GPT Architecture (Formal Study Notes)

Definition of GPT

GPT (Generative Pre-trained Transformer) is a **decoder-only Transformer architecture** designed for **autoregressive text generation**.

Formally, GPT processes previously generated tokens and predicts the subsequent token in a sequence.

Architecture Type

- Decoder-only Transformer architecture
- Employs masked multi-head self-attention
- Operates in an autoregressive manner, generating one token at a time

High-Level Architectural Flow

Input tokens

Token Embedding + Positional Encoding

Masked Multi–Head Self–Attention

Feed Forward Neural Network

Stacked Decoder Blocks (N layers)

Linear Projection + Softmax

Next Token Prediction

Core Components of GPT

1. Token Embedding

- Maps discrete tokens to continuous vector representations
- Each token is represented using a fixed-dimensional embedding

2. Positional Encoding

- Incorporates word order information into token embeddings
- Necessary because attention mechanisms lack inherent positional awareness

3. Masked Self-Attention

- Restricts each token to attend only to preceding tokens
- Future tokens are masked during attention computation
- Ensures causality during training and inference

Illustrative Example:

```
The bank approved the
```

Predicted output token:

```
loan
```

4. Feed Forward Neural Network

- Applies position-wise non-linear transformations
- Enhances representational capacity of the model

5. Repeated Decoder Blocks

Each decoder block in GPT consists of:

- Masked multi-head self-attention
- Feed forward neural network
- Residual connections
- Layer normalization

Rationale for Masked Attention

GPT is trained to model the probability:

$$\text{next\_token} = f(\text{previous\_tokens})$$

Accordingly, when predicting a token, the model must not access future tokens. Masked attention enforces this causal constraint.

Training Objective

GPT is trained using **Causal Language Modeling (CLM)**, defined as:

$$P(w_t \mid w_1, w_2, \ldots, w_{t-1})$$

That is, the model estimates the probability of the next token given all preceding tokens.

Training Example

Input sequence:

```
The bank approved
```

Predicted token:

```
the
```

Extended input:

```
The bank approved the
```

Predicted token:

```
loan
```

Applications of GPT

- Natural language generation
- Conversational agents
- Story and content generation
- Source code generation

- Generative question answering

Limitations of GPT

- Not optimized for sentence-level embeddings
- Lacks bidirectional contextual understanding
- Requires task-specific fine-tuning for classification tasks

Concise Definition for Exams or Interviews

> *GPT is a decoder-only Transformer architecture that generates text autoregressively using masked self-attention.*

Memory Aid

> **GPT processes tokens from left to right and predicts subsequent tokens.**