**五、RoBERTa: A Robustly Optimized BERT Pretraining Approach**

* RoBERTa之主要修改

將 BERT 的訓練過程做部分修改，包含增大 batch size，使用動態遮罩 (Dynamic Masking)，與應用更多語料進行訓練。

修改 BERT 之主因在於其自身訓練不足，因此 RoBERTa 使用多種不同的最佳化方法改善預訓練性能。

* 靜態遮罩與動態遮罩

原始 BERT 於資料預處理 (Data Preprocessing) 即使用遮罩。  
然而，RoBERTa 為了避免不同 epochs 使用到相同遮罩，提出 10 種不同的遮罩，共訓練 40 epochs，每筆資料用到 4 次相同遮罩，以於訓練過程增加更多變異。

* 動態遮罩

動態遮罩於模型訓練前動態生成遮罩模式 (Masking Pattern)，實驗顯示使用動態遮罩的訓練效能比靜態遮罩更好。

* 更大 batches

BERT 原始論文中使用的 batch size 為 256，RoBERTa 使用2K

及 8K減少運算的 steps。此外，更大之 batch-size 更容易進行平

行化運算。實驗過程發現，提高 batch size 至 8K能減低Masked

Language Model 的困惑度 (Perplexity)。