Review of Devlin et al. (2018)

Matthew Evans

April 5, 2025

Overview

Devlin et al. (2018) [1] address the challenge of learning rich language representations that fully leverage both left and right context. Existing pre-training methods either read text in a single direction or fuse unidirectional models in a superficial way, limiting their ability to capture nuanced dependencies. To solve this, the paper introduces BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers), which uses a simple "mask-and-predict" approach alongside a sentence-ordering task to pre-train deep, bidirectional Transformer encoders. These representations can then be fine-tuned with minimal task-specific architecture changes, yielding strong gains across diverse language understanding problems.

Approach

Prior Work

Prior to BERT, researchers pursued two main strategies for leveraging large unlabeled text: *Unsupervised feature-based approaches*, which extract contextual embeddings to feed into task-specific models, and *Unsupervised fine-tuning approaches*, which pre-train a language model and then tune all its parameters end-to-end on a downstream task.

Unsupervised feature-based approaches

- **ELMo (Peters et al., 2018a)**[2]: Builds context-sensitive word embeddings by concatenating independently trained left-to-right and right-to-left LSTM language models, and supplies these as additional features to downstream architectures.
- **Melamud et al. (2016)**[3]: Propose a "cloze"[1] pre-training task using LSTMs to predict a missing word from both left and right context; like ELMo, it remains feature-based rather than deeply bidirectional.
- **Fedus et al. (2018)**[4]: Show that cloze-style masking can improve the robustness of text-generation models by injecting noise during pre-training.

Unsupervised fine-tuning approaches

- **Collobert & Weston (2008)**[5]: Early work pre-training word-embedding parameters on large corpora, demonstrating that fixed embeddings can boost a variety of NLP tasks.
- **Dai & Le (2015); Howard & Ruder (2018)**[6]: Extend fine-tuning to sentence- and document-level encoders, showing that pre-training on unlabeled text followed by supervised fine-tuning yields strong performance with minimal task-specific parameters.
- **OpenAI GPT (Radford et al., 2018)**[7]: Trains a left-to-right Transformer language model on large text corpora, then fine-tunes it end-to-end for downstream tasks—achieving state-of-the-art results on several sentence-level benchmarks.

[1] A cloze task is a fill-in-the-blank exercise where selected words in a text are hidden and the model must predict the missing tokens using the surrounding context.

Novelty

BERT overcomes the unidirectionality constraint of prior pre-trained models by introducing a simple masking strategy and a sentence-ordering task that allow every Transformer layer to condition on both left and right context. Whereas models like GPT are strictly left-to-right and ELMo concatenates separate directional LMs only at the top of the architectural stack, BERT's *Masked Language Model* (MLM) and *Next Sentence Prediction* (NSP) objectives jointly pre-train a deep bidirectional encoder without architectural modifications.

During pre-training, BERT randomly masks 15% of tokens and trains the Transformer to predict the original tokens from their full context, while simultaneously classifying whether one sentence follows another. Because the same Transformer encoder is used in both pre-training and fine-tuning, downstream tasks require only a small task-specific output layer, and all parameters can be fine-tuned end-to-end with minimal model architecture engineering.

Deep Bidirectional Pre-training via Masked Language Modeling

BERT's primary innovation is its MLM objective, which randomly replaces 15% of input tokens with a special `[MASK]` token (80% of the time), a random token (10%), or leaves them unchanged (10%), then predicts the original tokens via cross-entropy loss. By masking tokens rather than shifting a unidirectional window, every layer combines left and right context, enabling truly bidirectional representations that improve both token and sentence-level tasks.

Next Sentence Prediction for Text-Pair Representations

To capture inter-sentence coherence, BERT adds an NSP task built on a special input format. Each example is constructed as:

$$[\texttt{CLS}]\ \text{Sentence A}\ [\texttt{SEP}]\ \text{Sentence B}\ [\texttt{SEP}]$$

where `[CLS]` is a special classification token added at the front of every input example and `[SEP]` tokens mark the boundary between segments. The final hidden state corresponding to `[CLS]`, often denoted $\mathbf{C}$, serves as an "aggregate" representation of the entire sequence. A binary classifier on $\mathbf{C}$ is trained to predict whether Sentence B actually follows Sentence A (labelled `IsNext`) or is a random sentence (labelled `NotNext`). This pre-training objective directly conditions the model to understand sentence-pair relationships, improving performance on tasks like *Question Answering* and *Natural Language Inference*.

