Our tensor notation is discussed in Section **??** of Bayesuvius.
ℓ = number of words in a sentence segment. $\alpha \in [\ell]$, $\ell \sim 100$
L = number of words in vocabulary, $\beta \in [L]$, $L >> \ell$
$d = d_{\underline{q}} = d_{\underline{k}} = d_{\underline{v}} = 64$, hidden dimension per head, $\delta \in [d]$.
$n_{\underline{h}} = 8$, number of heads, $\nu \in [n_{\underline{h}}]$
$D = n_{\underline{h}} d = 8(64) = 512$, hidden dimension for all heads. $\Delta \in [D]$
reshaping

$$T^{\nu,\delta} \to T^{\Delta(\nu,\delta)} \quad \left(T^{[n_{\underline{h}}],[d]} \to T^{[D]}\right) \tag{1}$$

$$T^{\Delta(\nu,\delta)} \to T^{\nu,\delta} \quad \left(T^{[D]} \to T^{[n_{\underline{h}}],[d]}\right) \tag{2}$$

concatenation

$$T^{[n]} = (T^0, T^1, \ldots, T^{n-1}) = (T^\nu)_{\nu \in [n]} \tag{3}$$

Hadamard product (element-wise, entry-wise multiplication)

$$T^{[n]} * S^{[n]} = (T^\nu S^\nu)_{\nu \in [n]} \tag{4}$$

Matrix multiplication ($T^{[n]} = T^{[n],[1]}$ is a column vector)

$$(T^{[n]})^T S^{[n]} = \text{scalar} \tag{5}$$

$$T^{[a],[b]} S^{[b],[c]} = \left[\sum_{\beta \in [b]} T^{\alpha,\beta} S^{\beta,\gamma}\right]_{\alpha \in [a], \gamma \in [c]} \tag{6}$$

$$e^{\delta,\alpha} = \sum_\beta E^{\delta,\beta} x^{\beta,\alpha} \quad \left(e^{[d],[\ell]} = E^{[d],[L]} x^{[L],[\ell]}\right) \tag{7}$$

$$Q^{\nu,\delta,\alpha} = \sum_{\delta'} W_{\underline{q}}^{\nu,\delta,\delta'} e^{\delta',\alpha} \quad \left(Q^{[D],[\ell]} = W_{\underline{q}}^{[D],[d]} E^{[d],[\ell]}\right) \tag{8}$$

$$K^{\nu,\delta,\alpha} = \sum_{\delta'} W_{\underline{k}}^{\nu,\delta,\delta'} e^{\delta',\alpha} \quad \left(K^{[D],[\ell]} = W_{\underline{k}}^{[D],[d]} E^{[d],[\ell]}\right) \tag{9}$$

$$V^{\nu,\delta,\alpha} = \sum_{\delta'} W_{\underline{v}}^{\nu,\delta,\delta'} e^{\delta',\alpha} \quad \left(V^{[D],[\ell]} = W_{\underline{v}}^{[D],[d]} E^{[d],[\ell]}\right) \tag{10}$$

$$B^{\nu,\alpha,\alpha'} = \frac{1}{\sqrt{d}} \sum_\delta Q^{\nu,\delta,\alpha} K^{\nu,\delta,\alpha'} \quad \left(B^{[n_h],[\ell],[\ell]} = \left[\frac{1}{\sqrt{d}} (Q^{\nu,[d],[\ell]})^T K^{\nu,[d],[\ell]}\right]_{\nu \in [n_{\underline{h}}]}\right) \tag{11}$$

$$
\begin{aligned}
A^{[n_{\underline{h}}],[d],[\ell]} &= \left[\sum_{\alpha} V^{\nu,[d],\alpha} \underbrace{\text{softmax}(B^{\nu,\alpha,[\ell]})}_{(B^*)^{\nu,\alpha,[\ell]}}\right]_{\nu\in[n_{\underline{h}}]} && (12)\\
&= \left[V^{\nu,[d],[\ell]}(B^*)^{\nu,[\ell],[\ell]}\right]_{\nu\in[n_{\underline{h}}]} && (13)
\end{aligned}
$$

$$
A^{[n_{\underline{h}}],[d],[\ell]} \to A^{[D],[\ell]} \tag{14}
$$

- **Positional Encoding Matrix**

  $E_{pos}^{[d],[\ell]}$

$$
E_{pos}^{\delta,\beta} = \begin{cases} \sin\left(\frac{\beta}{10^{4\delta/d}}\right) = \sin(2\pi\frac{\beta}{\lambda(\delta)}) & \text{if } \delta \text{ is even} \\ \cos\left(\frac{\beta}{10^{4(\delta-1)/d}}\right) = \cos(2\pi\frac{\beta}{\lambda(\delta)}) & \text{if } \delta \text{ is odd} \end{cases} \tag{15}
$$

  $E_{pos}^{\delta,\beta}$ changes in phase by $\pi/2$ every time δ changes by 1. Its wavelength λ is independent of β, but increases rapidly with δ, from $\lambda(\delta = 0) = 2\pi * 1$ to $\lambda(\delta = d) = 2\pi * 10^4$.

- **ReLU**

  For a tensor T of arbitrary shape

$$
ReLU(T) = (T)_+ = max(0, T) \tag{16}
$$

  max element-wise

- **Feed Forward neural net**

$$
F(x^{[1],[\ell]}) = ReLU(x^{[1],[\ell]}W_1^{[\ell],[d]} + b_1^{[1],[d]})W_2^{[d],[\ell]} + b_1^{[1],[\ell]} \tag{17}
$$

$$
F(x^{[\ell]}) = W_2^{[\ell],[d]}ReLU(W_1^{[d],[\ell]}x^{[\ell]} + b_1^{[d]}) + b_1^{[\ell]} \tag{18}
$$

- **softmax**

  softmax() takes a vector and returns a vector of probabilities of the same length

$$
x^{[n]} \to P^{[n]} \tag{19}
$$

  where

$$
P^{\alpha} = \frac{\exp(x^{\alpha})}{\sum_{\alpha\in[n]}\exp(x^{\alpha})} \quad \left(P^{[n]} = \frac{\exp(x^{[n]})}{\|\exp(x^{[n]})\|_0}\right) \tag{20}
$$

For example,

$$(1, 0, 0) \to (e, 1, 1)/norm \tag{21}$$

$$(10, 0, 0) \to (e^{10}, 1, 1)/norm \approx (1, 0, 0) \tag{22}$$

For any $a \in \mathbb{R}$,

$$(a, a, a) \to (1, 1, 1)/3 \tag{23}$$