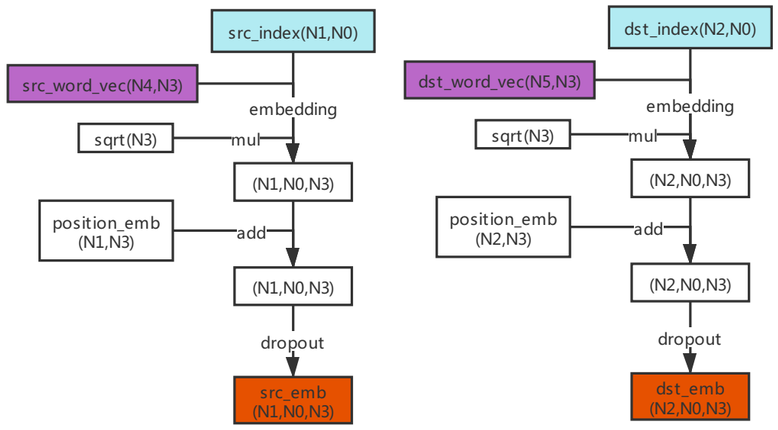
transformerN0: batch size
N1: src length
N2: dst length
N3: embedding size
N4: src vocab size
N5: dst vocab size



mutli-head-attentionN0:
batch size
N1: query length
N2: key length
N3: embedding size
N4: num_head
N5=N3/N4: head dim