LLM Basics

Subhasis Dasgupta

August 9, 2024

Evolution of Language Modeling

- **50s Era: The Beginning**
  - Claude Shannon: Applied information theory to human language.
  - Simple n-gram Models: Foundation for tasks like speech recognition and machine translation.
- **Advancement to Statistical Models**
  - Central to Natural Language Understanding (NLU) tasks.

The Four Waves of Language Modeling

- **Wave 1: Statistical Language Models (SLMs)**
  - Key Concept: n-gram Models
  - Challenges: Data sparsity, need for smoothing.
- **Wave 2: Neural Language Models (NLMs)**
  - Innovation: Low-dimensional word embeddings.
  - Advancement: Semantic similarity, cross-modal tasks.

The Rise of Pre-trained Language Models (PLMs)

- **Wave 3: PLMs**
  - Task-Agnostic Approach: Pre-training and fine-tuning.
  - Technology: RNNs and Transformers.
  - Significance: Generalization across tasks.

The Advent of Large Language Models (LLMs)

- **Wave 4: LLMs**
  - Scale: Tens to hundreds of billions of parameters.
  - Impact: Multi-step reasoning, diverse tasks.
  - Examples: PaLM, LLaMA, GPT-4.
  - Foundation for AGI: Paving the way for general-purpose AI systems.

The Future of LLMs and AI

- Rapid Innovation: The fast pace of developments in LLMs.
- Challenges: Keeping up with new techniques and models.
- Conclusion: The importance of ongoing research and adaptation.

Encoder-Only PLMs

- **Overview:**
  - Focus on understanding and representing input text.
  - Used for tasks like classification, question answering, and language inference.
- **Key Models:**
  - **BERT** (Bidirectional Encoder Representations from Transformers):
    - Revolutionized language understanding tasks.
    - Inspired similar models like RoBERTa and ALBERT.
  - **RoBERTa**:
    - Improved BERT with better hyperparameters and training strategies.
  - **ALBERT**:
    - Optimized BERT for memory and training efficiency.
  - **DeBERTa**:
    - Enhanced BERT with disentangled attention and improved decoding mechanisms.

Decoder-Only PLMs

- **Overview:**
  - Designed for generating text based on a given context.
  - Excel in tasks such as text generation, language modeling, and dialogue systems.
- **Key Models:**
  - **GPT-1** (Generative Pre-trained Transformer):
    - Introduced the concept of pre-training a Transformer on unlabeled text.
    - Laid the groundwork for subsequent models.
  - **GPT-2**:
    - Demonstrated that language models could perform specific tasks without explicit supervision.
    - Expanded on GPT-1 with larger datasets and architectural improvements.

Encoder-Decoder PLMs

- **Overview:**
  - Combine the strengths of both encoders and decoders.
  - Versatile, handling both natural language understanding and generation tasks.
- **Key Models:**
  - **T5** (Text-to-Text Transfer Transformer):
    - Unified framework for all NLP tasks as text-to-text problems.
    - **mT5**: A multilingual variant pre-trained on texts in 101 languages.
  - **MASS** (Masked Sequence to Sequence):
    - Pre-trains by reconstructing a masked sentence fragment, integrating both encoding and generation.
  - **BART** (Bidirectional and Auto-Regressive Transformers):
    - Pre-trained by corrupting and reconstructing text, useful for translation and summarization.

Key Models

- **GPT-1:**
  - The first in the GPT series, demonstrating the potential of pre-training a Transformer model on a large text corpus.
  - **Notable Feature:** Introduced generative pre-training followed by fine-tuning.
- **GPT-2:**
  - Expanded upon GPT-1 with a larger model and dataset, generating coherent, contextually relevant text.
  - **Notable Feature:** Human-like text generation, raising discussions on AI ethics.
- **GPT-3:**
  - Marked a leap with 175 billion parameters, enabling in-context learning for various tasks.
  - **Notable Feature:** Emergent abilities in translation, question answering, and reasoning.

Key Models (cont.)

- **CODEX:**
  - A fine-tuned version of GPT-3 for programming tasks, translating natural language into code.
  - **Notable Feature:** Powers GitHub Copilot, generating code snippets from descriptive prompts.
- **ChatGPT:**
  - An interactive chatbot based on GPT-3.5 and GPT-4, capable of performing a wide range of conversational tasks.
  - **Notable Feature:** Widely used for customer support, education, and as a general-purpose AI assistant.
- **GPT-4:**
  - The latest and most advanced model, with multimodal capabilities processing both text and images.
  - **Notable Feature:** Demonstrates human-level performance on various benchmarks, including professional exams.

