How to Scale Your EMA

Что такое EMA?

- EMA - Exponential Moving Average - взвешенное усреднение модели

$$\zeta_{t+1} = \rho\,\zeta_t + (1-\rho)\,\theta_t$$

- У EMA более широкий район оптимума.
- EMA - простой способ получить модель, отличающуюся от исходной. Например, для дистилляции.
- Авторы статьи предлагают при увеличении батча в k раз, заменять p на p^k

EMA для SGD

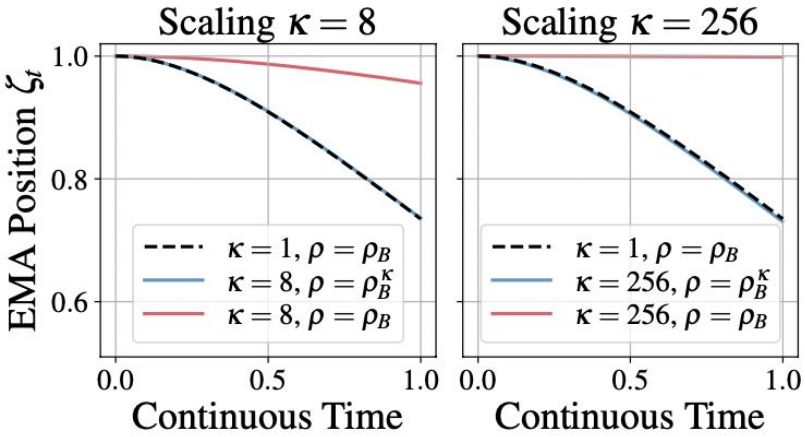
$$\theta_{t+1} = \theta_t - \eta \times \frac{1}{B} \sum_{x \in \mathbb{B}} \nabla_\theta \mathcal{L}(x; \theta_t),$$

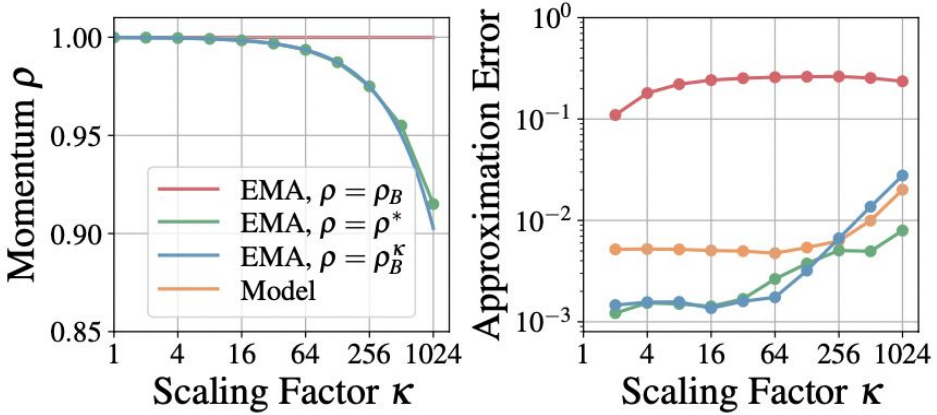
$$\nabla_\theta \mathcal{L}(x; \theta_{t+j}, \zeta_{t+j}) \approx \nabla_\theta \mathcal{L}(x; \theta_t, \zeta_t) \approx \mathbf{g},$$

$$\begin{bmatrix} \theta_{t+\kappa} \\ \zeta_{t+\kappa} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\eta \\ (1-\rho) & \rho & 0 \\ 0 & 0 & 1 \end{bmatrix}^{\kappa} \cdot \begin{bmatrix} \theta_t \\ \zeta_t \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} \theta_t - \eta\, \kappa\, \mathbf{g} \\ \rho^{\kappa}\, \zeta_t + (1-\rho^{\kappa})\, \theta_t + O\left(\eta \times \beta_\rho\right) \\ \mathbf{g} \end{bmatrix}.$$

Polyak Ruppert averaging

На inference используем модель с EMA весами





Semi-supervised learning

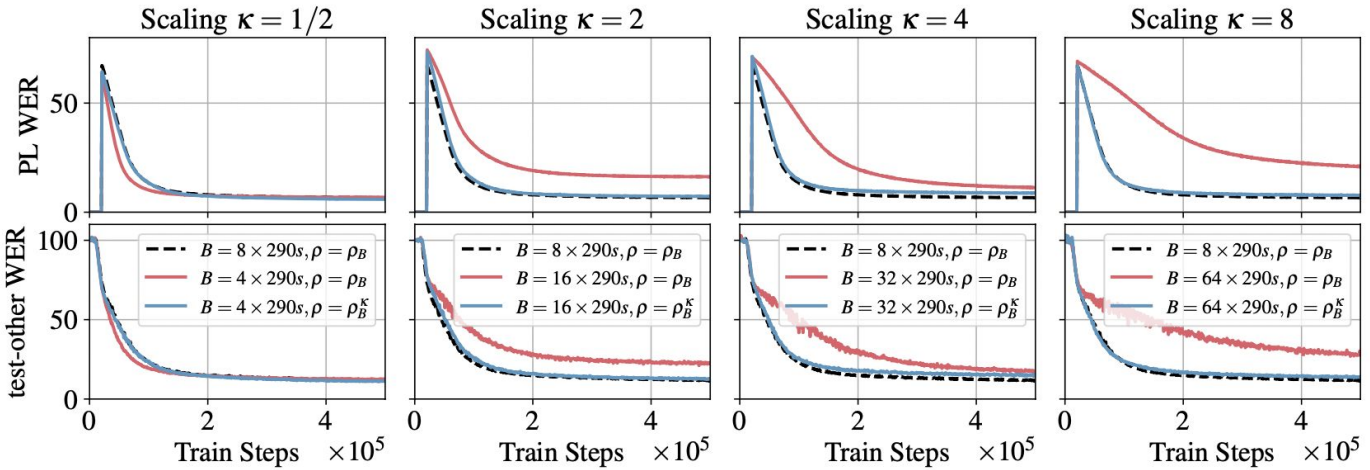


Figure 4: *Transformer pseudo-labeling on LibriSpeech* with different scalings κ. The baseline ($\kappa = 1$, black dashed) is trained with Adam at a *dynamic batch size* of 8×290 seconds, which corresponds to a single train step on the x-axis. The model EMA (*teacher*) is updated with momentum $\rho_B = 0.9999$. We investigate dynamic batch sizes down to $B = 4 \times 290s$ (left) and up to $B = 64 \times 290s$ (right), with (blue, $\rho = \rho_B^\kappa$) and without (red, $\rho = \rho_B$) the EMA Scaling Rule. The Adam Scaling Rule (Malladi et al. (2022), Definition C.3) is used throughout. For $\kappa \leq 2$, we start pseudo-labeling after $20\text{k}/\kappa$ training steps; while for $\kappa > 2$, we start when pre-training WER matches the baseline WER.

Self-supervised