Attention Is All You Need

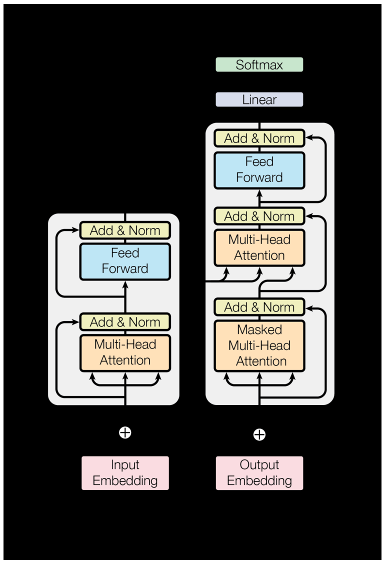
Introduction

In practice, we compute the attention function on a set of queries simultaneously, packed together
into a matrix Q. The keys and values are also packed together into matrices K and V . We compute
the matrix of outputs as:
$\text{Attention}(Q,K, V) = \text{softmax}(QKT\sqrt{}dk)V$
The two most commonly used attention functions are additive attention [2], and dot-product (multiplicative)
attention. Dot-product attention is identical to our algorithm, except for the scaling factor
of $1 \sqrt{}dk$
.
With a single attention head, averaging inhibits this.
$\text{MultiHead}(Q,K, V) = \text{Concat}(\text{head1}, ..., \text{headh})WO$
where $\text{headi} = \text{Attention}(QWQ$
i ,KWK
i , VWV
$i)$
Where the projections are parameter matricesWQ
$i \in Rdmodel\times dk ,WK$
$i \in Rdmodel\times dk ,WV$
$i \in Rdmodel\times dv$
and $WO \in Rhdv\times dmodel$.
In this work we employ $h = 8$ parallel attention layers, or heads. For each of these we use
$dk = dv = dmodel/h = 64$. Due to the reduced dimension of each head, the total computational cost
is similar to that of single-head attention with full dimensionality

ReLU activation in between.
$\text{FFN}(x) = \max(0,xW1 + b1)W2 + b2$ (2)
While the linear transformations are the same across different positions

In this work, we use sine and cosine functions of different frequencies:
$PE(pos,2i) = \sin(pos/100002i/dmodel)$
$PE(pos,2i+1) = \cos(pos/100002i/dmodel)$
where pos is the position and i is the dimension. That is, each dimension of the positional encoding

We used the Adam optimizer [20] with $\beta 1 = 0.9$, $\beta 2 = 0.98$ and $\epsilon = 10-9$. We varied the learning
rate over the course of training, according to the formula:
$lrate = d-0.5$
$model \cdot \min(step\_num-0.5, step\_num \cdot warmup\_steps-1.5)$ (3)
This corresponds to increasing the learning rate linearly for the first warmup_steps training steps,
and decreasing it thereafter proportionally to the inverse square root of the step number. We used
warmup_steps = 4000.



$$\bar{V}_S = \bar{A}\,\bar{V}_R + \bar{B}\,\bar{I}_R \quad \text{and} \quad \bar{I}_S = \bar{C}\,\bar{V}_R + \bar{D}\,\bar{I}_R$$

We have,

$$\boxed{\bar{A} = \bar{D} = \cosh\gamma l = \cosh(\sqrt{yz}\,l) = \cosh\sqrt{(ly)(lz)} = \cosh(\sqrt{YZ})}$$

$$\bar{B} = Z_C \sinh\gamma l = \sqrt{\frac{z}{y}}\,\sinh\left(\sqrt{yz}\,l\right)$$

$$\boxed{\bar{B} = \sqrt{\frac{l\,z}{l\,y}}\,\sinh(\sqrt{(ly)(lz)}) = \sqrt{\frac{Z}{Y}}\,\sinh(\sqrt{YZ})}$$

$$\bar{C} = \frac{1}{Z_C}\sinh\gamma l = \sqrt{\frac{y}{z}}\,\sinh(\sqrt{yz}\,l)$$

$$\therefore \quad \boxed{\bar{C} = \sqrt{\frac{l\,y}{l\,z}}\,\sinh(\sqrt{(ly)(lz)}) = \sqrt{\frac{Y}{Z}}\,\sinh(\sqrt{YZ})}$$

As $\bar{A} = \bar{D}$, we can say that the network is symmetrical.

Consider,

$$\bar{A}\,\bar{D} - \bar{B}\,\bar{C} = [(\cosh\sqrt{YZ})(\cosh\sqrt{YZ})] - \left[\sqrt{\frac{Z}{Y}}\,\sinh(\sqrt{YZ})\sqrt{\frac{Y}{Z}}\,\sinh(\sqrt{YZ})\right]$$

$$= \cosh^2\left(\sqrt{YZ}\right) - \sinh^2\left(\sqrt{YZ}\right) = 1$$

As $\bar{A}\,\bar{D} - \bar{B}\,\bar{C} = 1$ we can say that the network is reciprocal also.