Forward:

input (seq, vec)

Query (seq, vec) = input (seq, vec) x wQ (vec, vec)

Key (seq, vec) = input (seq, vec) x wK (vec, vec)

Value (seq, vec) = input (seq, vec) x wV (vec, vec)

TKey (vec, seq) = Transpose Key (seq, vec)

A (seq, seq) = Query (seq, vec) x Tkey (vec, seq)

S (seq, seq) = Softmax (A (seq, seq) / sqrt(vec))

O1 (seq, vec) = S (seq, seq) x Value (seq, vec)

O2 (seq, seq\_o) = Softmax (O1 (seq, vec) x output (vec, seq\_o))

input =

wV =

Value = input x wV =

Value12 = input11 x wV12 + input12 x wV22 +…

Value22 = input21 x wV12 + input22 x wV22 + …

Value11 = input12 x wV21 + …

Value21 = input22 x wV21 + …

total = wV row

S =

O1 = S x Value =

O1 12 = S11 x Value12n + S12 x Value22n + …

O1 22 = S21 x Value12n + S22 x Value22n + …

O1 11 = S11 x Value11 + S12 Value21 + …

total = Value row

output =

O2 = Softmax (O1 x output) =

O2 11 = O1 12n x output 21 + O1 11 x output 11 + …

O2 12 = O1 12n x output 22 + O1 11 x output 12 + …

O2 13 = O1 12n x output 23 + O1 11 x output 13 + …

O2 21 = O1 22n x output 21 + O1 21 x output 11 + …

O2 22 = O1 22n x output 22 + O1 21 x output 12 + …

O2 23 = O1 22n x output 23 + O1 21 x output 13 + …

total = O2 row x O2 col

Backward:

E/wV = E/O2 x O2/O1 x O1/Value x Value/wV

wV22 = E/O2 x O2/O1 x O1/Value x Value/wV

E11 = E11/O2 11 x O2 11/O1 12 x (O1 12/Value12 + O1 12/Value 22) x

= error 11 x output 21 x

wV 22 = E 1/O2 11 x O2 11/O1 12 x (O1 12/Value 12 + O1 12/Value 22) x (Value 12/wV 22 + Value 22/wV 22)

= error 1 x output 21 x (S11 + S12) x (input 12 + input 22)

= E 2/O2 12 x O2 12/O1 12 x (O1 12/Value 12 + O1 12/Value 22) x (Value 12/wV 22 + Value 22/wV 22)

= error 2 x output 22 x (S11 + S12) x (input 12 + input 22)

= E 3/O2 13 x O2 13/O1 12 x (O1 12/Value 12 + O1 12/Value 22) x (Value 12/wV 22 + Value 22/wV 22)

= error 3 x output 23 x (S11 + S12) x (input 12 + input 22)

= E 4/O2 21 x O2 21/O1 22 x (O1 22/Value 12 + O1 22/Value 22) x (Value 12/wV 22 + Value 22/wV 22)

= error 4 x output 21 x (S21 + S22) x (input 12 + input 22)

= E 5/O2 22 x O2 22/O1 22 x (O1 22/Value 12 + O1 22/Value 22) x (Value 12/wV 22 + Value 22/wV 22)

= error 5 x output 22 x (S21 + S22) x (input 12 + input 22)

= E 6/O2 23 x O2 23/O1 22 x (O1 22/Value 12 + O1 22/Value 22) x (Value 12/wV 22 + Value 22/wV 22)

= error 6 x output 23 x (S21 + S22) x (input 12 + input 22)