General Language Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Model** | **Origin** | **Release Time** | **Publication** | **Comment** | |
|  | Transformer | Google | 2017.06 | Attention is All You Need |  | The encoder-decoder structure inspires the development of the models such as BERT，GPT, GPT-2, -3, and ChatGPT. |
|  | GPT | OpenAI | 2018.06 | Improving Language Understanding by Generative Pre-Training | **Model** | the pretrained base model is a 12-layer decoder-only Transformer with masked self-attention heads; Besides unsupervised pre-training, it also adopts supervised fine-tuning. |
| **Data** | BooksCorpus with 7000 unique books; 1B Word Benchmark; ELMo |
|  | BERT | Google | 2018.10, 2019.05 | Pre-training of deep bidirectional transformers for language understanding |  |  |
|  | GPT-2 | OpenAI | 2019.02 | Language Models are Unsupervised Multitask Learners. | **Model** | 1.5B parameters. It follows the GPT model with modifications. |
| **Data** | WebText with over 8 million webpages for 40GB of texts.  Perform a wide range of tasks in a zero-shot setting. |
|  | GPT-3 | OpenAI | 2020.05, 2020.07 | Language Models are Few-Shot Learners | **Model** | It is an autoregressive model with 175B parameters and 96 layers. The model is similar to GPT-2 but with larger model size, dataset size, diversity, and length of training. |
| **Data** | a filtered CommonCrawl dataset together with an expanded version of WebText dataset, two books corpora, and English-language Wikipedia.  Test different levels of task-specific training: fine-tuning, few-shot, one-shot, and zero-shot. |
|  | GPT-4 | OpenAI | 2023.03 | Technical Report  https://openai.com/research/gpt-4?ref=emergentmind |  | Predictable scaling: predict GPT-4’s final loss by extrapolating from models trained using the same methodology but much less computing resource.  Fine-tune the base model using reinforcement learning with human feedback (RLHF)  Receive a prompt of text and images |
|  | BLOOM | BigScience  (Open-source) | 2022.11 | BLOOM: A 176B-Parameter Open-Access Multilingual Language Model | **Model**: | The model contains 176B parameters.  The decoder-only architecture, with ALiBi positional embedding instead of positional encoding and with an additional layer normalization after the embedding layer. |
| **Data** | Pretraining: the ROOTS corpus (498 Hugging Face datasets amounting to 1.6TB texts that span 46 natural languages and 13 programming languages);  Finetuning: xP3 (a subset of Public Pool of Prompts plus datasets in languages other than English and new tasks) |
|  | LLaMA | Meta AI  (Open-source) | 2023.02 | LLaMA: Open and Efficient Foundation Language. Models | **Model** | A series of models ranging from 7B to 65B parameters.  The LLaMA-\*B are pretrained base models; LLaMA-I is a fine-tuned one specific for the task of MMLU. |
| **Data** | English CommonCrawl, C4 (Filtered CommonCrawl), Github, Wiki covering 20 languages, Gutenberg and Books3 (book corpora), ArXiv latex, Stack Exchange Q&A. The dataset contains **1.4T tokens** |
|  | Chinchilla | DeepMind | 2022.03 | Training Compute-Optimal Large Language Models | **Model** | In this paper, the authors first discussed about the undertrained problem. Then, a model with 70B parameters was trained on 1.4T tokens. |
| **Data** |  |
|  | PaLM |  |  |  |  |  |

LLM for Code

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Model** | **Origin** | **Release Time** | **Publication** |  | **Comment** |
|  | Codex | OpenAI | 2021.07 | Evaluating Large Language Models Trained on Code | **Model** | It is a GPT model fine-tuned on publicly available code from GitHub. It contains 12B parameters |
| **Data** | 179GB of unique Python files under 1MB from 54 million public software repositories on GitHub. (159GB after filtering) |
|  | StarCoder | BigCode  (Open-source) | 2023.05 | StarCoder: May the Source Be With You! | **Model** | The StarCoderBase and StarCoder models both contain 15.5B parameters; They are decoder-only Transformers with Fill-in-the-Middle, Multi-Query-Attention, and learned absolute positional embeddings. |
| **Data** | The base model StarCoderBase was trained on 1T tokens sourced from 80+ programming languages, GitHub issues, Git commits, and Jupyter notebooks  The StarCoder model was achieved by fine-tuning the base model on another 35B Python tokens. |
|  |  |  |  |  | **Model** |  |
| **Data** |  |