Assignment5

1. Yes, after padding,we have same length
2. (kernel\_size - (m\_model+2))/2
3. Reduce the gradient vanishing and automatically choose the information to pass to another layer. Should start from positive because in order to start gradient descent,x gate should start with 1 and then decrese. If start with 0, it make no sense to introduce the new layer and x proj
4. LSTM vs Transformer

Transformer can run on multiple groups parallelly which is faster and different groups can capture different information

Difference between view, reshape, transpose and permute in PyTorch

<https://jdhao.github.io/2019/07/10/pytorch_view_reshape_transpose_permute/>

