

# Intro to Diff Attention

written by Sabarish Kuduwa on Functor Network  
original link: <https://functor.network/user/1090/entry/560>

---

Recently, Microsoft (Tianzhu Ye et. al.) released a paper called differential attention which is supposed to cancel noise in attention module. The concept is similar to how noise cancelling headphones work.

In the attention block,  $Q$  and  $K$  are projected to  $W_1^Q, W_2^Q, W_1^K, W_2^K$ . The idea here is we can use  $W_1^Q$  and  $W_1^K$  to be coming from “one microphone” and  $W_2^Q$  and  $W_2^K$  to be coming from “other microphone”. The attention is modified as follows,

$$A_{diff} = softmax(\frac{W_1^Q(W_1^K)^T}{\sqrt{d}}) - \lambda softmax(\frac{W_2^Q(W_2^K)^T}{\sqrt{d}})$$

Here,  $\lambda$  is learnable scalar and authors have defined it as

$$\lambda = exp(\lambda_{q1} \cdot \lambda_{k1}) - exp(\lambda_{q2} \cdot \lambda_{k2}) + \lambda_{init}$$

$\lambda_{q1}, \lambda_{k1}, \lambda_{q2}, \lambda_{k2}$  are learned through backprop. Whereas,  $\lambda_{init}$  is defined as  $0.8 - 0.6 \times exp(-0.3 \cdot (l - 1))$ .

## My thoughts

- Based on the equation, this does seem promising in minimizing noise in attention blocks. I’m eager to try this out in encoder-decoder model and visualize attention maps.
- Not sure how authors have reached to initialization of  $\lambda_{init}$ , but they claim it is robust to any initializations.