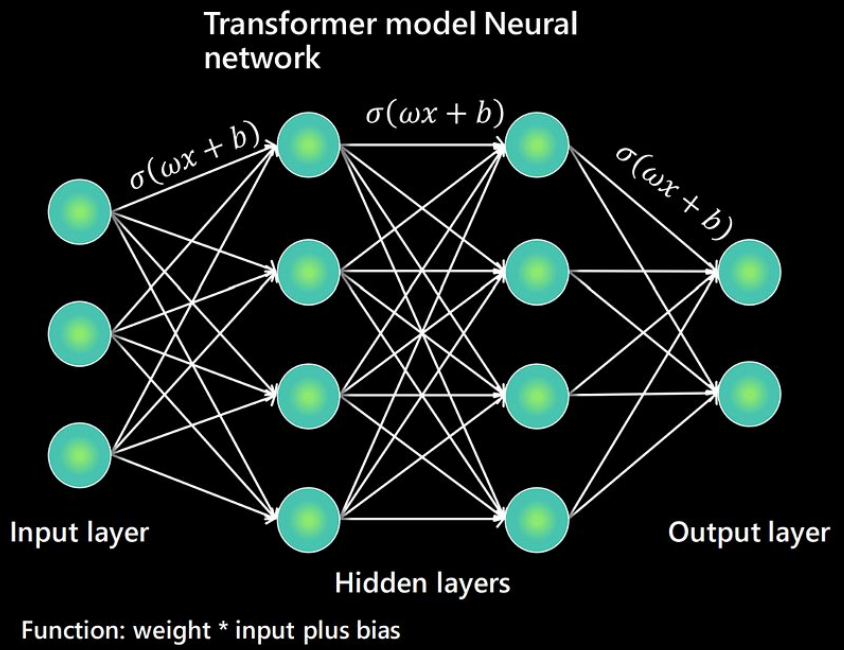
LLMs

1. ML algorithm is run on this (eg. attention) → generates a sequence of vectors

2. Vectors are decoded into their alphabetical representation

1. relationship between two word vectors (Q, K, V) = softmax($QK^T/\sqrt{(d^k)}$)*V
   => testing the model =>
   [[0.12 0.91 0.60],
   [0.21 0.45 0.37],
   [0.22 0.59 0.09],
   [0.24 0.25 0.01]
   ]
2. "It is cloudy today"

How large are they?



BERT Large - 2018
345M

GPT2 - 2019
1.5B

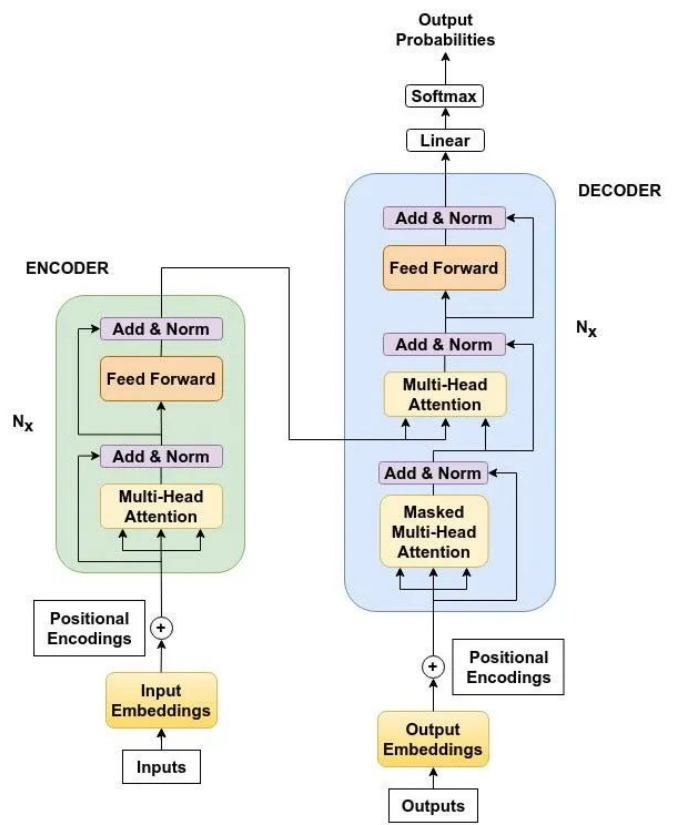
GPT3 - 2020
175B

Turing Megatron NLG 2021
530B

GPT4 – 2023
1.4T (estimated)

Source: https://bobrupakroy.medium.com/what-are-the-parameters-in-llm-76da7040e607

LLM Parameters



Transformer Architecture

Source: https://bobrupakroy.medium.com/what-are-the-parameters-in-llm-76da7040e607

RAG

1. Facts are stored as vectors in a vector database.
2. User asks a question/ prompt
3. Prompt text processing, conversion to numerical (vector representation)
4. Retrieval: of top k similar vectors, through cosine similarity ⇒ cosine similarity is computed for every fact_vector with query_vector, and top k most similar vectors are chosen.

1. `Huge collection of fact_vector (one vector for each fact)`
2. `question = "What is a smartphone battery made of?"`
3. `Result: query_vector = [0.72 0.96 0.83 0.96 0.87 0.67 0.34 0.23 0.12 0.99]`
4. `cosine_similarity(query_vector, fact_vector) = dot_product(query_vector, fact_vector)`

RAG

1. **Retrieval**: of top k similar vectors, through cosine similarity.

2. Decoding these vectors into facts.

1. `if k = 2, let's say the most similar vectors were fact_vector_5 = [0.92 0.16 0.03 0.50 0.17 0.65 0.12 0.13 0.42 0.49], and fact_vector_8 = [0.22 0.66 0.73 0.13 0.91 0.65 0.39 0.13 0.01 0.02]`

2. `fact_5 = "Smartphones typically feature lithium-ion or lithium-polymer batteries."`
   `fact_8 = "Smartphones are typically equipped with a power button and volume buttons."`

RAG

1. **Augmentation**: of prompt to LLM, by providing context of facts retrieved.

2. **Generation**: of output by LLM.

1.
```
context = fact_5 + fact_8
prompt = f"Using the
provided context:
{context}, answer this
question: {question}"
```

2.
```
generated_response = "The
answer is lithium-ion or
lithium-polymer batteries.
The answer: Lithium-ion
batteries are made of
lithium, a type of metal."
```

AI Models vs AI Agents

AI models-

- require human intervention
- predefined constraints

AI agents-

- have more autonomy
- require limited human intervention
- have goal-driven behaviour
- have adaptability

source:
https://www.ibm.com/think/topics/agentic-ai

AI Models vs AI Agents

AI model example:

- User: "What is the capital of France?"
- AI Model: "The capital of France is Paris."

source: https://dev.to/abhishekjaiswal_4896/the-difference-between-ai-agents-and-traditional-ai-models-1aj

AI agents example:

- User: "Research the latest trends in AI, summarize key points, and email me a report."
- AI Agent's Steps:
  - Searches online for the latest AI research papers and news articles.
  - Summarizes key trends and insights from multiple sources.
  - Generates a well-structured report.
  - Sends the report to the user's email automatically.

An Agent in a Project: Demo

Thank You!