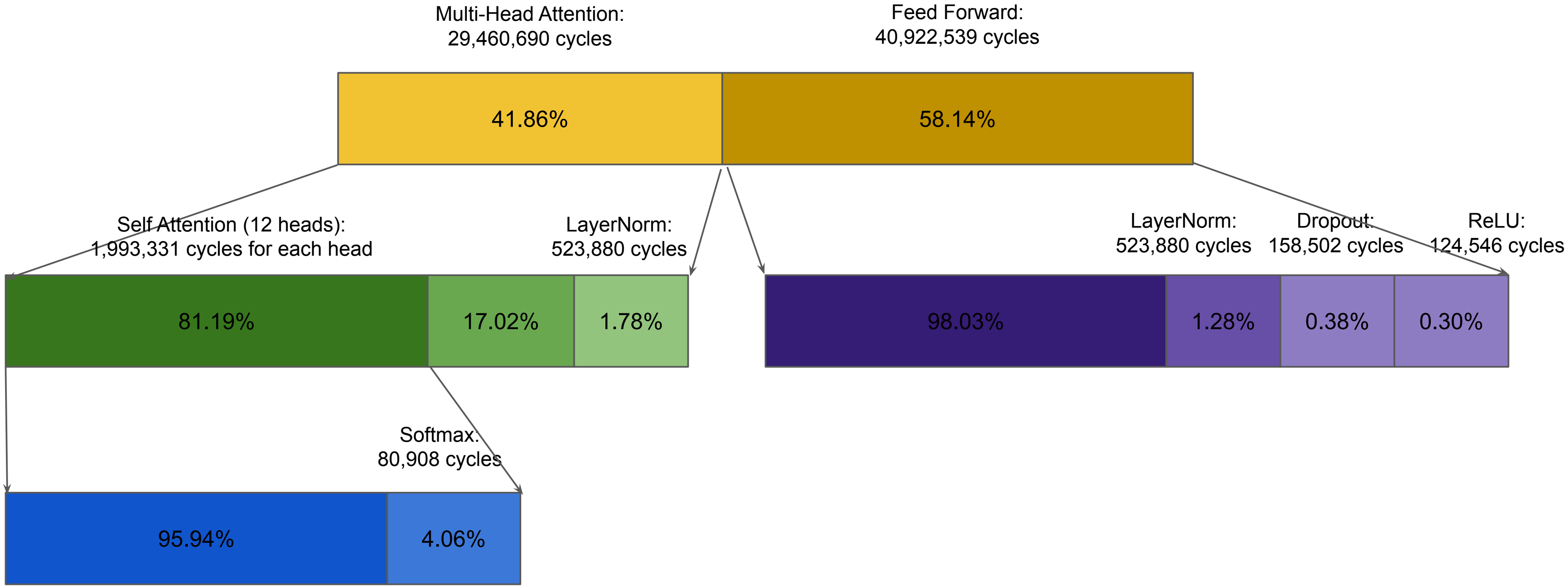
Week 9

Transformer Performance



Multi-Head Attention:
29,460,690 cycles

Feed Forward:
40,922,539 cycles

Self Attention (12 heads):
1,993,331 cycles for each head

LayerNorm:
523,880 cycles

LayerNorm:
523,880 cycles

Dropout:
158,502 cycles

ReLU:
124,546 cycles

Softmax:
80,908 cycles

SP-FLOP & Utilization

kernel	execution cycles	utilization	performance (SP-FLOP/cycle)
layernorm	523880	9.39%	0.844651
softmax	80908	44.30%	4.606912
self-attention	1993331	64.34%	10.1879
multihead-attention	29460690	68.47%	10.8529
feed_forward	40922539	92.56%	14.78552
dropout	158502	15.51%	1.240413
relu	124546	19.73%	1.578597

New MatMul

Example: A * B + bias + C, A: D1xD2, B: D2xD3, bias:D3x1, C: D1xD3

- In the innermost loop, if it is the last calculation, the bias and C are loaded while the MAC is executed

kernel	execution cycles	utilization	performance (SP-FLOP/cycle)
MatMul	598694	65.68%	10.508634
MatMul-biased	608641	64.69%	10.343621
MatMul-biased-MatAdd	610002	64.63%	10.327258

64x768x64, lower performance than 64x64x64 (~15 FLOP/Cycle, 80%)