Linear

Concat

Scaled Dot-Product Attention

h

Linear

Linear

Linear

V

K

Q

n x d_v

n x n

n x n

n x n

n x n

MatMul

SoftMax

Mask (opt.)

Scale

MatMul

n x d_k

n x d_v

Q   K   V

Answer Key

n x d_model

n x h*d_v

h*d_v x d_model

Linear

n x d_v

Concat

Answer Key

Scaled Dot-Product Attention

h

n x d_v

n x d_k

Linear   Linear   Linear

d_model x d_v

d_model x d_k

V   K   Q

n x d_model   n x d_model   n x d_model