Supplementary Materials

Table 7. Effect of removing the recurrent connections in the LMU, measured on psMNIST. We see that removing the recurrent connections $(\boldsymbol{e}_h, \boldsymbol{W}_h)$ is beneficial, and although there is utility to having $\boldsymbol{e}_m$, it unfortunately hinders parallelization.

Model	Accuracy
LMU (original)	97.15
LMU (no $\boldsymbol{e}_h$)	97.27
LMU (no $\boldsymbol{e}_m$)	96.82
LMU (no $\boldsymbol{W}_h$)	97.42
LMU (no $\boldsymbol{W}_m$)	20.10
LMU (no f)	89.81

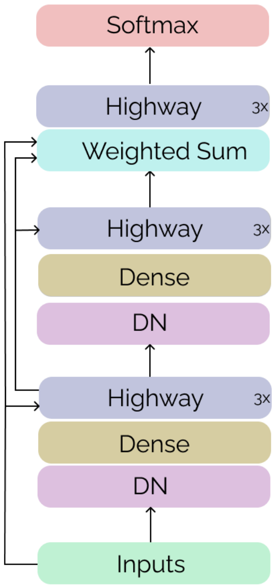

Figure 2. Illustration of the language model used for pre-training on the Amazon Reviews dataset. Although the actual model uses five blocks (combination of DN, Dense and Highway), we only show two blocks in the above figure.