大型語言模型 (LLM) 背後的建模技術

自從 ChatGPT 等大型語言模型 (LLM) 爆紅之後，自然語言處理 (NLP) 一直是深度學習的熱門研究話題。GPT (Generative Pre-trained Transformer) 的名號應該不用多說了，它隨著 ChatGPT 而聲名大噪，ChatGPT 背後的技術就是 GPT。企業開發人員想奠定對大型語言模型的建模基礎，就不得不好好認識 GPT 這個熱門模型的底層知識。

文／旗標科技

認識 GPT 模型

如同其名「Generative (生成式)」，GPT 除了能照其預訓練目標完成 Auto-Complete 字詞預測外，其憑空生成文字的能力也非常出色。

要訓練出大規模的高效能模型，一般作法是先以大型資料集預訓練 (pretrain) 模型、熟稔資料基本特徵與結構後，再轉用於不同類型任務。有些模型可單憑之前累積的知識，輕鬆應對新工作，有些可能得加上幾層神經層、再以任務資料集微調 (fine-tune) 參數才能應付該任務，這些都算是遷移學習 (transfer learning) 的技巧。而 GPT 就是應用遷移學習的精神，先設計一些任務來預訓練模型，以綜合各層能力應對多種任務。

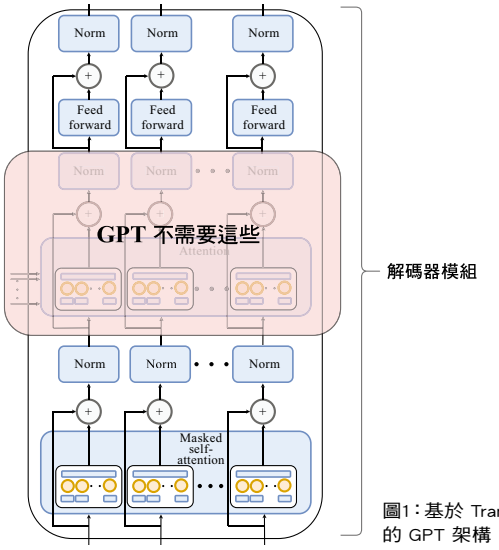

圖1：基於 Transformer 的 GPT 架構

GPT 的架構

GPT 的架構是由 Transformer 這個 seq2seq 模型的「解碼器」模組稍作修改、再堆疊而成的獨立語言模型，圖 1 是聚焦解碼器部分的架構。

圖 2 則是 GPT 整體架構，由多個解碼器重覆運算組合而成。

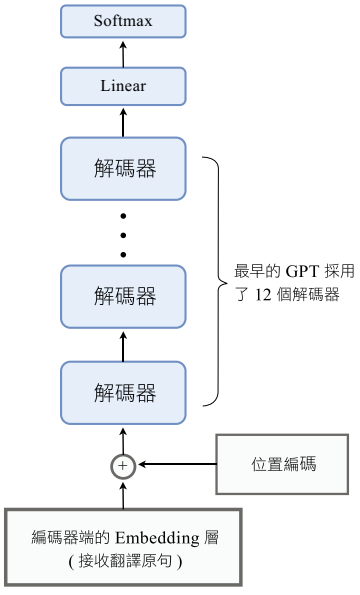

圖2：GPT 整體架構

GPT 的預訓練 (pre-train) 及微調 (fine-tune) 做法

GPT 模型的預訓練做法如圖 3 所示，是採用一般語言模型任務：

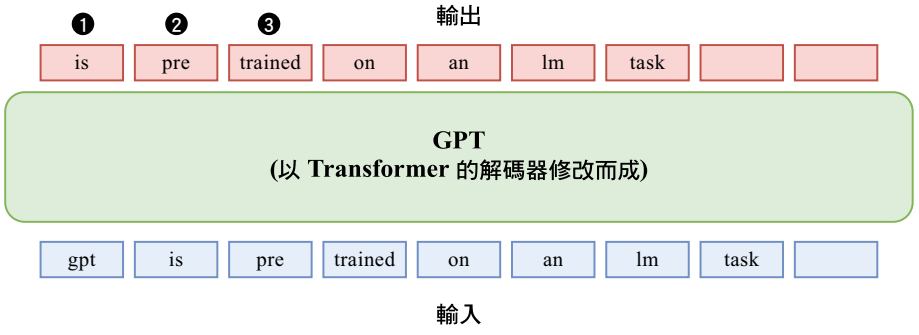

圖3：GPT 的預訓練示意圖