Significance

- The GPT family has set new standards in AI language models, influencing the development of subsequent LLMs across the industry.
- Each iteration has expanded the boundaries of what AI can achieve in natural language understanding and generation.

Overview of LLaMA

- **Developed by:** Meta
- **Model Type:** Open-source foundation language models
- **Purpose:** To provide powerful, open-access LLMs for research and development, competing with closed-source models.

Key Features of LLaMA Models

- Open-source: Model weights are available under a noncommercial license.
- Architecture: Based on GPT-3 with modifications such as SwiGLU activation, rotary positional embeddings, and RMS layer normalization.
- LLaMA-13B: Outperforms GPT-3 (175B) on most benchmarks.

Key Models in the LLaMA Family

- **LLaMA-2 (2023):**
  - Includes both foundation models and fine-tuned Chat models.
  - LLaMA-2 Chat models excel in dialogue tasks with iterative refinement using RLHF and other optimization techniques.
- **Guanaco:**
  - Efficiently fine-tuned using QLoRA on a single GPU.
  - Reaches 99.3
- **Koala:**
  - Focused on interaction data from closed-source chat models.
  - Competitive performance with state-of-the-art chat models.

Advanced LLaMA Models

- **Mistral-7B:**
  - A 7B-parameter model engineered for efficiency and performance.
  - Outperforms larger models in reasoning, mathematics, and code generation.
  - Features grouped-query attention and sliding window attention for faster inference.
- **Other Models:**
  - The LLaMA family includes a variety of specialized models such as Code LLaMA, Gorilla, Giraffe, Vigogne, and more.
  - These models are designed for specific tasks, including coding, long-text processing, and multilingual support.

Significance of the LLaMA Family

- LLaMA models promote open research and transparency in the AI community.
- Rapid growth due to open-source access, enabling wide adoption and innovation.
- LLaMA models set new standards for efficiency and performance in open-source LLMs.

Overview About PaLM

- **Developed by:** Google
- **Model Type:** Transformer-based LLMs
- **Purpose:** To achieve state-of-the-art performance across language understanding and generation tasks through large-scale, efficient training.

Key Models in the PaLM Family

- **PaLM-540B (2022):**
  - 540 billion parameters, trained on 780 billion tokens.
  - Utilizes Google's Pathways system for highly efficient training.
  - Achieves state-of-the-art few-shot learning results on numerous benchmarks.
- **U-PaLM:**
  - Continually trained on PaLM with UL2R, achieving approximately 2x computational savings.
  - Scales include 8B, 62B, and 540B parameters.
- **Flan-PaLM:**
  - Instruction-finetuned version of U-PaLM.
  - Outperforms PaLM-540B by a large margin (+9.4
  - Finetuned on 1.8K tasks, with a large and diverse dataset.

Advanced PaLM Models

- **PaLM-2:**
  - More compute-efficient with better multilingual and reasoning capabilities.
  - Trained using a mixture of objectives for enhanced performance on downstream tasks.
- **Med-PaLM:**
  - Domain-specific model for medical applications.
  - Finetuned using instruction prompt tuning for alignment with healthcare tasks.
  - Med-PaLM 2 further improves performance with med-domain finetuning and ensemble prompting.

Significance of the PaLM Family

- PaLM models set new benchmarks for few-shot learning, multilingual understanding, and specialized domains like healthcare.
- PaLM-2 and Med-PaLM demonstrate the versatility and scalability of the PaLM architecture in both general and domain-specific tasks.
- Continuous innovation with U-PaLM and Flan-PaLM emphasizes Google's commitment to pushing the boundaries of LLM capabilities.

Dominant Architectures in Language Models (1/2)

1) Transformer

- Proposed by Vaswani et al.
- Utilizes self-attention mechanism for capturing long-term contextual information.
- Composed of an encoder and a decoder, each with multi-head self-attention layers.
- Suitable for tasks like machine translation.

2) Encoder-Only

- Attention layers access all words in the sentence.
- Pre-training involves reconstructing corrupted sentences.
- Effective for understanding full sequences.
- Example: BERT (Bidirectional Encoder Representations).

