(a) Interleaved Transformer

(b) Sandwich Transformer

1. 如果一個模型中有很多層的 transformer layer，那正常的 transformer 是採用先經過 self attention layer 再經過 fc layer，所以會呈現 s f 交替的架構。但是如果是 sandwich transformer，他會先將較多層連續的 self attention layer 擺在底部 (靠近 input)，並且將較多層連續的 fc layer 擺在尾端 (靠近 output)，中間再採用 s f 交替的架構。從論文的實驗結果也可以得知前面擺 6 層 self attention layer 和後面擺 6 層 fc layer 的結果是最好。

2. Sandwich transformer 的好處是由於模型的參數和原本的 transformer 一樣，只是交換了 layer 的順序，所以 sandwich transformer 不需要額外的參數，memory，或是 runtime。在語言模型等問題上可以在不用額外 cost 的情況下就可以 improve performance.