GPT-2 Model Summary

2_GPT2language_models_are_unsu...
PDF

Extract and explain contents

The document **"Language Models are Unsupervised Multitask Learners"** by Alec Radford et al. introduces GPT-2, an advanced unsupervised language model trained on a large dataset called **WebText**. The paper explores how GPT-2 can perform multiple NLP tasks **without explicit supervision**, demonstrating the potential of **zero-shot learning**.

Key Points:

1. **Unsupervised Learning & Zero-Shot Transfer**
    - Traditional NLP tasks (question answering, summarization, translation) require supervised datasets.
    - GPT-2, trained on WebText (millions of webpages), learns these tasks **without explicit supervision** by leveraging context.
    - Achieves **55 F1 on CoQA** without training on its 127,000 examples.
2. **Model Architecture & Training**
    - GPT-2 is a **1.5 billion parameter Transformer model**.
    - It improves performance in a **log-linear fashion** as model size increases.
    - Uses **Byte Pair Encoding (BPE)** for handling diverse text inputs.
3. **Performance on NLP Benchmarks**
    - **Outperforms previous state-of-the-art (SOTA) models** on **7 out of 8** language modeling datasets.
    - Excels in **reading comprehension, translation, question answering, and summarization**.
    - Demonstrates **strong generalization** but still underfits WebText.
4. **Training Dataset - WebText**
    - Scraped from **Reddit links with at least 3 karma** to ensure quality.
    - Contains **40GB of text** from over **8 million documents**.
    - **Wikipedia was excluded** to prevent overlap with existing NLP benchmarks.

5. **Experiments & Insights**
    - **Reading Comprehension:** Matches or exceeds **3 out of 4 baselines** without fine-tuning.
    - **Summarization:** Performs well but struggles with **factual consistency**.
    - **Question Answering:** Answers **4.1% of Natural Questions dataset correctly**, improving significantly with model size.
    - **Translation:** Achieves **5 BLEU on English-French**, improving when given context.
6. **Generalization vs. Memorization**
    - Overlap analysis shows **minimal memorization**, meaning GPT-2 genuinely **learns patterns** rather than memorizing text.
    - Some memorization exists for **high-frequency phrases** like famous speeches.
7. **Conclusions & Future Work**
    - **High-capacity language models can learn tasks naturally from diverse datasets**.
    - Zero-shot performance indicates **unsupervised learning has vast potential**.
    - Fine-tuning could further improve performance across tasks.

Would you like a more detailed explanation on any of these points? 🚀

Explain it's architecture

GPT-2 Architecture Explained 🏗️

GPT-2 (Generative Pre-trained Transformer 2) is a **Transformer-based** neural network model designed for **unsupervised language modeling**. It improves upon the original **GPT (Radford et al., 2018)** by significantly increasing model capacity and dataset diversity.

---

1. Transformer-Based Architecture

GPT-2 is built on the **Transformer decoder** architecture, which uses **self-attention** and **feed-forward layers** to model text. It follows the **stacked multi-layer Transformer decoder** structure.

- Unlike BERT (which uses both encoder and decoder), **GPT-2 uses only the decoder**.

- It generates text **autoregressively**, meaning it predicts the next word based on previous words.

2. Model Components

(a) Tokenization: Byte-Pair Encoding (BPE)

- Uses **Byte-Pair Encoding (BPE)** to break words into subword units.
- Helps handle **out-of-vocabulary (OOV)** words efficiently.
- Expands the vocabulary size to **50,257 tokens**.

(b) Positional Embeddings

- Since Transformers **do not have recurrence** (like RNNs), GPT-2 uses **positional encodings** to retain word order.
- Unlike sinusoidal embeddings used in BERT, GPT-2 **learns positional embeddings** during training.

(c) Multi-Head Self-Attention

GPT-2 uses the **self-attention mechanism** to learn contextual relationships between words.

- **Multi-Head Attention**: Instead of attending to all words equally, the model assigns different attention weights to different words in the sequence.
- GPT-2 uses **masked self-attention**, meaning it can only attend to **previous** tokens (left-to-right generation).

(d) Layer Normalization

- Instead of Batch Normalization, GPT-2 uses **Layer Normalization (LN)** after each self-attention and feed-forward layer.
- Helps stabilize training by normalizing activations.

(e) Feed-Forward Network (FFN)

- Each attention block is followed by a **fully connected feed-forward network (FFN)**.
- Consists of **two linear layers** with a **non-linearity (ReLU or GELU)** in between.

- Increases model capacity and non-linearity.

(f) Residual Connections

- Uses **skip connections (residual connections)** around attention and FFN layers.
- Helps gradient flow and speeds up convergence.

---

3. Model Hyperparameters

GPT-2 was trained in **four different model sizes**, each with a different number of layers and parameters.

Model Size	Layers	Hidden Size (d_model)	Attention Heads	Parameters
117M	12	768	12	117M
345M	24	1024	16	345M
762M	36	1280	20	762M
1.5B (GPT-2)	48	1600	25	**1.5B**

- **Increasing model size** leads to **log-linear improvements** in performance.
- The largest model, **GPT-2 (1.5B parameters)**, outperforms previous state-of-the-art models on multiple tasks.

---

4. Training Process

- **Objective:** Trains using the **causal language modeling (CLM) objective**, predicting the **next word** given previous words.
- **Dataset: WebText** (~40GB of high-quality text from the internet).
- **Optimizer: Adam with weight decay** (variant of AdamW).
- **Batch Size:** 512
- **Context Length: 1024 tokens** (longer than GPT-1's 512 tokens).

- **Training Data Cleaning:** Removed **Wikipedia** to avoid overlap with standard NLP benchmarks.

---

5. Differences from Previous Models

Feature	GPT-1	GPT-2	BERT
Architecture	Transformer Decoder	Transformer Decoder	Transformer Encoder
Training Data	BooksCorpus (5GB)	WebText (40GB)	Wikipedia + Books
Training Objective	Causal LM	Causal LM	Masked LM (MLM)
Context Window	512 tokens	**1024 tokens**	512 tokens
Fine-tuning Needed?	Yes	**No (zero-shot learning)**	Yes
Use Case	Text Generation	Unsupervised Task Learning	NLP Understanding

---

6. Strengths & Weaknesses

✅ Strengths:

- **Unsupervised multitask learning**: Learns tasks without labeled data.
- **State-of-the-art (SOTA) language modeling**: Achieves high scores in multiple benchmarks.
- **Scalability**: Larger models improve performance **log-linearly**.
- **Zero-shot generalization**: No need for fine-tuning to perform well.

❌ Weaknesses:

- **Computationally expensive**: Requires **huge GPU resources** for training and inference.
- **Prone to hallucinations**: Generates fluent but sometimes factually incorrect text.
- **Bias in training data**: Inherits biases from WebText corpus.

7. Summary

GPT-2 is a **massive, unsupervised, autoregressive Transformer** trained to predict text in a **left-to-right** manner. Its architecture scales well and demonstrates **zero-shot task transfer**, revolutionizing NLP. However, it also requires **significant computing power** and **suffers from biases** due to its training data.

Would you like a **code breakdown** of how GPT-2 processes text? 🚀