Batch-wise Convergent Pre-training:
Step-by-Step Learning Inspired by Child Language Development

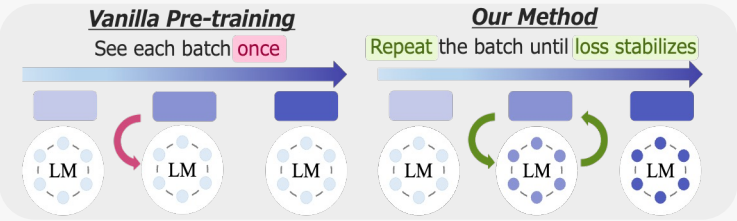
○Ko Yoshida[1], Daiki Shiono[1], Kai Sato[1], Toko Miura[1], Momoka Furuhashi[1], Jun Suzuki[1,2,3]
[1]Tohoku University, [2]RIKEN, [3]NII LLMC

Overview

Concept: What if a LM could gradually ***accumulate knowledge*** through ***repetitions*** within ***limited contexts*** – like children?

Method :
Learn ***each batch repeatedly*** with ***regularization*** to ***reduce forgetting***

○Method Overview

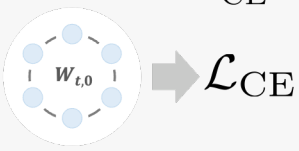


Method:Batch-wise Convergent Pre-training

Step1 Get $\mathcal{L}_{\mathrm{CE}}$ for batch $\mathcal{X}_t$

Let's Say:
- $\mathcal{X}_t$: Target Batch
- $\boldsymbol{W}_{t,0}$: Learnable Params*1
- α : Adaptive learning strength*2

Caluculate $\mathcal{L}_{\mathrm{CE}}$

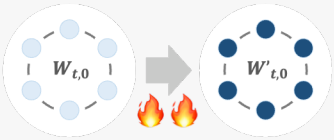


Step2 Parameter Update with α

①How far from $\mathcal{L}_{\mathrm{CE}_{\mathrm{goal}}}$ ②Accelerate with α'*4

$$\alpha' \leftarrow f(\mathcal{L}_{\mathrm{CE}}, \alpha)^{*3}$$

Boost learning for this batch!🔥🔥



*1
- Start with $\boldsymbol{W}_{t,0} = \boldsymbol{W}_{t-1,n}$
- Update iteratively to $\boldsymbol{W}_{t,k}$

*2
- Start with $\alpha = 1$
- α grows when $\mathcal{L}_{\mathrm{CE}} > \mathcal{L}_{\mathrm{CE}_{\mathrm{goal}}}$

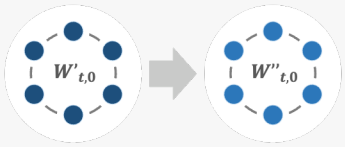
*3
- What is $f(\mathcal{L}_{\mathrm{CE}}, \alpha)$?
  $\alpha' \leftarrow \alpha + \eta(\mathcal{L}_{\mathrm{CE}} - \mathcal{L}_{\mathrm{CE}_{\mathrm{goal}}})$
- η is learning rate for α update

*4
- Use AdamW
  $W' \leftarrow \mathrm{AdamW}(\alpha' \mathcal{L}_{\mathrm{CE}})$

Step3 Pull back W' toward $W_{t,0}$

①Regularization term ②🧲 Regularize W'*6

$$\nabla R = \frac{C}{p} \left\| \mathbf{W}' - \mathbf{W}_{t,0} \right\|_p^{p\,*5}$$



*5
- C controls the regularization strength
- L_1 when $p = 1$, L_2 when $p = 2$

*6 Regularization
$$\mathbf{W}'' = \mathbf{W}' - \nabla R$$

Step4 Convergence Check

Has the model learned $\mathcal{X}_t$ sufficiently?
→ Check α' , Not $\mathcal{L}_{\mathrm{CE}}$

Check List
- ☐ Is α decreasing continuously?
- ☐ α < threshold?

✓ Next Batch $\mathcal{X}_{t+1}$
$\mathbf{W}_{t+1,0} \leftarrow \mathbf{W}_{t,0}$

✗ Repeat Batch $\mathcal{X}_t$
$\mathbf{W}_{t,1} \leftarrow \mathbf{W}_{t,0}$

Experiments/Analysis

Comparison with Official Baselines

Setup:
- We compare our model with **GPT-2**, **GPT-BERT**
- Model size is 117M, Qwen2.5 architecture
- A curriculum based on our original difficulty score

Result: No significant improvement

Model	BLiMP↑	BLiMP-S↑	WUG-ADJ↑	Text-Avg.↑
GPT-BERT	**80.5**	**73.0**	41.2	**70.9**
GPT-2	74.9	63.3	50.2	54.7
Ours(p=1)	49.2	50.4	**57.5**	32.8
Ours(p=2)	52.2	50.2	57.1	32.5

Next TODO : Isolated batch learning breaks the distributional assumption
→ Batch Design? or Architecture Design?

Training Orders & Repetition Strategies

Setup:
1. ***Random***: 10 epochs with the random data
2. ***Curriculum***: 10 epochs with the curriculum data
3. ***Curriculum-Repeat***: the curriculum data + repeat each batch 10 times in a row
4. ***Our Proposed Method***

Result: Comparable but not better

Analysis:

😕 **Our Proposed Method** performs slightly better in Text-Avg, by only 2 points

👀
Random performed best on BLiMP
Curriculum performed best on WUG-ADJ
Curriculum-Repeat performed best on Entity