
Algorithm 1 Forward Attention

```
1  def attention(x):
2      Q = x @ W_q
3      K = x @ W_k
4      V = x @ W_v
5      S = Q @ K.T / D**.5
6      P = softmax(S)
7      O = P @ V
8
9
10     return O
```

Algorithm 2 Backward Attention

```
1  def attention_backward(dO):
2      dV = P.T @ dO
3      dP = dO @ V.T
4      dS = dsoftmax(dP)
5      dQ = dS @ K / D**.5
6      dK = dS.T @ Q / D**.5
7      dW_v = x.T @ dV
8      dW_q = x.T @ dQ
9      dW_k = x.T @ dK
10     return dW_v, dW_q, dW_k
```