input-tensor depth:0 | (64, 10)

Stem

Norm depth:2 | input: (64, 10) | output: (64, 10)

reshape depth:2 | input: (64, 10) | output: (64, 10, 1)

Linear depth:2 | input: (64, 10, 1) | output: (64, 10, 32)

GELU depth:2 | input: (64, 10, 32) | output: (64, 10, 32)

output-tensor depth:0 | (64, 10, 32)

TransformerBlock

MultiHeadedAttention

AttentionHead

Linear depth:4 | input: (64, 10, 32) | output: (64, 10, 4)

einsum depth:4 | input: 2 x (64, 10, 4) | output: (64, 10, 10)

div depth:4 | input: (64, 10, 10) | output: (64, 10, 10)

softmax depth:4 | input: (64, 10, 10) | output: (64, 10, 10)

einsum depth:4 | input: (64, 10, 10), (64, 10, 4) | output: (64, 10, 4)

stack depth:3 | input: 8 x (64, 10, 10) | output: (64, 8, 10, 10)

output-tensor depth:0 | (64, 8, 10, 10)

stack depth:3 | input: 8 x (64, 10, 32) | output: (8, 64, 10, 32)

sum depth:3 | input: (8, 64, 10, 32) | output: (64, 10, 32)

add depth:2 | input: 2 x (64, 10, 32) | output: (64, 10, 32)

Sequential

Linear depth:3 | input: (64, 10, 32) | output: (64, 10, 128)

GELU depth:3 | input: (64, 10, 128) | output: (64, 10, 128)

Linear depth:3 | input: (64, 10, 128) | output: (64, 10, 32)

LayerNorm depth:2 | input: (64, 10, 32) | output: (64, 10, 32)

Linear depth:1 | input: (64, 10, 32) | output: (64, 10, 1)

output-tensor depth:0 | (64, 10, 32)

squeeze depth:1 | input: (64, 10, 1) | output: (64, 10)

output-tensor depth:0 | (64, 10)