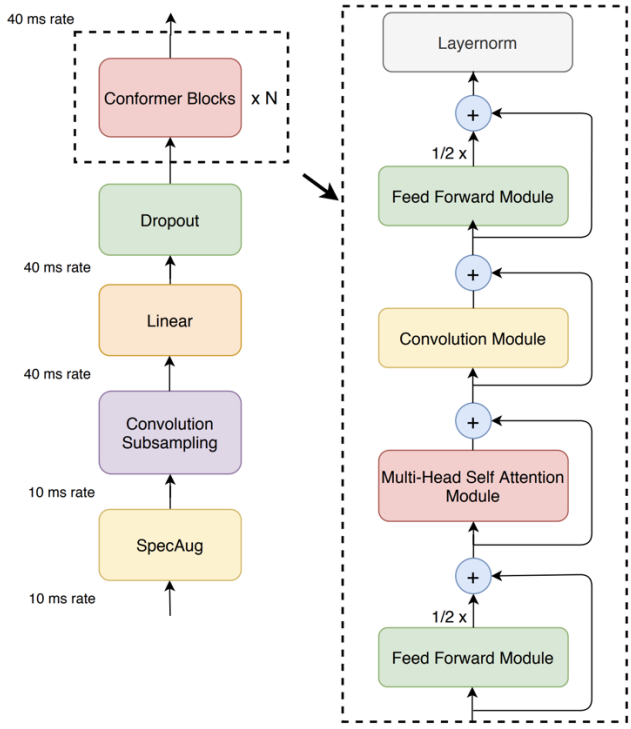
1. Make a brief introduction about a variant of Transformer.

Ans:

Transformer 是 Google 提出的一種不使用 RNN 和 CNN，而使用 self-attention 的架構，在處理 NLP 這種有前後順序、上下文關係很有效。而在一般的圖像處理中，大多使用的網路是有卷積層的 CNN 架構，利用卷積來提取局部特徵。Conformer 就是結合了原本的 Transformer 與卷積這兩種特性，應用在 NLP 上。



Conformer 主要是由 4 個 block 組成，也就是上圖右邊的 feed forward module、multi-head self attention module、convolution module 最後再接一個 feed forward module。

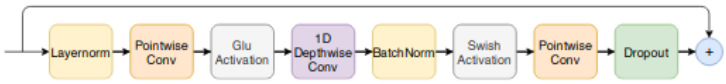


Figure 2: ***Convolution module.*** *The convolution module contains a pointwise convolution with an expansion factor of 2 projecting the number of channels with a GLU activation layer, followed by a 1-D Depthwise convolution. The 1-D depthwise conv is followed by a Batchnorm and then a swish activation layer.*

convolution module 中先對 layer 做 normaliztion，使用 pointwise convolution、GLU activation layer，一個近年發現能收斂比 Relu 快的 Activation，然後是一維的 depthwise convolution、Batchnorm，然後是 Swish activation layer。

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

Ans:
上述有提到 Transformers 很適合用於長形的 context 中，但是對於較局部的 local feature pattern 就很難提取出。而將 convolutional layers 加入到 Transformers 中便能解決這個缺點，因為 CNN 的 convolutional layers 能有效提取 local information，且使用 position-based kernels，可以維持 translation equivariance，加強整體 performance。