Week 19

Summary of the New MatMul

A: d1xd2, B: d2xd3, C = A * B: d1xd3

- matmul_t: unit-stride load A
  - matmul_t(A^T, B) = C^T
- matmul: strided load A
  - matmul(A, B) = C^T
- Negligible performance difference

Summary of the New MatMul

- From Linear Algebra:

$$C = A * B$$
$$C^T = (A * B)^T = B^T * A^T$$

matmul(B^T, A^T) = (C^T)^T = C

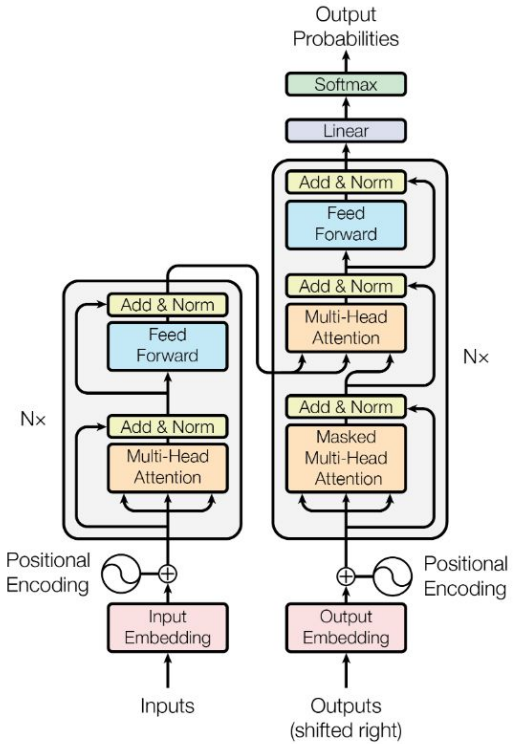
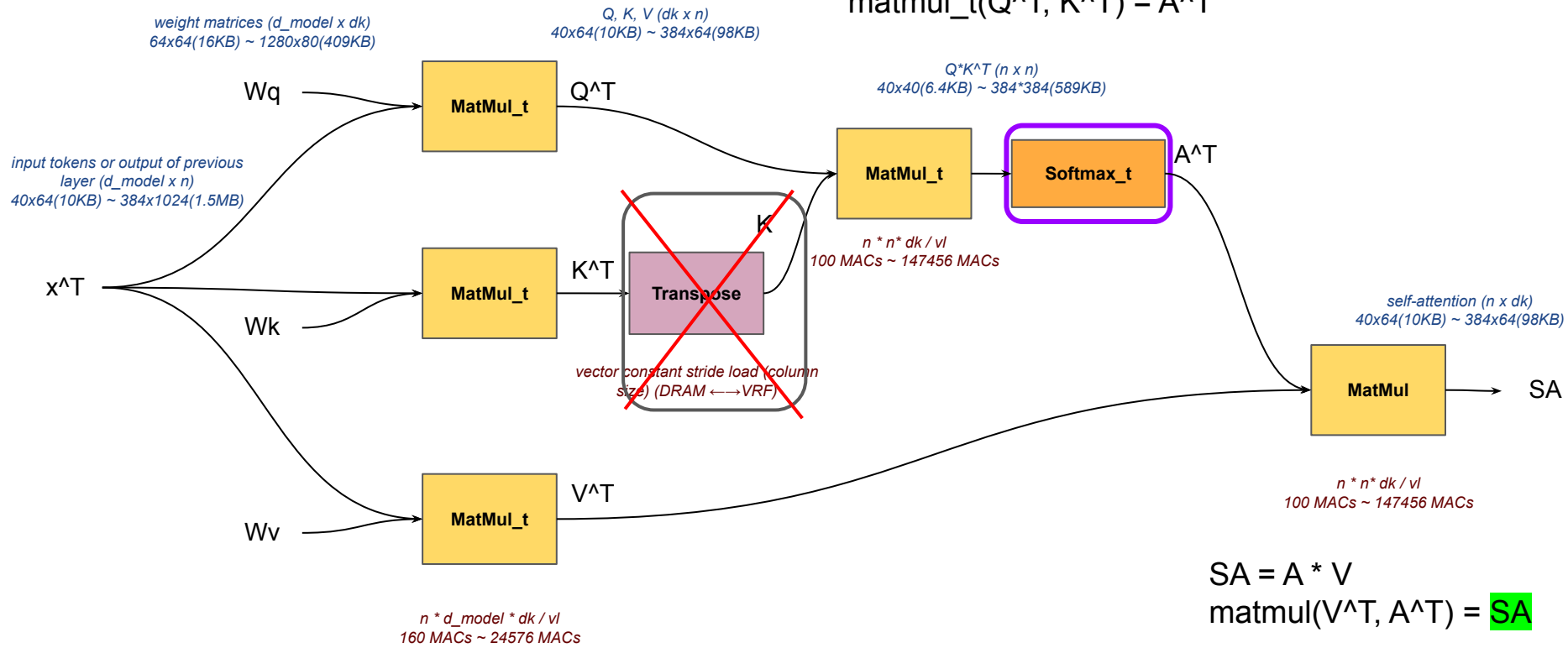
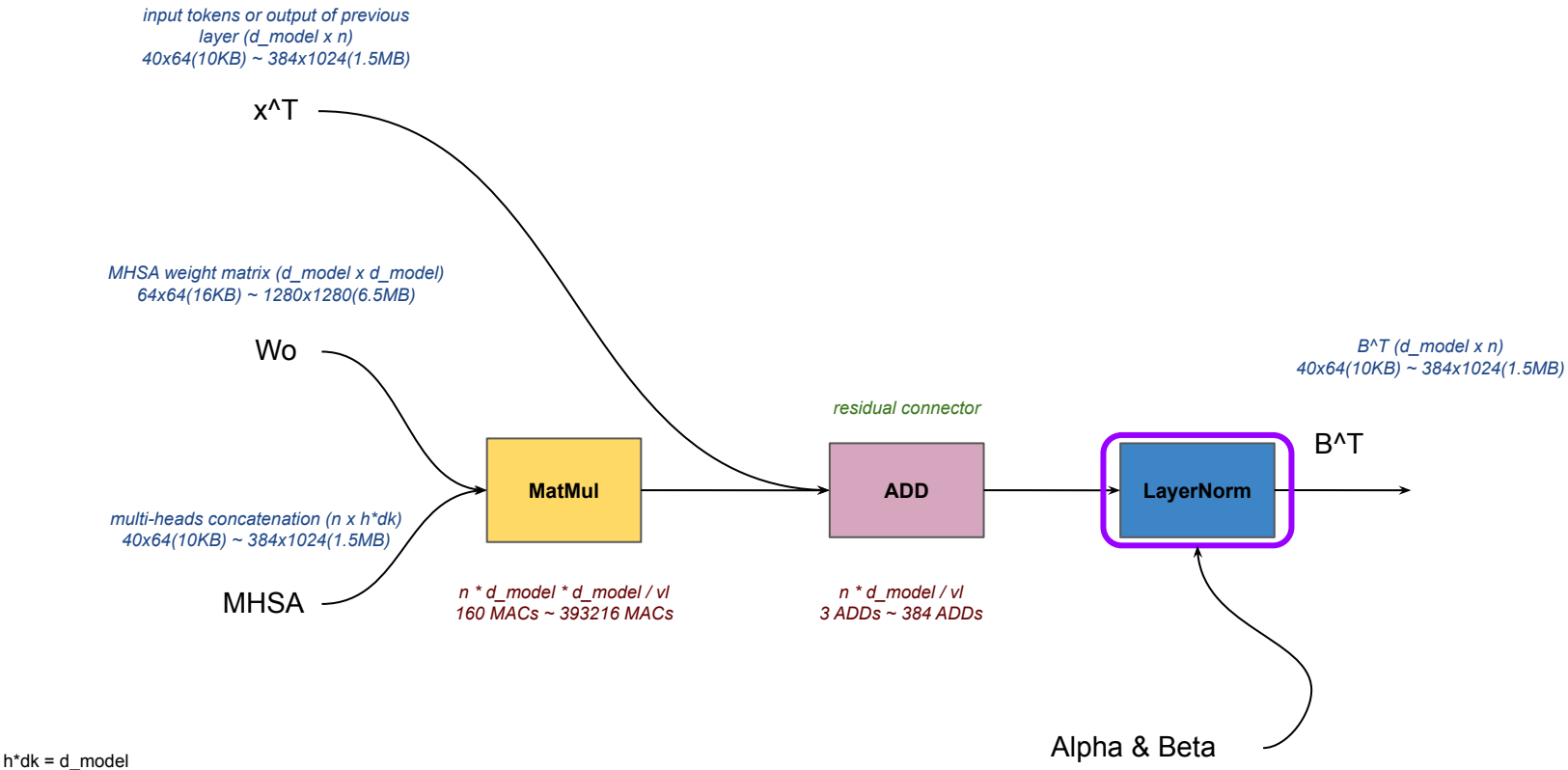


