PerDucer Formulations

2023 11033

May 2024

1 PerDucer Encoder Architecture

1.1 Behavior Tier (b-Tier) Architecture

$$\mathbf{e_{hd}}, \mathbf{e_{rel}}, \mathbf{e_{tl}} \in \mathbb{R}^d;$$

Head-Cell

$$\mathbf{e'_{hd}}_j = \tanh(W_{hd} \cdot \mathbf{e_{hd}}_j + \mathbf{b_{hd}});$$
$$\mathbf{h'_{b}}_{j-1} = \tanh(W_{h_b} \cdot \mathbf{h_b}_{j-1} + \mathbf{b_{h_b}});$$
$$\mathbf{h_{hd}}_j = \mathbf{e'_{hd}}_j + \mathbf{h'_{b}}_{j-1}$$

Relation-Cell

$$\mathbf{e'_{rel}}_j = \tanh(W_{rel} \cdot \mathbf{e_{rel}}_j + \mathbf{b_{rel}});$$
$$\mathbf{h'_{hd}}_j = \tanh(W_{h_b} \cdot \mathbf{h_{hd}}_{j_{\perp(\mathbf{e'_{rel}}_j)}} + \mathbf{b_{h_b}});$$
$$\mathbf{h_{rel}}_j = \mathbf{e'_{rel}}_j + \mathbf{h'_{hd}}_j$$

Tail-Cell

$$\mathbf{e'_{tl}}_j = \tanh(W_{tl} \cdot \mathbf{e_{tl}}_j + \mathbf{b_{tl}});$$
$$\mathbf{h'_{rel}}_j = \tanh(W_{h_b} \cdot \mathbf{h_{rel}}_{j_{\perp(\mathbf{e'_{tl}}_j)}} + \mathbf{b_{h_b}});$$
$$\mathbf{h_b}_j = \mathbf{e'_{tl}}_j + \mathbf{h'_{rel}}_j;$$
$$\mathbf{b}_j = \tanh((\mathbf{p}_j + W_b \cdot \mathbf{h_b}_j) + \mathbf{b_b})$$

where: One-Hot Positional Encoding: $\mathbf{p}_j = \mathbb{1}_j$

Where,

$$\mathbf{h'_{i}}_{\perp(\mathbf{r_j})} = \mathbf{h'_i} - (\frac{\mathbf{r_j}}{\|\mathbf{r_j}\|} \cdot \mathbf{h'_i}) \cdot (\frac{r_j}{\|\mathbf{r_j}\|})$$

1.2 Run Tier (r-Tier) Architecture

$$\begin{aligned}
&\mathbf{b'}_j = \tanh(W_{b_r} \cdot \mathbf{b}_j + \mathbf{b}_{\mathbf{b_r}});\\
&\mathbf{h'}_{\mathbf{r}_{j-1}} = \tanh(W_{h_r} \cdot \mathbf{h}_{\mathbf{r}j-1} + \mathbf{b}_{\mathbf{h_r}});\\
&\mathbf{h}_{\mathbf{r}_j} = \mathbf{b'}_j + \mathbf{h'}_{\mathbf{r}_{j-1}};\\
&\mathbf{r}_m = \tanh((\mathbf{p}_m + W_r \cdot \mathbf{h}_{\mathbf{r}_j}) + \mathbf{b}_{\mathbf{r}})
\end{aligned}$$

where: One-Hot Positional Encoding: $\mathbf{p}_m = \mathbb{1}_m$

2 PerDucer Decoder Architecture

2.1 MEGA-based Encoding of Run History

Learnable damped-EMA

$$\begin{aligned} \boldsymbol{r_t^{EMA}} &= \boldsymbol{\alpha_t} \odot \boldsymbol{r_t} + (\mathbf{1} - \boldsymbol{\alpha_t} \odot \boldsymbol{\delta_t}) \odot \boldsymbol{r_{t-1}^{EMA}} \\ \text{where: } \boldsymbol{\alpha_t} &= \tanh\left(W_\alpha \cdot \left(\boldsymbol{r_{t-1}^{EMA}} \| \boldsymbol{r_t}\right) + \boldsymbol{b_\alpha}\right); \\ \boldsymbol{\delta_t} &= \tanh\left(W_\delta \cdot \left(\boldsymbol{r_{t-1}^{EMA}} \| \boldsymbol{r_t}\right) + \boldsymbol{b_\delta}\right) \end{aligned}$$

