**二、 Bidirectional Encoder Representations from Transformers (BERT)**

* 預訓練 (Pre-Training) 雙向 Transformer

傳統自回歸語言模型(Autoencoder)因數學定義為單向且LSTM(Long Short-Term Memory) 只能完成淺層訓練並擷取特徵，導致對於不同位置方向的單詞而言，在編碼過程看不到反向的單詞預測與理解模式，無法滿足語言處理所需的雙向共同編碼。

在訓練模式中，句子中有些單詞會依賴鄰近左右側的單詞，故僅僅從單方向編碼無法滿足上下文雙向訓練需求。  
 使用 Transformer進行自然語言處理任務與RNN不同之處在於能將網路做得更深，不同位置的詞能不受位置距離和方向因素而進行編碼。然而，在語言生成組合方面，縱使BERT做詞嵌入時有加入位置編碼 (Position Encoding) ，其原理是被用來與輸入嵌入求平均，因此語言組合也涉及詞序推理，並非僅需注意力機制，因此 “(Attention isn't all you need.) ”。

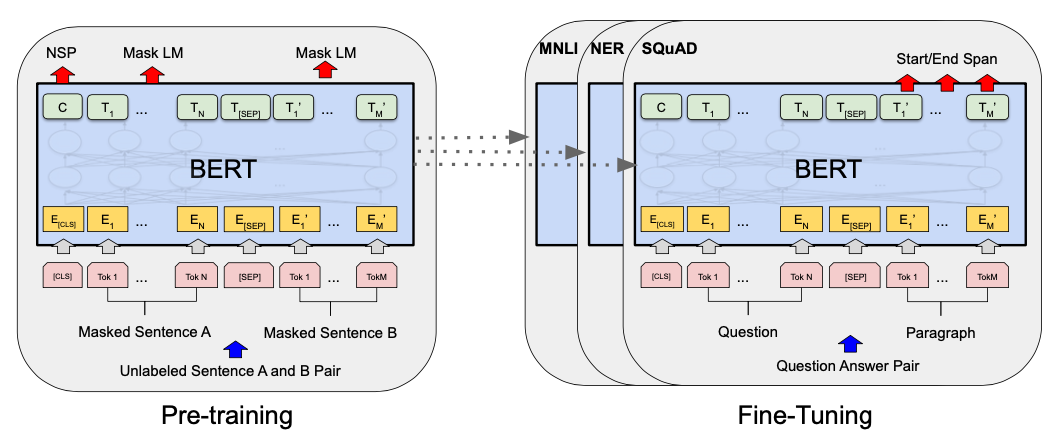


圖 4,BERT預訓練及微調結構圖, 來源: BERT 論文

* Pre-training Method＃1 Masked Language Model (Masked LM)  
   預訓練即將大量未標註之數據進行無監督學習，使其學習語法結構、解讀語義，以完成後續有效之自然語言任務。

Masked LM在預訓練前對訓練集中的文本序列隨機遮蔽 15% 的單詞，而非如以往將每個詞都預測一次。最後，損失函數只計算被遮蔽之詞塊 (token)，其中被遮蔽的15%中有10%被替換成其他單詞，另外 10% 不替換，剩餘 80% 被替換為 [MASK] 。

預訓練過程中，模型將猜測 15% 中的所有詞塊，對每個詞均計算損失與注意力機制，易造成當 [MASK] 出現過多，如此將影響模型收斂速度，甚至導致比傳統RNN (LSTM)模型從左至右的模型更慢。

* Pre-training Method ＃２: Next Sentence Prediction (NSP)

判斷第二個語句在原始文本中是否為第一句子之後續語句， RoBERTa模型將其視為無效之訓練方法。

* Transfer Learning on NLP

以語言模型預訓練方法訓練出對自然語言有相當程度理解之語言模型，將其用以做特徵擷取並針對下游任務進行微調模型。BERT能同時完成無監督學習和監督式微調。