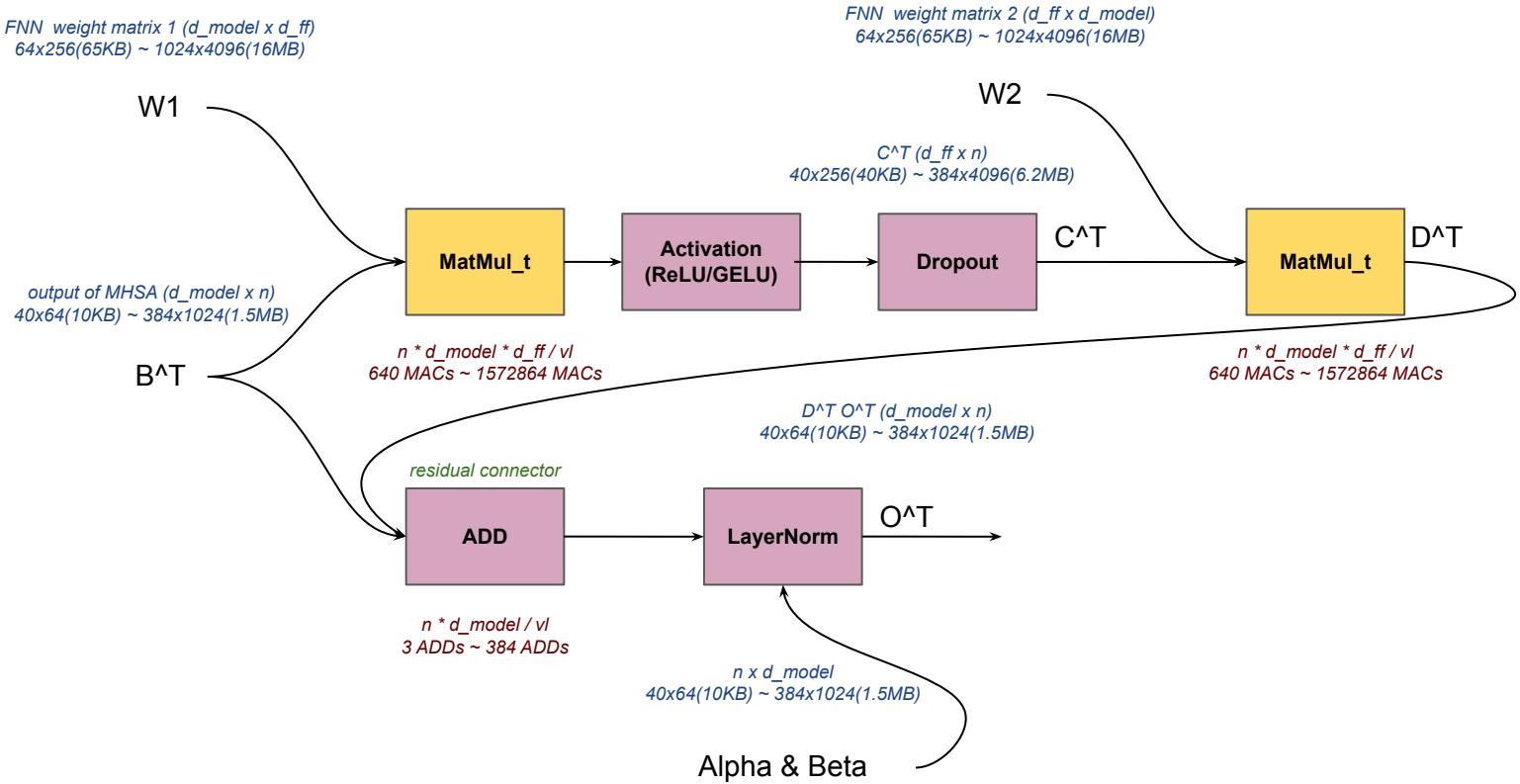
Figure 1: The Transformer - model architecture.



