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

Google AI Language

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

INTRODUCTION TO BERT

Context and Motivation:

- Pre-training language models has significantly advanced the performance of many NLP tasks.
- Typical tasks improved by language model pre-training include natural language inference, paraphrasing, named entity recognition, and question answering.
- Pre-trained models enhance the understanding of language by analyzing relationships between sentences and producing fine-grained output at the token level.

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 pre-trained models to downstream tasks with minimal task-specific parameters by

INTRODUCING BERT: BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Key Innovations of BERT:

- Overcomes the unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task.
- Enables the pre-training of deeply bidirectional Transformers by predicting the identity of masked tokens based on their context.
- Incorporates a "next sentence prediction" task that jointly pre-trains text-pair representations.

Contributions and Impact:

- Demonstrates the critical importance of bidirectional pre-training for language representations.
- Simplifies the architecture requirements for specific NLP tasks by reducing the need for heavily-engineered, task-specific models.

RELATED WORK ON LANGUAGE MODEL PRE-TRAINING

Unsupervised Feature-based Approaches:

- Initial focus on word embeddings using non-neural and neural methods (Brown et al., 1992; Mikolov et al., 2013).
- Evolution to sentence and paragraph embeddings with various training objectives, including:
    - Ranking candidate next sentences.
    - Generating next sentence words from sentence representations.
    - Denoising auto-encoder objectives.
- ELMo introduced context-sensitive features from bidirectional language models, significantly advancing NLP benchmarks.

Unsupervised Fine-tuning Approaches:

- Early models only pre-trained word embeddings (Collobert and Weston, 2008).
- Development of contextual token representations pre-trained and

OVERVIEW OF BERT’S FRAMEWORK

BERT (Bidirectional Encoder Representations from Transformers) introduces a novel approach in handling language representations by employing two main phases: **pre-training** and **fine-tuning**. During pre-training, BERT is trained on a large corpus of unlabeled data with two innovative tasks designed to capture the deep bidirectional context of language. Once pre-trained, BERT models are fine-tuned with labeled data specific to downstream tasks, adapting its versatile architecture to achieve state-of-the-art results across a wide range of NLP challenges.

BERT MODEL ARCHITECTURE

BERT utilizes a multi-layer bidirectional Transformer encoder architecture, which is based on the original Transformer model but modified to support bidirectional context processing, critical for understanding the full scope of language semantics. Key configurations include:

- Number of layers (L): different versions have varying depths 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 which uses a unidirectional architecture, BERT’s bidirectional Transformer encoder allows each token to attend to all tokens in the input sequence, enhancing its context understanding.

INPUT/OUTPUT REPRESENTATIONS

BERT’s input representation is 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)**: Random tokens are masked in the input, and the model learns to predict them based on the context provided by the non-masked tokens.
- **Next Sentence Prediction (NSP)**: BERT learns to predict if a sentence logically follows another, which is crucial for tasks that require understanding the relationship between sentences.

These pre-training tasks help BERT develop a profound understanding of language structure and context before any task-specific fine-tuning.

FINE-TUNING PROCEDURE

Post pre-training, BERT can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference. 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.

Fine-tuning allows BERT to adapt to any specific task with minimal changes to the underlying model, demonstrating its versatility and power.

GLUE BENCHMARK

The General Language Understanding Evaluation (GLUE) benchmark is a collection of diverse NLP tasks designed to evaluate and promote models' general language understanding capabilities. These tasks include question answering, sentiment analysis, and textual entailment among others. BERT is fine-tuned on each of these tasks using the representations learned during pre-training, with only a simple classification layer added for predictions.

FINE-TUNING BERT ON GLUE

Fine-tuning BERT on the GLUE tasks involves:

- Using the `[CLS]` token's final hidden state as the input for classification.
- Introducing a small number of new parameters specific to each task, mainly the classification layer weights.
- Optimizing the model on each task's training data, with hyperparameters like learning rate selected based on the development set performance.

BERT's performance on GLUE significantly surpasses previous state-of-the-art models, demonstrating its effective 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

- **Effect of Pre-training Tasks:**
  - Removing NSP significantly decreases performance on tasks requiring understanding of sentence relationships (e.g., QNLI, MNLI).
  - Bidirectional models (MLM without NSP) perform better than unidirectional models (LTR & No NSP) across all tasks, particularly on tasks like SQuAD and MRPC.
- **Effect of Model Size:**
  - Larger models consistently improve task performance across various GLUE tasks.
  - Model scaling shows substantial benefits even on small-scale tasks, enhancing the capacity to capture complex features and dependencies.
- **Feature-based vs. Fine-tuning Approach:**