T5 Architecture: Formal Study Notes

Definition of T5

T5 (Text-to-Text Transfer Transformer) is a **Transformer-based encoder–decoder architecture** developed to address a wide range of natural language processing tasks through a unified formulation.

The defining principle of T5 is that **all tasks are expressed as mappings from input text to output text**.

Architectural Classification

T5 belongs to the class of **encoder–decoder Transformer models** and is characterized by the following properties:

- The encoder is responsible for processing and interpreting the input text.
- The decoder is responsible for generating the output text sequence.
- The architecture incorporates self-attention and cross-attention mechanisms.

High-Level Architectural Workflow

The overall processing pipeline of T5 can be summarized as follows:

```
Input text

Encoder (semantic interpretation)

Contextual representations

Decoder (text generation)

Output text
```

Unified Text-to-Text Formulation

In T5, every task is reformulated as a text-to-text transformation problem, regardless of its original nature.

Illustrative Examples

Machine Translation

```
Input : translate English to French: The bank approved the loan
Output: La banque a approuv  le pr t
```

Text Summarization

```
Input : summarize: Transformers are very powerful models...
Output: Transformers are powerful NLP models.
```

Question Answering

```
Input : question: What is NLP? context: NLP is a field of AI...
Output: NLP is a field of artificial intelligence.
```

Core Architectural Components

Encoder Component

The encoder component performs the following functions:

- Processes the complete input sequence simultaneously.
- Employs bidirectional self-attention to capture global contextual information.
- Produces high-dimensional, context-aware representations of the input text.

Encoder Processing Flow:

```
Input tokens

Token embedding with positional encoding

Multi-head self-attention

Feed Forward Neural Network

Encoder output representations
```

Decoder Component

The decoder component is responsible for output sequence generation and exhibits the following characteristics:

- Generates output tokens in an autoregressive manner.
- Applies masked self-attention to ensure causal generation.
- Utilizes cross-attention to incorporate encoder-derived contextual information.

Decoder Processing Flow:

```
Previously generated tokens

Masked self-attention

Cross-attention over encoder outputs

Feed Forward Neural Network

Next-token prediction
```

Attention Mechanisms Employed in T5

Encoder Attention

- Bidirectional self-attention allowing each token to attend to all other tokens in the input sequence.

Decoder Attention

The decoder employs two distinct attention mechanisms:

1. **Masked Self-Attention**, which restricts attention to previously generated tokens.
2. **Cross-Attention**, which enables the decoder to attend to encoder output representations.

This dual-attention structure differentiates T5 from decoder-only architectures.

Training Objective

T5 is trained using a supervised **text-to-text learning objective** with teacher forcing.

Formally, the model learns to approximate the conditional probability distribution:

$$P(\text{output\_text} \mid \text{input\_text})$$

Advantages of the T5 Architecture

The T5 architecture offers several advantages:

- A unified framework capable of handling multiple NLP tasks.
- Simplified task adaptation through a consistent text-to-text format.
- Strong performance on structured input-to-output transformation tasks.

Application Domains

T5 is commonly applied to the following tasks:

- Machine translation
- Text summarization
- Question answering
- Text rewriting and paraphrasing
- Instruction-conditioned language tasks

Limitations of T5

Despite its strengths, T5 exhibits certain limitations:

- It is not optimized for open-ended conversational generation.
- It is less suitable for unconditional text generation.
- Long conversational contexts require architectural extensions or modifications.

Comparative Summary

Model	Architecture	Primary Objective
GPT	Decoder-only	Text generation
BERT	Encoder-only	Language understanding
T5	Encoder–Decoder	Text transformation

Concise Examination Definition

T5 is an encoder–decoder Transformer architecture that formulates all natural language processing tasks as text-to-text transformations.

Conceptual Memory Aid

GPT generates text, BERT interprets text, and T5 transforms text.