data format: fp32, vector length (vl): 1024, MAC: multiply-accumulate operation


input tokens or output of previous layer (d_model x n)
40x64(10KB) ~ 384x1024(1.5MB)
x^T
MHSA weight matrix (d_model x d_model)
64x64(16KB) ~ 1280x1280(6.5MB)
Wo
multi-heads concatenation (n x h*dk)
40x64(10KB) ~ 384x1024(1.5MB)
MHSA
MatMul
n * d_model * d_model / vl
160 MACs ~ 393216 MACs
residual connector
ADD
n * d_model / vl
3 ADDs ~ 384 ADDs
LayerNorm
B^T (d_model x n)
40x64(10KB) ~ 384x1024(1.5MB)
B^T
Alpha & Beta

h*dk = d_model


FNN weight matrix 1 (d_model x d_ff)
64x256(65KB) ~ 1024x4096(16MB)
W1
output of MHSA (d_model x n)
40x64(10KB) ~ 384x1024(1.5MB)
B^T
MatMul_t
Activation (ReLU/GELU)
Dropout
C^T
C^T (d_ff x n)
40x256(40KB) ~ 384x4096(6.2MB)
n * d_model * d_ff / vl
640 MACs ~ 1572864 MACs
FNN weight matrix 2 (d_ff x d_model)
64x256(65KB) ~ 1024x4096(16MB)
W2
MatMul_t
D^T
n * d_model * d_ff / vl
640 MACs ~ 1572864 MACs
D^T O^T (d_model x n)
40x64(10KB) ~ 384x1024(1.5MB)
residual connector
ADD
LayerNorm
O^T
n * d_model / vl
3 ADDs ~ 384 ADDs
n x d_model
40x64(10KB) ~ 384x1024(1.5MB)
Alpha & Beta