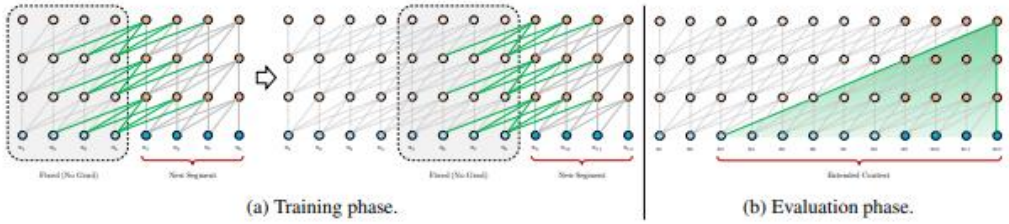
HW4

R10945001
陳欣頤

1. **Make a brief introduction about a variant of Transformer.**

   Transformer-XL (XL = extra long)
   Transformer 結構會為輸入序列的所有位置對計算相似度，然而這種方法在輸入序列的長度較長時效果不佳，對長距離序列的建模能力不足。
   Transformer-XL 提供 segment-level recurrence mechanism，引入一個記憶模塊(memory)，循環用來建構片段之間的聯繫，使得片段之間有交互關係，能夠達到長距離依賴的建模。另外也提出 relative position embedding scheme 代替絕對位置編碼，透過權重學習不同片段的位置編碼。



   因此，Transformer-XL 能夠處理更長的序列結構，也提高了運行的速度。

   Paper:
   [Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context]

2. **Briefly explain why adding convolutional layers to Transformer can boost performance.**

   Transformer 是基於 Self-attention 設計的 model，擅長捕獲全局的交互訊息，對於大範圍前後相關的特徵做訓練具有較好的效果，但較缺乏提取局部細微的特徵的能力。而 CNN 擅長提取局部細微的特徵，只考慮 receptive field 範圍裡的資訊，像是影像的一些邊緣特徵。因此將兩者結合能兼具局部與全局的資訊，得到更好的效果。