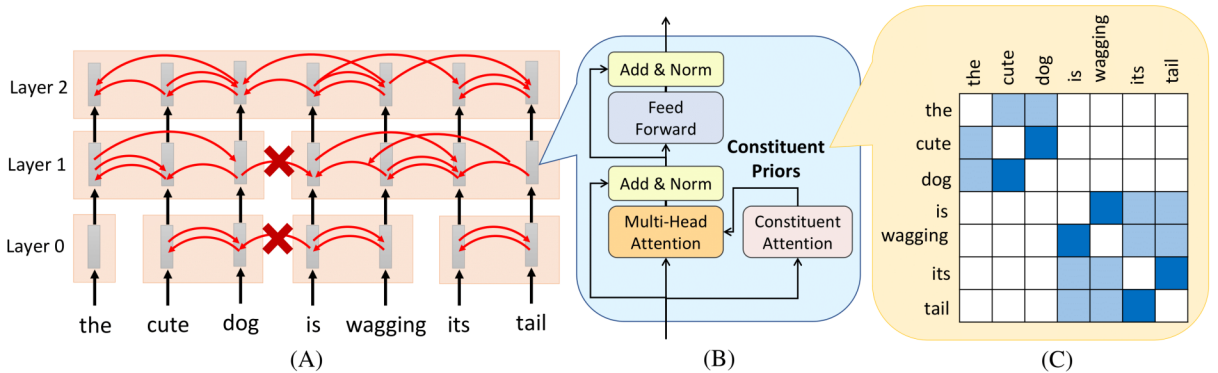

Figure 1: (A) A 3-layer Tree Transformer, where the blocks are constituents induced from the input sentence. The two neighboring constituents may merge together in the next layer, so the sizes of constituents gradually grow from layer to layer. The red arrows indicate the self-attention. (B) The building blocks of Tree Transformer. (C) Constituent prior C for the layer 1.