Let $\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}}$ be a matrix composed of $\tau_{\mathcal{H}}$ rows of run $\boldsymbol{r}$. Each row is denoted by $\boldsymbol{r_t^{EMA}}$.

$$\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'} = \phi_{\text{silu}}(\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}} \cdot W_{EMA} + \boldsymbol{b_{EMA}})$$

Contextualizing the $\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'}$ via self-attention:

$$\begin{aligned} \boldsymbol{Q}_{\tau_{\mathcal{H}}}^{\text{EMA}} &= \boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'} \cdot W_q + \boldsymbol{b_q}; \\ \boldsymbol{K}_{\tau_{\mathcal{H}}}^{\text{EMA}} &= \boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'} \cdot W_k + \boldsymbol{b_k}; \\ \boldsymbol{V}_{\tau_{\mathcal{H}}}^{\text{EMA}} &= \boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'} \cdot W_v + \boldsymbol{b_z}; \\ \boldsymbol{Z}_{\tau_{\mathcal{H}}}^{\text{EMA}} &= \text{softmax}\left(\frac{\boldsymbol{Q}_{\tau_{\mathcal{H}}}^{\text{EMA}} \cdot (\boldsymbol{K}_{\tau_{\mathcal{H}}}^{\text{EMA}})^T}{\sqrt{\boldsymbol{d_Q}}}\right) \cdot \boldsymbol{V}_{\tau_{\mathcal{H}}}^{\text{EMA}} \end{aligned}$$

Applying forget gate to contextualized $\boldsymbol{Z}_{\tau_{\mathcal{H}}}^{\text{EMA}}$:

$$\boldsymbol{Z}_{\tau_{\mathcal{H}}}^{\text{EMA}^f} = \boldsymbol{f} \odot \boldsymbol{Z}_{\tau_{\mathcal{H}}}^{\text{EMA}}$$

$$\text{where: } \boldsymbol{f} = \phi_{\text{silu}}(\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'} \cdot W_f + \boldsymbol{b_f})$$

Combining contextualized (w/ forget gate) and non-contextualized versions of $\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'}$:

$$\boldsymbol{Z}_{\tau_{\mathcal{H}}}^{\text{EMA}^C} = \phi_{\text{silu}}(\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'} \cdot W_{EMA}^C + \boldsymbol{Z}_{\tau_{\mathcal{H}}}^{\text{EMA}^f} \cdot W_z^C + \boldsymbol{b_C})$$

Gated Residual ($\boldsymbol{i}$) based hidden-state representation of user preference-history ($\boldsymbol{R}_{\tau_{\mathcal{H}}}^h$):

$$\boldsymbol{R}_{\tau_{\mathcal{H}}}^h = \textbf{MEGA}(\boldsymbol{R}_{\tau_{\mathcal{H}}}) = \boldsymbol{i} \odot \boldsymbol{Z}_{\tau_{\mathcal{H}}}^{\text{EMA}^C} + (1 - \boldsymbol{i}) \odot \boldsymbol{R}_{\tau_{\mathcal{H}}}$$

where $\boldsymbol{i} = \sigma(\boldsymbol{R}_{\tau_{\mathcal{H}}}^{\text{EMA}'} \cdot W_i + \boldsymbol{b_i})$

2.2 Prediction of upcoming runs:

$$\hat{\boldsymbol{r}}_{\tau_{\mathcal{F}}} = W_{\tau_{\mathcal{F}}} \cdot \boldsymbol{R}^{h}_{\tau_{\mathcal{H}}} + \boldsymbol{b}_{\boldsymbol{\tau_{\mathcal{F}}}}; \qquad \in \mathbb{R}^{|R_{KG}|\times 1}$$

$$\hat{\boldsymbol{p}}_{\tau_{\mathcal{F}}} = \mathrm{sigmoid}(\hat{\boldsymbol{r}}_{\tau_{\mathcal{F}}}) \qquad \in [0,1]^{|R_{KG}|\times 1}$$

