

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.