BERT: PRE-TRAINING OF DEEP BIDIRECTIONAL TRANSFORMERS FOR LANGUAGE UNDERSTANDING

Google AI Language

Available at https://aclanthology.org/N19-1423.pdf

INTRODUCTION

Context and Motivation:

- Pre-training language models has significantly advanced the performance of many NLP tasks.

Existing Strategies:

- Feature-based approach (e.g., ELMo): Integrates pre-trained representations as additional features into task-specific architectures.
- Fine-tuning approach (e.g., OpenAI GPT): Adapts to downstream tasks with minimal task-specific parameters by fine-tuning all layers.

Limitations of Current Techniques:

- Constrained by unidirectionality: Restricts learning of contextual relationships essential for comprehensive language understanding, especially in token/sentence-level tasks such as question answering.

BERT: BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Key Innovations of BERT:

- Using a "masked language model" (MLM) pre-training objective for deep bidirectional Transformer.
- Also incorporates a "next sentence prediction" task that jointly pre-trains text-pair representations.

Contributions and Impact:

- Demonstrates bidirectional pre-training for language representations.
- Simplifies the architecture requirements for specific NLP tasks.
- Achieves state-of-the-art performance across eleven tasks.

RELATED WORK

Unsupervised Feature-based Approaches:

- Initial focus on word embeddings
- ELMo introduced context-sensitive features from bidirectional language models, significantly advancing NLP benchmarks.

Unsupervised Fine-tuning Approaches:

- Early models only pre-trained word embeddings
- Development of contextual token representations pre-trained and fine-tuned from unlabeled text for downstream tasks:

Transfer Learning from Supervised Data:

OVERVIEW OF BERT

- Handles language representations through two main phases: **pre-training** and **fine-tuning**.
- **Pre-training**
  - Trained on a large corpus of unlabeled data.
  - Employs two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).
  - Captures deep bidirectional context of language.
- **Fine-tuning**
  - Uses labeled data specific to downstream tasks.
  - Adapts BERT's architecture to achieve state-of-the-art results.
  - Applicable to a wide range of NLP challenges.

BERT MODEL ARCHITECTURE

BERT based on the original Transformer model but modified to support bidirectional context processing. Key configurations include:

- Number of layers (L) e.g., BERT BASE has 12 layers.
- Hidden size (H): size of the hidden layers. BERT BASE uses 768.
- Self-attention heads (A): BERT BASE uses 12 attention heads.
- Total Parameters: BERT BASE has 110M parameters.

Unlike OpenAI GPT unidirectional architecture, BERT's bidirectional Transformer encoder allows each token to attend to all tokens in the input sequence.

INPUT/OUTPUT REPRESENTATIONS

Designed to handle single sentences or a pair of sentences seamlessly by incorporating special tokens and embeddings:

- Special tokens: `[CLS]` for classification tasks and `[SEP]` for separating sentence pairs.
- Segment embeddings: distinguish between sentences in tasks involving comparisons.
- Positional embeddings: maintain the positional context of words.

The sum of these embeddings provides a rich representation of the input tokens, which is critical for the model to understand the language context fully.

PRE-TRAINING TASKS OF BERT

BERT is pre-trained using two unsupervised tasks:

- **Masked LM (MLM)**: Learns to predict randomly masked tokens in input based on the context provided by the non-masked tokens.
- **Next Sentence Prediction (NSP)**: Learns to predict if a sentence logically follows another.

FINE-TUNING PROCEDURE

Can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. The fine-tuning adjusts all the pre-trained parameters to make them task-specific:

- Input during fine-tuning is adjusted according to the task (e.g., sentence pairs for NLI).
- The model outputs are tailored to the specific needs of the task, using the representations learned during pre-training.

GLUE BENCHMARK

The GLUE benchmark evaluates models' language understanding across diverse NLP tasks like question answering and sentiment analysis. BERT, fine-tuned on these tasks, uses pre-trained representations with a simple classification layer for predictions. Fine-tuning BERT involves:

- Using the `[CLS]` token's final hidden state for classification.
- Adding task-specific parameters, primarily the classification layer.
- Optimizing on each task's training data with task-specific hyperparameters.

BERT's performance on GLUE exceeds previous state-of-the-art models, showcasing its transfer learning capabilities.

SQUAD V1.1

The Stanford Question Answering Dataset (SQuAD v1.1) challenges models to answer questions based on content from Wikipedia articles, where the answer to each question is a segment of text, or "span", from the corresponding reading passage.

- BERT reformulates question answering as a span prediction task.
- It predicts the start and the end of the answer span within the passage.
- The model is fine-tuned to maximize the log-probability of the correct answer span.

BERT’s fine-tuning approach allows it to outperform previous models on SQuAD v1.1, achieving new state-of-the-art results.

SQUAD V2.0

SQuAD v2.0 extends v1.1 by adding questions that do not have an answer in the provided passage, making it essential for models to determine not only the answers but also when no answer is supported by the text.

- For no-answer predictions, BERT compares the score of a null answer (based on the `[CLS]` token) to the best non-null span score.
- A threshold value (τ) is tuned on the development set to decide when to predict no answer.

This adjustment makes SQuAD v2.0 a more challenging and realistic task, which BERT handles effectively, significantly improving over prior best models.

SWAG

The Situations With Adversarial Generations (SWAG) dataset aims to evaluate a model's ability to predict the most plausible continuation of a sentence among four given options.

- BERT is fine-tuned to select the most plausible sentence continuation, employing its pre-trained contextual understanding enhanced with the SWAG dataset.
- Performance on SWAG highlights BERT's capacity for commonsense reasoning and contextual inference.

BERT's results on SWAG dramatically surpass previous approaches, underlining its robustness in handling complex language understanding tasks.

ABLATION STUDIES ON BERT

- **Pre-training Tasks:**
  - Removing NSP decreases performance on tasks needing sentence relationship understanding (e.g., QNLI, MNLI).
  - Bidirectional models (MLM) outperform unidirectional models (LTR) across tasks, especially SQuAD and MRPC.
- **Model Size:**
  - Larger models improve performance on GLUE tasks.
  - Scaling enhances feature and dependency capture, even for small-scale tasks.
- **Feature-based vs. Fine-tuning:**
  - Feature-based methods using BERT's layers are competitive, especially in NER task.
  - Fine-tuning generally outperforms.