3 Task 1 Training

3.1 Decoder Loss: Predicting future run sequence

$$\mathcal{L^F}_{\mathbf{r}}(\mathbf{p}, \hat{\mathbf{p}}) = [...\ \mathcal{L^F}_r(p_i, \hat{p_i})\ ...]$$

$$\mathcal{L^F}_r(p_r^t, \hat{p}_r^t) = \sum_{j=1:d} \left[p_r^{t^{(j)}} \log \frac{1}{\hat{p}_r^{t^{(j)}}} + (1 - p_r^{t^{(j)}}) \log \frac{1}{(1 - \hat{p}_r^{t^{(j)}})} \right]$$

$$\mathcal{L^F}_r(p_r^{\tau_\mathcal{F}}, \hat{p}_r^{\tau_\mathcal{F}}) = \sum_{t=(\tau_\mathcal{H}+1):\tau_\mathcal{F}} \mathcal{L^F}_b(p_r^t, \hat{p}_r^t)$$

3.2 Run-tier (r-Tier) History Encoding Loss

$$\mathcal{L^H}_r(p_i, \hat{p_i}) = \sum_j p_i^{(j)} \log \frac{1}{\hat{p}_i^{(j)}} + (1 - p_i^{(j)}) \log \frac{1}{(1 - \hat{p}_i^{(j)})}$$

$$\mathcal{L^H}_{\mathbf{r}}(\mathbf{p}, \hat{\mathbf{p}}) = [...\ \mathcal{L^H}_r(p_i, \hat{p_i})\ ...]$$

$$\begin{aligned}
\nabla^t_{W_r} \mathcal{L^H}_r(p_\tau, \hat{p_\tau}) &= \nabla^{t+1}_{W_r} \mathcal{L^H}_r(p_\tau, \hat{p_\tau}) + \nabla^t_{W_r} \mathcal{L^H}_r(p_t, \hat{p_t}) \\
&= \nabla^{t+1}_{W_r} \mathcal{L^H}_r(p_\tau, \hat{p_\tau}) + \left[\left(\nabla^t_{h_t} \mathcal{L^H}_r(p_t, \hat{p_t}) + \nabla^{t+1}_{W_r} \mathcal{L^H}_r(p_{t+1}, \hat{p}_{t+1}) \right) \cdot \nabla^t_{W_r} h_t \right]
\end{aligned}$$

3.3 Behavior-tier (b-Tier) History Encoding Loss (BPTT)

3.3.1 Local loss at time-step t:

$$\mathcal{L}^{\mathcal{H}}{}_{b}(p_b^t, \hat{p}_b^t) = \sum_{j=1:d} \left[p_b^{t^{(j)}} \log \frac{1}{\hat{p}_b^{t^{(j)}}} + (1 - p_b^{t^{(j)}}) \log \frac{1}{(1 - \hat{p}_b^{t^{(j)}})} \right]$$

$$\mathcal{L}^{\mathcal{H}}{}_{\mathbf{b}}(\mathbf{p}, \hat{\mathbf{p}}) = [...\ \mathcal{L}^{\mathcal{H}}{}_{b}(p_i, \hat{p}_i)\ ...]$$

3.3.2 Loss for Back-propagation-through-time (BPTT):

b-tier Encoder Loss: $\mathcal{L}^{\mathcal{H}}{}_b(p_b^{\tau_{\mathcal{H}}}, \hat{p}_b^{\tau_{\mathcal{H}}}) = \sum\limits_{t=1:\tau_{\mathcal{H}}} \mathcal{L}^{\mathcal{H}}{}_b(p_b^t, \hat{p}_b^t)$

Decoder Loss: $\mathcal{L}^{\mathcal{F}}{}_r(p_r^{\tau_{\mathcal{F}}}, \hat{p}_r^{\tau_{\mathcal{F}}})$

r-tier Cumulative Loss: $\mathcal{L}_r(p_r^{\tau_{\mathcal{F}+\mathcal{H}}}, \hat{p}_r^{\tau_{\mathcal{F}+\mathcal{H}}})$

b-tier Cumulative Loss: $\mathcal{L}_b(p_b^{\tau_{\mathcal{H}+\mathcal{F}}}, \hat{p}_b^{\tau_{\mathcal{H}+\mathcal{F}}}) = \mathcal{L}_r(p_r^{\tau_{\mathcal{F}+\mathcal{H}}}, \hat{p}_r^{\tau_{\mathcal{F}+\mathcal{H}}}) + \mathcal{L}^{\mathcal{H}}{}_b(p_b^{\tau_{\mathcal{H}}}, \hat{p}_b^{\tau_{\mathcal{H}}})$

