SWINBURNE UNIVERSITY OF TECHNOLOGY

COS30082
Applied Machine Learning

Lecture 11
Large Language Models (LLMs)

What Are Large Language Models (LLMs)

- Large Language Models (LLMs) are advanced AI models, often rooted in the Transformer architecture, created to comprehend and produce human language, code, and beyond.
- Trained on extensive text data, they capture language nuances.
- Versatile in tasks, they excel in accuracy, fluency, and style.

Terminologies

1. **Token**: The smallest unit of semantic meaning in text analysis, tokens are derived by segmenting text into smaller parts. These serve as the foundational input for an LLM.
2. **Vocabulary**: Defined as a specific and limited collection of tokens that a model recognizes and utilizes.
3. **Tokenization**: This process divides text into tokens, enabling LLMs to efficiently interpret and manage textual data.
4. **Autoregressive Models**: These models predict subsequent tokens based solely on prior ones within a sentence, making them highly effective for generating text. GPT is a prominent example.
5. **Autoencoding Models**: Designed to restore the original text from a modified input, these models primarily focus on tasks like sentence or token classification. BERT exemplifies this model type.

Three Categories of LLMs

- **Autoregressive models**, such as GPT, which predict the next token in a sentence based on the previous tokens. These LLMs are effective at generating coherent free-text following a given context.
- **Autoencoding models**, such as BERT, which build a bidirectional representation of a sentence by masking some of the input tokens and trying to predict them from the remaining ones. These LLMs are adept at capturing contextual relationships between tokens quickly and at scale, which makes them great candidates for text classification tasks.
- **Combinations** of autoregressive and autoencoding, such as T5, which can use the encoder and decoder to be more versatile and flexible in generating text. Such combination models can generate more diverse and creative text in different contexts compared to pure decoder-based autoregressive models due to their ability to capture additional context using the encoder.

Pre-training

- Every LLM on the market has been **pre-trained** on a large corpus of text data and on specific language modeling-related tasks.
- During pre-training, the LLM tries to learn and understand general language and relationships between words.
- Every LLM is trained on different corpora and on different tasks.

Pre-training

- BERT was originally pre-trained on two publicly available text corpora:
  - English Wikipedia: A comprehensive collection from Wikipedia's English articles, spanning diverse topics and styles, encompassing about 2.5 billion words.
  - BookCorpus: An extensive dataset of both fiction and nonfiction books, covering various genres, totaling around 800 million words.

- Modern LLMs are trained on datasets vastly larger, thousands of times the size of traditional ones.

Pre-training

- BERT was pre-trained on two specific language modeling tasks:
  - Masked Language Modeling (MLM) task (autoencoding task): helps BERT recognize token interactions within a single sentence.
  - Next Sentence Prediction (NSP) task: helps BERT understand how tokens interact with each other between sentences.

Transfer Learning

- Transfer learning is a machine learning technique that harnesses knowledge from one task to enhance performance in a related task.
- For LLMs, it involves fine-tuning a pre-trained model for a specific task, like text classification or generation, utilizing its already acquired language knowledge.
- This approach significantly reduces training time and resources compared to training models from scratch.

Fine-Tuning

- Fine-tuning involves training the LLM on a smaller, task-specific dataset to adjust its parameters for the specific task at hand.
- This allows the LLM to leverage its pre-trained knowledge of the language to improve its accuracy for the specific task.
- Fine-tuning has been shown to drastically improve performance on domain-specific and task-specific tasks and lets LLMs adapt quickly to a wide variety of NLP applications.

Fine-Tuning Steps

1. Define the model we want to fine-tune as well as any fine-tuning parameters (e.g., learning rate).
2. Aggregate some training data (the format and other characteristics depend on the model we are updating).
3. Compute losses (a measure of error) and gradients (information about how to change the model to minimize error).
4. Update the model through backpropagation—a mechanism to update model parameters to minimize errors.

Popular Modern LLMs

- BERT, developed by Google, is an autoencoding model with bidirectional sentence representation, optimized for tasks like classification, utilizing the Transformer's encoder for efficient text processing.
- ChatGPT, part of the GPT-3 family by OpenAI, operates as an autoregressive model, excelling in text generation with a focus on the Transformer's decoder and large context window.
- T5, a Google creation, stands out as an encoder/decoder Transformer model, showcasing versatility in NLP tasks from classification to summarization and generation, handling multiple tasks without fine-tuning.

Applications of LLMs

- **Using Pre-trained LLMs in Larger Architectures**

  Example: Building an information retrieval system using pre-trained BERT/GPT; Fine-Tuning Pre-trained LLMs for Specific Tasks

- **Employing transfer learning to adapt models for specialized tasks**

  Example: Customizing T5 for domain-specific document summarization; Direct Application of Pre-trained LLMs

- **Leveraging models for tasks they're pre-trained to solve**

  Examples: Prompting GPT-3 for blog post writing; Using T5 for language translation.

- … …

Classical NLP Tasks

- Text Classification: Assigns labels to text, used in sentiment analysis and topic classification.
- Translation Tasks: Translates text between languages, preserving meaning and context.
- SQL Generation: Converts natural language to SQL, enabling language-to-code translation.
- Free-Text Generation: Writes blogs, emails, academic papers, aiding in content creation.
- Information Retrieval/Neural Semantic Search: Dynamically updates information using LLMs for accurate document retrieval.
- Chatbots: Facilitates conversational AI, replacing traditional chatbot design with more advanced, context-aware systems.

GPT (Generative Pretrained Transformer)

- GPT stands for Generative Pretrained Transformer, developed by OpenAI.
- Built on the Transformer architecture.
- Designed for natural language understanding and generation, making it versatile for various applications.

- History
- GPT1, 2018, parameters 117M, training data about 1GB
- GPT2, 2019, parameters 1.5B, training data 40GB
- GPT3, 2020, parameters 175B, training data 580GB
- GPT4, GPT-4, released in March 2023 by OpenAI. Architecture details and parameter count not published.