A APPENDIX

Here we provide the derivation for Equation 1. Consider the Taylor expansion of this expression around $\rho = 0$

$$\|\texttt{Swish}(\mathbf{x})\|^2 = \frac{1}{4}\|\mathbf{x}\|^2 + O(|\rho|). \tag{8}$$

Putting this together with the mean squared norm of the feature embedding

$$E[\|\texttt{Emb}(\mathbf{x})\|^2] = E[\|\texttt{Swish}(\gamma\texttt{BatchNorm}(\texttt{PreEmb}(\mathbf{x})) + \beta)\|^2], \tag{9}$$

we have

$$E[\|\texttt{Emb}(\mathbf{x})\|^2] = \frac{1}{4}E[\|\gamma\texttt{BatchNorm}(\texttt{PreEmb}(\mathbf{x})) + \beta)\|^2] + O(|\rho|). \tag{10}$$

This further simplifies due to the normalizing effect of BatchNorm

$$E[\|\gamma\texttt{BatchNorm}(\texttt{PreEmb}(\mathbf{x})) + \beta\|^2] = \sum_i E\left[(\gamma_i\texttt{BatchNorm}(\texttt{PreEmb}(\mathbf{x}))_i + \beta_i)^2\right] \tag{11}$$

$$= \|\gamma\|^2 + \|\beta\|^2, \tag{12}$$

yielding the result

$$E[\|\texttt{Emb}(\mathbf{x})\|^2] = \frac{1}{4}\|\gamma\|^2 + \frac{1}{4}\|\beta\|^2 + O(|\rho|). \tag{13}$$