$$
\begin{aligned}
\nabla_\theta^t \mathcal{L}_b(p_b^{\tau_{\mathcal{H}+\mathcal{F}}}, \hat{p}_b^{\tau_{\mathcal{H}+\mathcal{F}}}) &= \nabla_\theta^t \mathcal{L}^{\mathcal{H}}{}_b(p_b^{\tau_{\mathcal{H}}}, \hat{p}_b^{\tau_{\mathcal{H}}}) + \nabla_\theta^t \mathcal{L}_r(p_r^{\tau_{\mathcal{F}+\mathcal{H}}}, \hat{p}_r^{\tau_{\mathcal{F}+\mathcal{H}}}) \\
&= \left[\nabla_\theta^{t+1} \mathcal{L}^{\mathcal{H}}{}_b(p_b^{\tau_{\mathcal{H}}}, \hat{p}_b^{\tau_{\mathcal{H}}}) + \nabla_\theta^t \mathcal{L}^{\mathcal{H}}{}_b(p_b^t, \hat{p}_b^t)\right] + \nabla_\theta^t \mathcal{L}_r(p_r^{\tau_{\mathcal{F}+\mathcal{H}}}, \hat{p}_r^{\tau_{\mathcal{F}+\mathcal{H}}}) \\
&= \nabla_\theta^{t+1} \mathcal{L}^{\mathcal{H}}{}_b(p_b^{\tau_{\mathcal{H}}}, \hat{p}_b^{\tau_{\mathcal{H}}}) + \left[\left(\nabla_{h_b^t}^t \mathcal{L}^{\mathcal{H}}{}_b(p_b^t, \hat{p}_b^t) + \nabla_\theta^{t+1} \mathcal{L}^{\mathcal{H}}{}_b(p_b^{t+1}, \hat{p}_b^{t+1})\right) \cdot \nabla_\theta^t h_b^t\right] \\
&\quad + \nabla_\theta^t \mathcal{L}_r(p_r^{\tau_{\mathcal{F}+\mathcal{H}}}, \hat{p}_r^{\tau_{\mathcal{F}+\mathcal{H}}})
\end{aligned}
$$

where: $\theta \in \{W_{b_{tl}}, W_{b_{hd}}, W_{n_{hd}}, W_{n_{tl}}, W_{n_{rel}}\}$

4 Task 2: Next Run prediction Training

4.1 Query Summary Embedding Generation: Backward Translation

$$\hat{\mathbf{r}}^{\mathcal{F}}_{\tau_{\mathcal{H}}+1} = f_r(\hat{\mathbf{b}}_q, \hat{\mathbf{h}}_{r_{q-1}}) \implies \hat{\mathbf{b}}_q = f_r^{-1}(\hat{\mathbf{r}}^{\mathcal{F}}_{\tau_{\mathcal{H}}+1}, \hat{\mathbf{h}}_{r_{q-1}})$$

$$\hat{\mathbf{b}}_q = g_b(\hat{\mathbf{s}}_q, \hat{\mathbf{h}}_{summGen}) = g_b(\hat{\mathbf{s}}_q, g_{rel}(\hat{\mathbf{d}}_q, \hat{\mathbf{rel}}_{summGen}))$$

$$\because \hat{\mathbf{h}}_{summGen} = g_{rel}(\hat{\mathbf{d}}_q, \hat{\mathbf{rel}}_{summGen})$$

$$\therefore \hat{\mathbf{s}}_q = g_b^{-1}(g_{rel}(\hat{\mathbf{d}}_q, \hat{\mathbf{rel}}_{summGen}), \hat{\mathbf{b}}_q) = g_b^{-1}(g_{rel}(\hat{\mathbf{d}}_q, \hat{\mathbf{rel}}_{summGen}), f_r^{-1}(\hat{\mathbf{r}}^{\mathcal{F}}_{\tau_{\mathcal{H}}+1}, \hat{\mathbf{h}}_{r_{q-1}}))$$

5 Miscellaneous

$\hat{\mathbf{r}}$

$\hat{\mathbf{r}}_{l+1}$

$\hat{\mathbf{r}}_{l+2}$

$\hat{\mathbf{r}}_{\tau}$

$\hat{\mathbf{p}}_{l+1}$

$\hat{\mathbf{p}}_{l+2}$

$\hat{\mathbf{p}}_{\tau}$

$\hat{\mathbf{P}}_{\text{next}}$

$\mathcal{L}^{\mathcal{H}}{}_{b}(p_b^{\tau\mathcal{H}}, \hat{p}_b^{\tau\mathcal{H}})$