# Continual Training of Language Models for Few-shot Learning

**Zixuan Ke[1], Haowei Lin[2], Yijia Shao[2], Hu Xu[3], Lei Shu[4] and Bing Liu[1]**
University of Illinois at Chicago[1], Peking University[2], Meta AI[3] and Google Resaerch[4]

## Continual Post-Training

- Post-training (a.k.a., domain-adaptive pre-training or pre-finetuning) helps an LM achieves better results
- CPT goes a step further
  - Continually improves an LM's ability to handle new and emerging domains.
- This is importance because
  - The world is dynamic (think about the ever-changing variants of COVID)
  - As re-training an LM from scratch is extremely expensive, incrementally updating the LM with the latest data is critical
- **Our goal:** Continually post-train an LM in a sequence of domains, *without* forgetting its learned skills

## Proposed CPT Model: Parallel CL-plugin and Task Masking

- **CL-plugin**
  - Parallel adapters (shared by all domains)
- **Task Mask**
  - Train the mask ($e_l^{(t)}$: task embedding for layer $l$ in task $t$)

$$m_l^{(t)} = \sigma(e_l^{(t)}/\tau)$$

  - Forward pass ($k_l^{(t)}$: output of layer $l$ in task $t$)

$$o_l^{(t)} = (k_l^{(t)} \otimes m_l^{(t)})$$

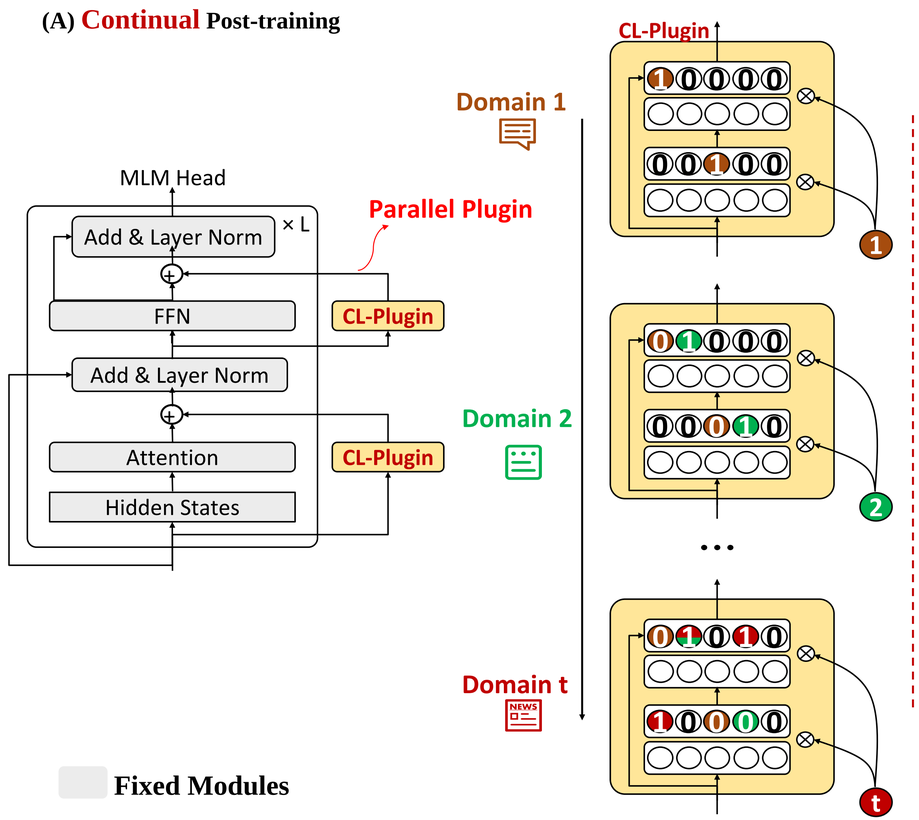
  - Backward pass ($i_{\text{prev}}$: all old tasks)

