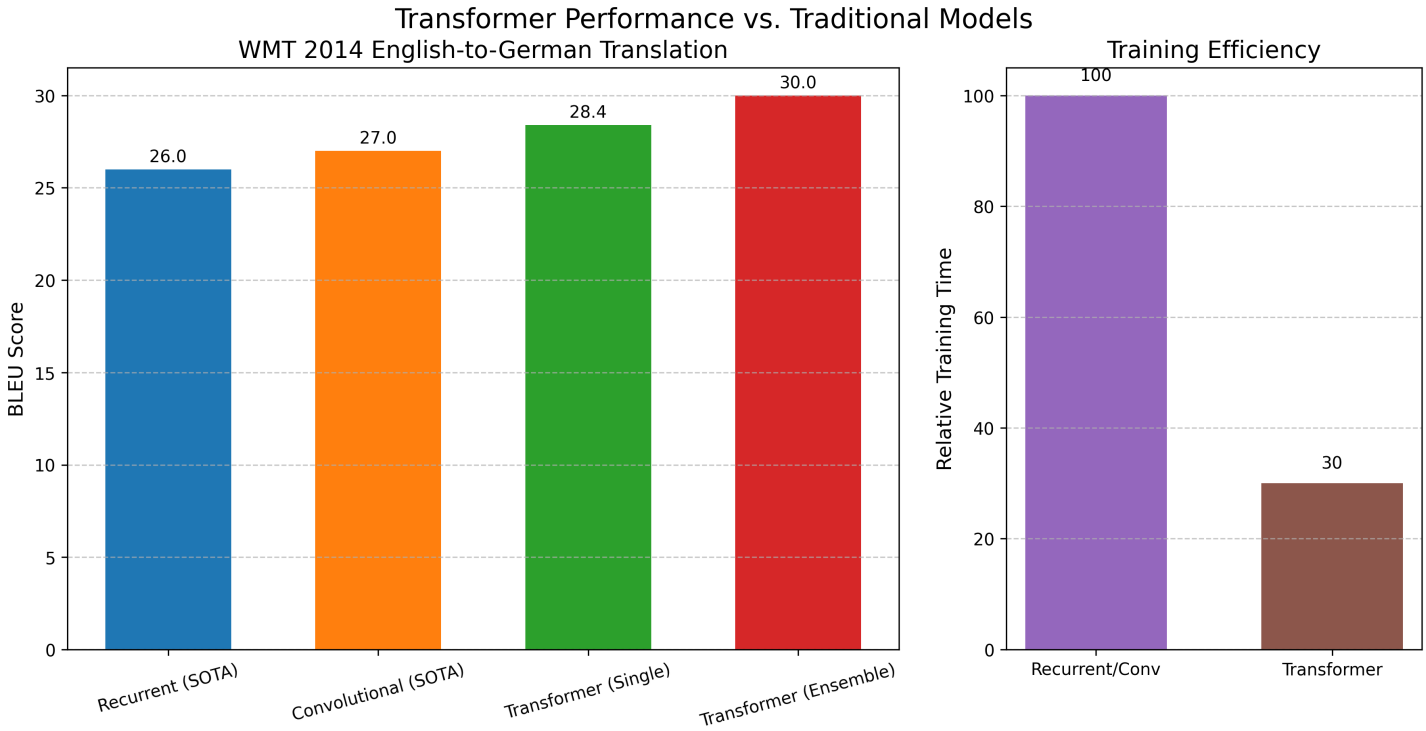
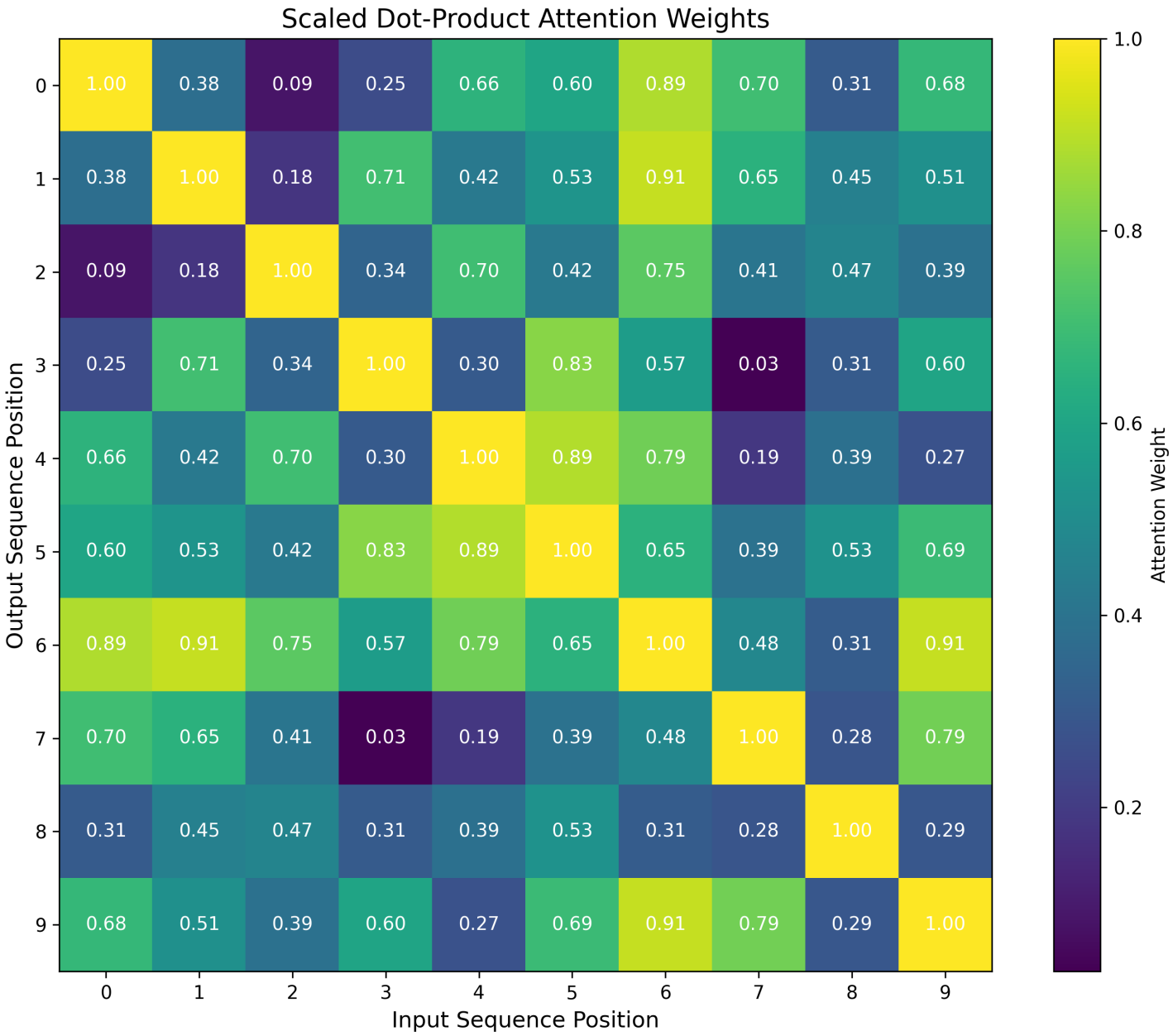
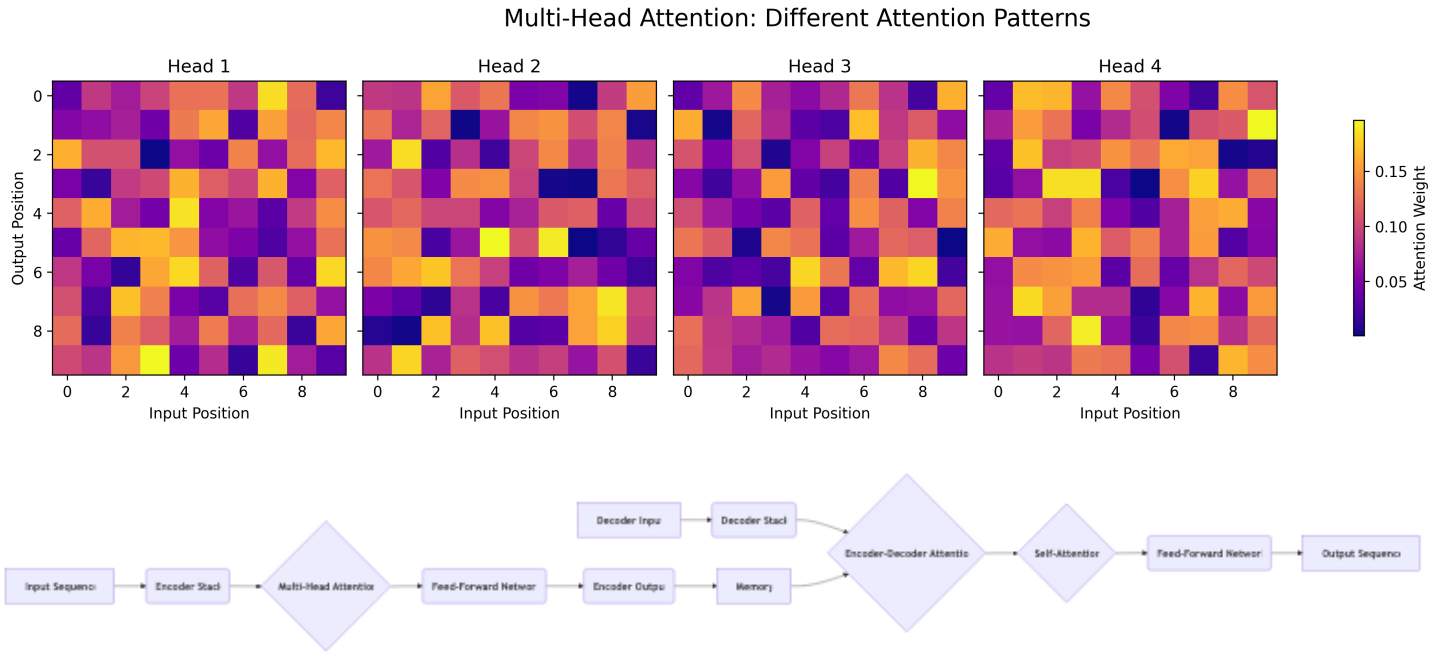

Transformer Performance vs. Traditional Models
WMT 2014 English-to-German Translation
BLEU Score
26.0
27.0
28.4
30.0
Recurrent (SOTA)
Convolutional (SOTA)
Transformer (Single)
Transformer (Ensemble)
Training Efficiency
Relative Training Time
100
30
Recurrent/Conv
Transformer


Scaled Dot-Product Attention Weights