Dominant Architectures in Language Models (2/2)

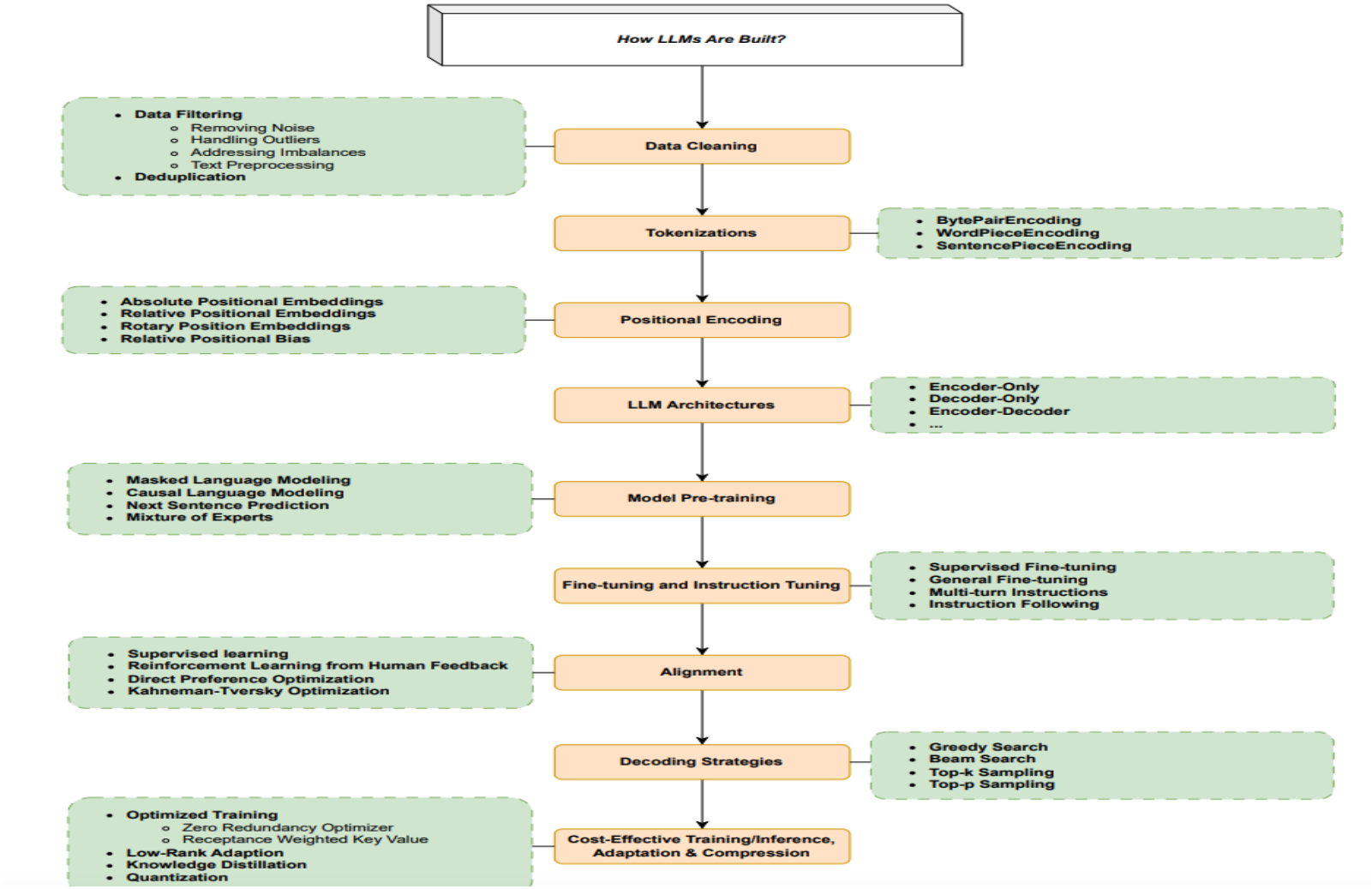
3) Decoder-Only

- Attention layers only access previous words in the sequence.
- Pre-training involves predicting the next word in the sequence.
- Ideal for text generation tasks.
- Example: GPT models (Generative Pre-trained Transformers).

4) Encoder-Decoder

- Combines encoder and decoder; also known as sequence-to-sequence models.
- Encoder accesses all words; decoder accesses previous words in the sequence.
- Best for tasks like summarization, translation, and generative question answering.

LLM Path Overview



Minaee S, et.al. Large language models: A survey 2024

References I

- Vaswani, A., et al. (2017). "Attention is all you need." *Advances in Neural Information Processing Systems*, 5998-6008.
- Devlin, J., et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*.
- Brown, T., et al. (2020). "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Raffel, C., et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research*, 21(140), 1-67.
- Chowdhery, A., et al. (2022). "PaLM: Scaling Language Modeling with Pathways." *arXiv preprint arXiv:2204.02311*.
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). "Large language models: A survey." *arXiv preprint arXiv:2402.06196*.

Thank You!

For any questions, feel free to contact me:

sudasgupta@ucsd.edu