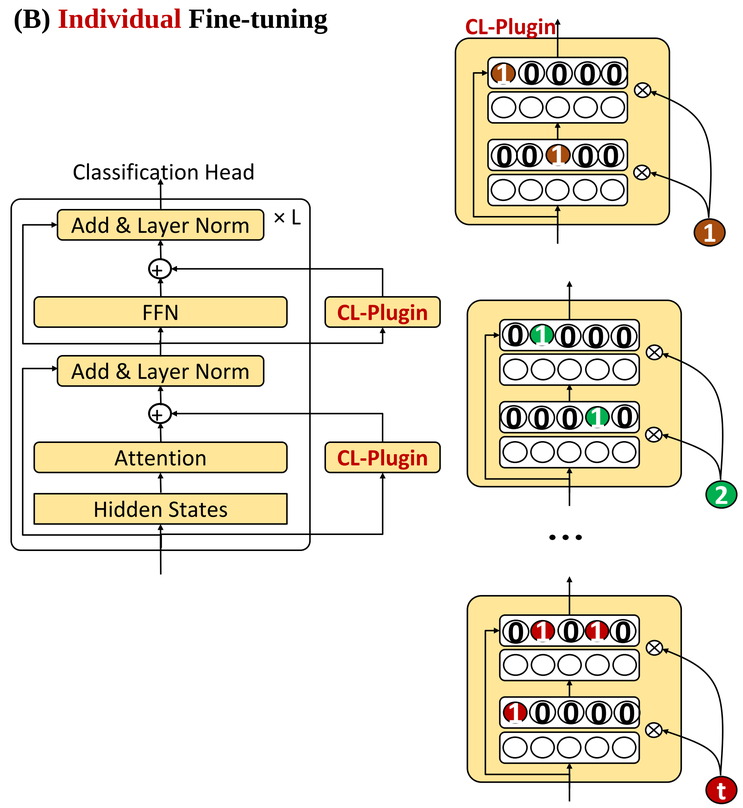
$$\widehat{\nabla}_l^{(t)} = \nabla_l^{(t)} \otimes (1 - \max(\{m_l^{(i_{\text{prev}})}\}))$$

  - Catastrophic Butterfly Effect (make the mask a hard mask; used in gradient manipulation and forward pass in end-task fine-tuning)

$$m_l^{(t)} = \begin{cases} 1, & m_l^{(t)} > \theta, \\ 0, & \text{Otherwise} \end{cases}$$

**Code, data, post-trained models: https://github.com/UIC-Liu-Lab/CPT**




**MLM head for unsupervised post-training of the plugins only**

**Use the final post-trained model (with different masks) to evaluate the post-trained CPT**

## Datasets

| Unlabeled Domain Datasets | | | End-Task Classification Datasets | | | | |
|---|---|---|---|---|---|---|---|
| Dataset | Source | #training | Dataset | Task | #training | #testing | #classes |
| Yelp Restaurant | Yelp Review | 1,132,359 | SemEval-res | Aspect Sentiment Classification | 32 | 1,120 | 3 |
| AI | AI Papers | 707,368 | SCIERC | Relation Classification | 56 | 2,388 | 7 |
| ACL | ACL Papers | 1,208,449 | ACL-ARC | Citation Intent Classification | 48 | 421 | 6 |
| AGNews | News Article | 73,750 | AGNews-FT | News Classification | 32 | 7,568 | 4 |

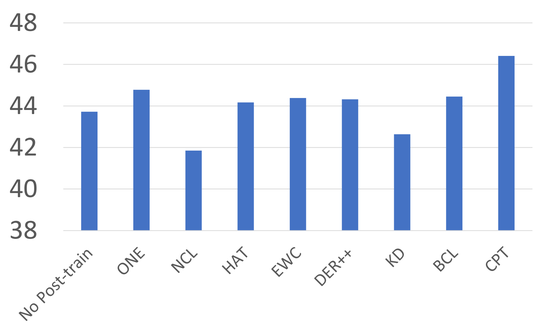
## Summary

✓ Proposed the problem of **Continual Post-training**

✓ Proposed an effective system CPT
- The key is to use **task masks** to protect the learned knowledge and prevent butterfly effect

✓ Experimental results show that
- It achieves no forgetting and outperforms a large number of baselines

## Experimental Results




✓ **0** forgetting rate

✓ Outperforms **13** SOTA baselines, including MLM, HAT, DER++, EWC, DEMIX etc.