Output Sequence Position
Input Sequence Position
Attention Weight
1.00 0.38 0.09 0.25 0.66 0.60 0.89 0.70 0.31 0.68
0.38 1.00 0.18 0.71 0.42 0.53 0.91 0.65 0.45 0.51
0.09 0.18 1.00 0.34 0.70 0.42 0.75 0.41 0.47 0.39
0.25 0.71 0.34 1.00 0.30 0.83 0.57 0.03 0.31 0.60
0.66 0.42 0.70 0.30 1.00 0.89 0.79 0.19 0.39 0.27
0.60 0.53 0.42 0.83 0.89 1.00 0.65 0.39 0.53 0.69
0.89 0.91 0.75 0.57 0.79 0.65 1.00 0.48 0.31 0.91
0.70 0.65 0.41 0.03 0.19 0.39 0.48 1.00 0.28 0.79
0.31 0.45 0.47 0.31 0.39 0.53 0.31 0.28 1.00 0.29
0.68 0.51 0.39 0.60 0.27 0.69 0.91 0.79 0.29 1.00


Multi-Head Attention: Different Attention Patterns
Head 1
Head 2
Head 3
Head 4
Output Position
Input Position
Attention Weight
0
2
4
6
8
0.15
0.10
0.05

Input Sequenc
Encoder Stacl
Multi-Head Attentior
Feed-Forward Networ
Encoder Outpu
Decoder Inpu
Decoder Stacl
Memory
Encoder-Decoder Attentio
Self-Attentior
Feed-Forward Networ
Output Sequenci