1. **Transformer**

Transformer 是一種基於自注意力機制的 Seq2Seq(Sequence to Sequence)模型，用於提高神經網路機器翻譯的性能及模型訓練速度，在特定任務中的表現優於傳統的機器翻譯模型。

* 自注意力機制 (Self-Attention)

Transformer編碼器的輸入首先經過自注意力層，對特定單詞進行編碼並查看輸入語句的其他單詞，並輸出至前饋神經網路。  
 解碼器包含自注意力機制、前饋層，與用以編解碼之注意力層，協助解碼器將注意力集中於輸入語句的相關部分，避免前饋層當前解碼如圖1。

  
圖１, 編碼器與解碼器架構圖, 來源: jalammer.github.io  
  
 輸入語句的每個單詞將轉換成三個向量 (query查詢向量、key鍵向量、value值向量) ，將嵌入向量對訓練過程建立的三個矩陣 (query, key, value) 進行點乘 (Dot-Production) 產生向量。

透過查詢向量之點乘與各個單詞鍵向量內積得出注意力得分，為 q1⋅kn、q2⋅kn……分別除以向量維度的平方根以獲得更穩定之梯度。

透過softmax對分數進行標準化，顯示該單詞對目標單詞的相關性。將每個值向量乘以softmax分數，並加權總和就能得到自注意力輸出，如圖2。



圖 2, 自注意力結構圖, 來源: jalammer.github.io

* 多頭注意力機制 (Multi-Head Attention)

多頭注意力為單頭注意力的原理延伸，其著重於單一查詢矩陣和多個鍵向量進行點乘並一起考慮整個輸入語句的單詞，如圖3。

頭關注的點隨著下游任務而不同，其中包含以單頭取局部的資訊或以多頭取全局的資訊。因此無論是單頭注意力機制或多頭注意力機制，在特定問題的解決方案都會被用到，如圖4。

  
圖 3, 多頭注意力結構圖, 來源: jalammer.github.io



圖4, 注意力機制比較圖, 來源: Attention Is All You Need 論文

編碼過程中與傳統 RNN 相比能降低順序輸入問題，但也因無序問題衍生出位置編碼 (Position Encoding) 的必要性。在未加入位置編碼的情況下，自注意力機制會將句子中出現重複的詞賦予一樣的權重，造成相近之詞應較重要，其權重卻與另一個較遠之詞一樣重要，產生 Transformer 弱於捕獲文本中的短期依賴問題，這種對注意力的依賴會導致 Transformer 在語法敏感任務上的性能不如 RNN 模型。