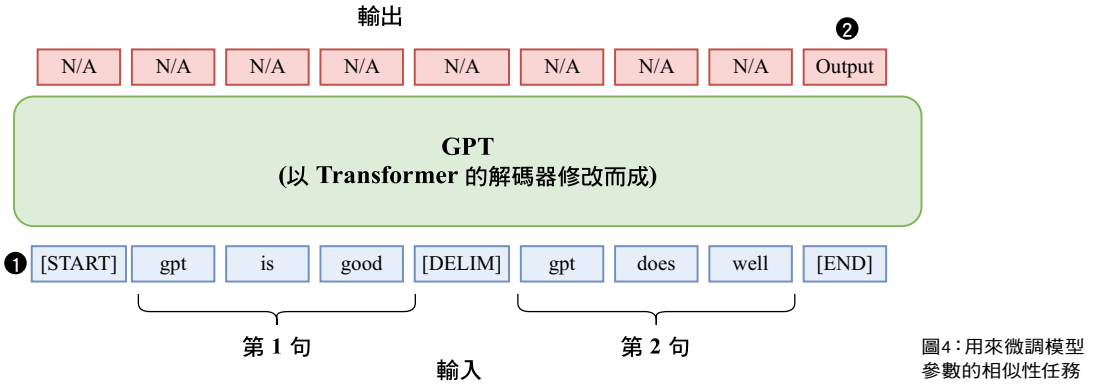

圖4：用來微調模型參數的相似性任務

圖 3 以「gpt is pre trained on an lm task」為例句示範。GPT 模型要學會預測的正解跟輸入是同一句，也就是用 gpt 預測 is、用 gpt is 預測 pre、用 gpt is pre 預測 trained … 依此類推。而在 GPT 架構圖應該有注意到，GPT 的 self-attention 機制保留了屏蔽 (masked) 功能，意思是當用 gpt 預測 is 時，後面的「pre trained on an lm task」是要遮起來的，以避免較早生成的單字去關注後續單字，讓模型提早關注到正解而作弊。

預訓練結束後，接著會再用監督式訓練進一步微調 (fine-tune) 模型參數。此階段要用標記過的資料搭配指定任務訓練模型，故輸入、輸出資料格式就得針對任務量身修改，這些任務包括分類 (classification)、文章相似性(similarity) 分析、自然語言推理 (textual entailment)… 等等。例如圖 4 展示的是文章相似性 (similarity) 分析任務，模型要學習判斷任意兩句子是否相似。為方便任務進行，兩句子會先串接、中間以分隔標記 (DELIM) 隔開 (註：Delimit, 劃界的意思)，並在開頭與結尾各加上 START、END 標記，以此輸入模型。

模型接收完輸入，便會在對應的 END 這個時步輸出判斷結果。除了得調整輸入輸出資料格式，亦得修改輸出層，初代 GPT 論文 (Radford et al., 2018) 描述了幾種因應不同類型任務的修改方式。而對相似性任務來說，兩句誰先誰後都沒差，故訓練時可將兩個句子以相反順序分別輸入 GPT，再將兩邊輸出相加、送至獨立線性分類器，使之學習判斷兩句的相似性。

《**跟 NVIDIA 學深度學習！**》一書涵蓋穩機器視覺與大型語言模型 (LLM) 的重要建模知識，從基本神經網路到 CNN・RNN・LSTM・seq2seq・Transformer・GPT，是企業開發人員、資料科學家、分析師等人的理想之選。