TÜBİTAK

ITU NLP NATURAL LANGUAGE PROCESSING RESEARCH GROUP

EURO^2

Lect. Tuğba Pamay Arslan

ITUNLP Research Group

AI & Data Engineering, İstanbul Technical University

ncc@ulakbim.gov.tr

Available Large Language Models

BERT

BERT: Bidirectional **E**ncoder **R**epresentations from **T**ransformers

Developer: Google AI (2018)

Architecture: Encoder-only Model

Training Objective: Pre-trained using

1. Masked Language Modeling (MLM): Randomly masks 15% of tokens in a sentence and trains the model to predict them.
2. Next Sentence Prediction (NSP): Determines whether one sentence logically follows another.

https://huggingface.co/docs/transformers/en/model_doc/bert

BERT

Strength:

- Deep bidirectional context understanding.
- Used in various semantic-level NLP tasks :

  Sentiment analysis, Question answering, Text classification.

Disadvantages:

- Limited Generative Capabilities

  As an encoder-only model, BERT is NOT designed for <u>text generation</u> tasks.

- Overfitting on Short Sequences

  Its performance may degrade for tasks requiring longer context handling.

- Relies on Static Pre-training

  Domain-specific fine-tuning is required for best performance, which can be resource-intensive.

BERT - Variants

DistilBERT:

A lighter, faster version of BERT with 40% fewer parameters.

Trained using knowledge distillation* for smaller devices.

RoBERTa (Robustly Optimized BERT):

Removes the NSP objective.

Uses a larger dataset and longer training for improved performance.

multilingual version of **RoBERTa**

XLM-RoBERTa:

Multilingual version of **RoBERTa**

BioBERT/ClinicalBERT:

Fine-tuned on biomedical and clinical text for domain-specific tasks.

BERTweet:

Pre-trained on social media data (tweets) to handle informal language.

BERT - Tokenization

Tokenization: Splitting words into smaller components, i.e., tokens.

- **WordPiece:**
  - A subword tokenization method used in BERT.
  - Creates a smaller vocabulary while remaining flexible and efficient
  - Enables the model to handle rare words more effectively.

- How Does It Work?
  1. **Initial Vocabulary**: (Assumes a space-separated input.) Create a vocab by all letters and some common character combinations.
  2. **Splitting into Subwords**:
     The tokenizer first checks if the word exists in the vocabulary.
     If not, it splits the word into smaller subwords.
     Example: ‘playing’ -> ‘play’ & ’##ing'

BERT - Tokenization

- How Does It Work?
  3. **Handling Rare Words**: Rare words are broken down into even smaller units
     Example: ‘unhappiness’ → ‘un’, ‘##happy’, ‘##ness’

- Advantages
  - Requires less memory due to small vocabulary
  - **Handling Rare Words**:
  - Works effectively across different languages by splitting words appropriately.

Example:

Original Text: "I am playing football."

WordPiece Tokenization:

Word-level: ["I", "am", "playing", "football"]

Subword-level: ["I", "am", "play", "##ing", "foot", "##ball"]

T5

T5: Text-**to**-**T**ext **T**ransfer **T**ransformer

Architecture: Encoder-decoder (seq2seq) model

Training Objective: Reformulates all NLP tasks into a **text-to-text** framework, unified framework.

Example:

"*Translate English to French*: The book is on the table" → "Le livre est sur la table."

Key Models: Ranges from T5-small (60M parameters) to T5-11B (11 billion parameters).

https://huggingface.co/docs/transformers/en/model_doc/t5

T5

Strength:

Unified Framework

Works across diverse NLP tasks, including translation, summarization, and even regression problems.

Disadvantages:

More complex tokenization step: SentencePiece

Latency Issues: Much higher inference time

Resource-Intensive: **Its** larger variants require substantial computational resources for fine-tuning and inference.

T5 - Variants

Original T5 Variants:

- **T5-Small**: 60M parameters.
- **T5-Base**: 220M parameters.
- **T5-Large**: 770M parameters.
- **T5-3B**: 3 billion parameters.
- **T5-11B**: 11 billion parameters (largest variant).

mT5 (Multilingual T5):

- Trained on a multilingual dataset covering 101 languages.
- Suitable for cross-lingual and multilingual tasks.

Flan-T5

- Fine-tuned on a mixture of tasks using instruction tuning.:

T5 - SentencePiece

- **SentencePiece: L**anguage-agnostic tokenizer
  - Unlike traditional tokenizers that rely on whitespace or linguistic rules-
  - Useful for languages without whitespace (e.g., Chinese, Japanese) or languages with complex tokenization rules.

- How Does It Work?

1. SubWord Segmentation:

   Statistical segmentation methods: **Byte Pair Encoding (BPE)**

   BPE: Merges the most frequent character pairs iteratively.

2. Vocabulary Creation
3. Special Tokens

   <pad>: Padding <unk>: Unknown tokens

   <s> and </s>: Sentence start and end tokens.

SentencePiece vs WordPiece

- Language Assumptions
  - WP: Assumes a space-separated input, Requires splitting text into words first.
  - SP: Operates directly on raw text without requiring spaces.

- Subword Segmentation Algorithm
  - WP: Greedy BPE; the most frequent pairs merged iteratively. Merges the most frequent character pairs.
  - SP: BPE Model; for merging frequent subwords in a statistical manner. Keeps the most probable subwords.

- Computational Efficiency
  - WP: Lower Training and Faster Inference, as it assumes pre-tokenized input.
  - SP: Higher Training, Slower Inference, due to processing raw text.

Llama

Llama: Large Language Model Meta AI

Developer: Meta AI (2023, versions: Llama, Llama 2, Llama 3).

Architecture: Decoder-only Model

Training Objective: Pre-trained on a large corpus to model language generation tasks.

Recent Variants:

LlaMA 3.1-8B (8 billion parameters)

LlaMA 3.1-70B

LlaMA 3.1-405B

…And their Instructed-Tuned versions

https://huggingface.co/meta-llama

Llama 3

- **Advantages:**
  1. **Performance Efficiency**: The smaller versions can be highly efficient for most tasks,
  2. **Reduced Resource Requirements**: Compared to models like GPT-4, it offers impressive results with fewer parameters.
  3. !!! **Open Source** !!!
  4. **Versatility:** can be fine-tuned for a wide variety of **downstream tasks** (e.g., sentiment analysis, summarization, and domain-specific applications) with minimal additional resources.

- **Disadvantages:**
  1. Training Resources and Fine-Tuning Cost
  2. Limited Task-Specific Training: It may still require additional fine-tuning for specific tasks, particularly for specialized fields like medicine, law, or finance.

THE END

Thank you ☺

My mail address: pamay@itu.edu.tr