Unified Fine-Tuning Paradigm Reducing Task-Specific Engineering

Unlike feature-based methods that require separate architectures per task, BERT demonstrates that fine-tuning a single pre-trained model with just one added output layer suffices to achieve strong results across a wide variety of tasks—from classification to span selection—dramatically reducing the need for heavily engineered, task-specific models.

Considerations

Strengths

- **Truly bidirectional context**: By masking tokens rather than using a unidirectional window, BERT learns from both left and right context at every layer.
- **Unified pre-training objectives**: Combines *Masked Language Modeling* and *Next Sentence Prediction* in a single model, boosting both token and sentence-level understanding.
- **Minimal task-specific engineering**: Supports a wide range of downstream tasks with only one added output layer and end-to-end fine-tuning.
- **Flexibility across tasks**: Effective for classification, tagging, span selection, and sentence-pair tasks without architecture changes.

Weaknesses

- **Mask-fine-tune discrepancy**: The special `[MASK]` token used during pre-training never appears at fine-tuning time, creating a representational mismatch that must be mitigated via random and same-token replacement strategies.
- **Simplistic sentence-pair objective**: Next Sentence Prediction is a binary, randomly sampled task that may not fully capture nuanced discourse or complex inter-sentence relationships.
- **No free-text generation**: As an encoder-only, bidirectional model, BERT cannot perform autoregressive next-token generation or summarization without significant architectural changes.
- **Homogenous training corpus**: Pre-trained solely on BooksCorpus and Wikipedia, BERT's representations might degrade on out-of-scope corpora-specific tasks.
- **Fine-tuning sensitivity**: Although fine-tuning is relatively fast, small datasets are highly sensitive to hyperparameter choices—necessitating extensive tuning to reach peak performance.

Measures of Success

- **GLUE**[2]: $\text{BERT}_{\text{BASE}}$ achieves an average score of 79.6 and $\text{BERT}_{\text{LARGE}}$ 82.1 on the benchmark—surpassing previous state of the art by 4.5 % and 7.0 % respectively.
- **SQuAD**[3]: On the SQuAD v1.1 reading-comprehension tasks, $\text{BERT}_{\text{LARGE}}$ attains 90.9 F1 on dev, beating the next best by +1.3 F1; even greater improvements were seen on enhanced models (e.g., $\text{BERT}_{\text{LARGE}}$ ensemble). For SQuAD v2.0, $\text{BERT}_{\text{LARGE}}$ achieves 83.1, a +5.1 F1 improvement over the previous best.

Impact

Since its introduction, BERT has served as the foundation for a vibrant ecosystem of improved pre-trained language models. RoBERTa[8] showed that tuning hyperparameters, removing Next Sentence Prediction, and training on more data can yield further gains. ALBERT[9] introduced parameter-sharing and factorized embeddings to reduce memory usage and accelerate training while maintaining or improving accuracy. DistilBERT[10] applied knowledge distillation to compress BERT into a smaller, faster model retaining most of its performance. Subsequent innovations like ELECTRA[11] replaced masked language modeling with a more sample-efficient replaced token detection objective, inspiring research into alternative pre-training tasks. Collectively, these works underscore BERT's pivotal role in shaping modern approaches to contextual language representation.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018.

[3] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional LSTM. In Stefan Riezler and Yoav Goldberg, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] William Fedus, Ian Goodfellow, and Andrew M. Dai. Maskgan: Better text generation via filling in the ..., 2018.

[2]The GLUE benchmark (General Language Understanding Evaluation) is a collection of diverse natural language understanding tasks designed to evaluate and compare the performance of language models across broad NLU capabilities.

[3]The SQuAD benchmark is a reading-comprehension dataset of crowd-sourced question-answer pairs on Wikipedia passages, where models must predict the exact answer span in the text.

[5] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery.

[6] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[7] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[9] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[11] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.