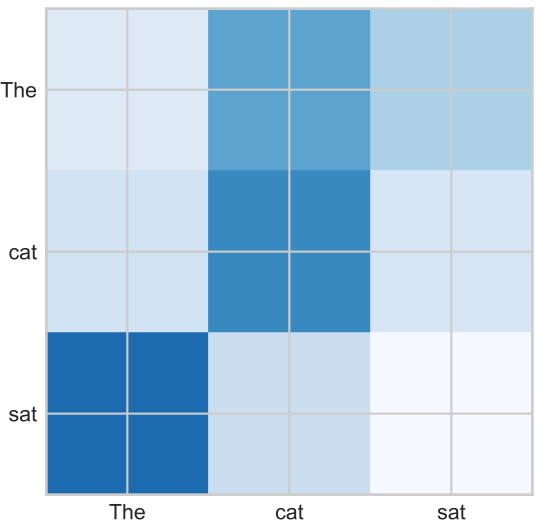
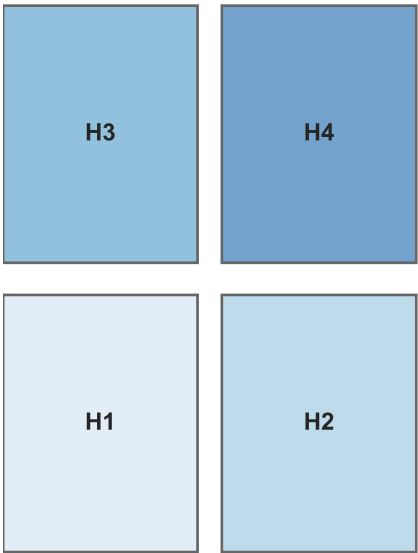


Building a Transformer: From Single Head to GPT Scale

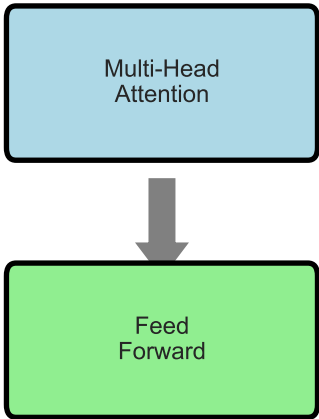
Step 1: Single Attention Head



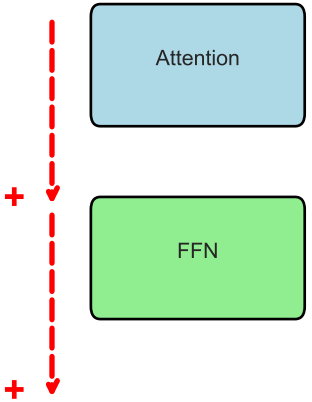
Step 2: Multi-Head (4 heads)



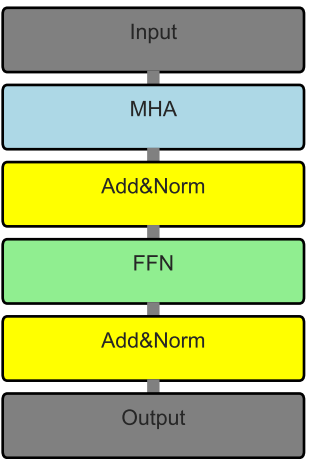
Step 3: Add Feed Forward



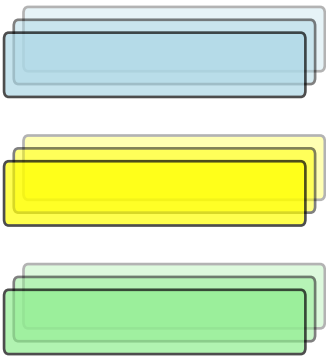
Step 4: Add Residuals



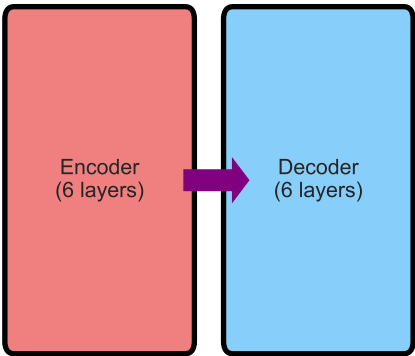
Step 5: One Complete Layer



Step 6: Stack 6 Layers



Step 7: Full Enc-Dec



Step 8: Scale to Billions

