Summary of the paper

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

Link:- https://arxiv.org/pdf/1810.04805

BERT:-

BERT introduced bidirectional training of transformers, meaning it reads text in both directions (left-to-right and right-to-left) simultaneously during pre-training. This was revolutionary because previous models like GPT only read left-to-right.

Model Architecture:-

BERT is based on the Transformer architecture and comes in two sizes:

- **BERT Base:-** 12 layers, 768 hidden units, 12 attention heads, 110 million parameters.
- **BERT Large:-** 24 layers, 1024 hidden units, 16 attention heads, 340 million parameters.

Training Approach

- **Pre-training:** The pre-training procedure largely follows the existing literature on language model pre-training. For the pre-training corpus we use the Books Corpus (800M words) and English Wikipedia
- **Fine-tuning:** is straightforward since the self-attention mechanism in the Transformer allows BERT to model many downstream tasks whether they involve single text or text pairs—by swapping out the appropriate inputs and outputs. For applications involving text pairs, a common pattern is to independently encode text pairs before applying bidirectional cross attention.

Pre-training Tasks:-

- **Masked Language Modeling (MLM):-**

  Randomly masks some tokens in the input and trains the model to predict them, enabling deep bidirectional understanding.

- **Next Sentence Prediction (NSP):-**

  Trains the model to understand the relationship between two sentences, which is useful for tasks like Question Answering and Natural Language Inference.