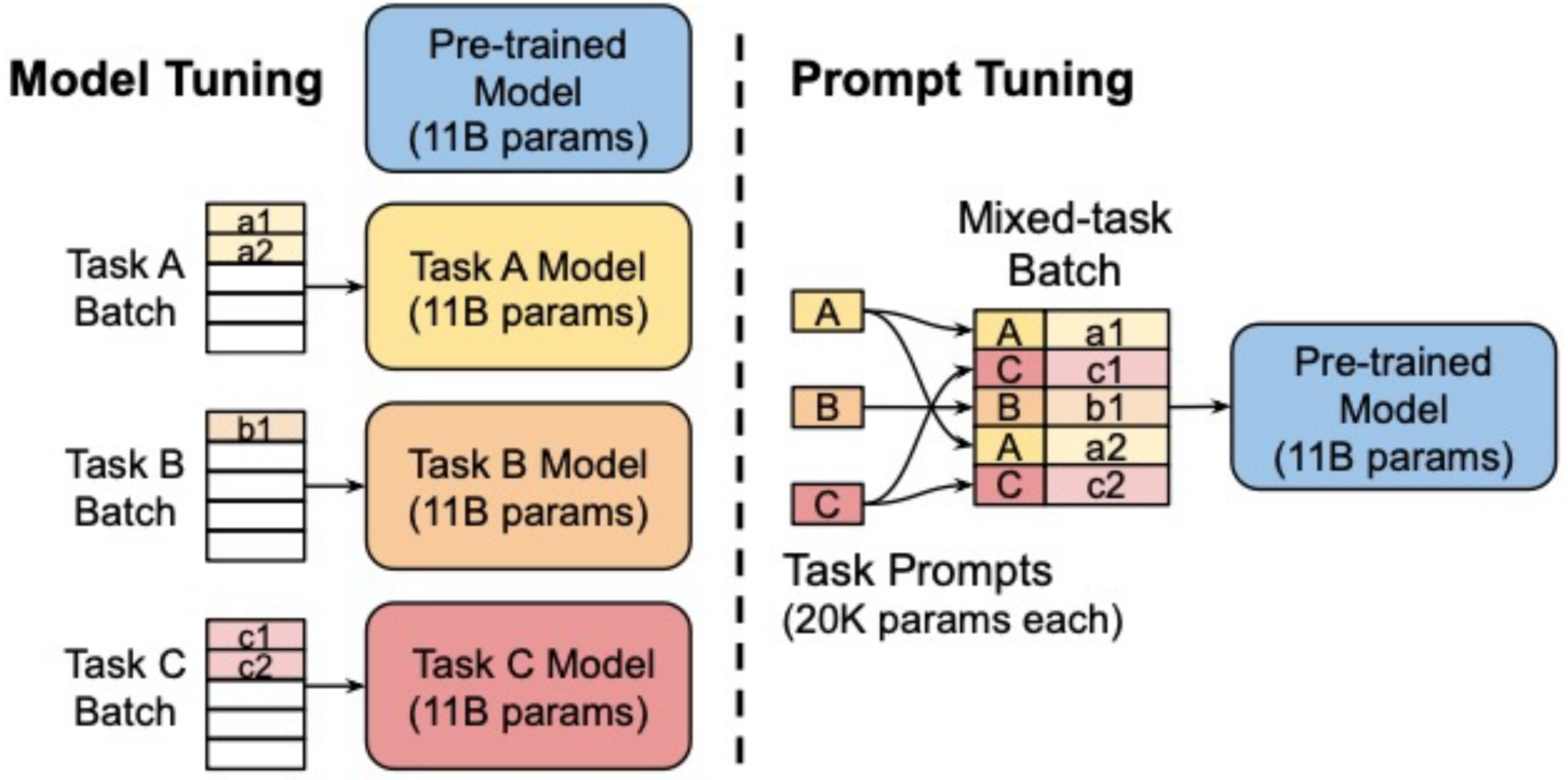
# The Power of Scale for Parameter-Efficient Prompt Tuning

HUMANE Lab
김건수
2025.02.07

# Introduction

- **Adaptation Techniques:**
  - Earlier methods (e.g., ELMo) used frozen pre-trained models with task-specific adjustments.
  - Fine-tuning (as used in GPT and BERT) adjusts all model parameters.
- **Prompt-Based Methods:**
  - GPT-3's prompting shows that text prompts can control model behavior, yet suffers from input leng th limits and human dependency.
- **Motivation for Prompt Tuning:**
  - Need for an efficient, parameter-light method that enables a single frozen model to serve multiple tasks.

# Introduction



# Prompt Tuning

- **Text-to-Text Framework:**
  - Casts all tasks as text generation following T5.
- **Method:**
  - Prepend a set of tunable soft prompt tokens to the input.
  - Only update soft prompt parameters ($\theta_p$), keeping the rest of the model fixed.
- **Design Considerations:**
  - Prompt initialization options: random, vocabulary-based, or class label initialization.
  - Exploration of prompt lengths to balance parameter cost and performance.

# Pre-Training Objective & LM Adaptation

- **T5's Pre-training:**
  - Uses span corruption with sentinel tokens, not ideal for natural text generation.
- **Adaptation Strategies:**

  1. Off-the-shelf span corruption.

  2. Adding a sentinel to downstream targets.

  3. LM Adaptation: Fine-tune T5 with a language modeling objective for more natural outputs.
- **Finding:**
  - LM adaptation significantly improves prompt tuning, especially in mid-sized models.

# Experimental Results

- **SuperGLUE Benchmark:**
  - Experiments conducted across T5 models from Small to XXL.
- **Key Findings:**
  - Prompt tuning becomes more competitive with full model tuning as model size increases.
  - XXL models match multi-task tuned performance with far fewer task-specific parameters.
  - Outperforms GPT-3 few-shot prompting by a significant margin.

# Comparison with Similar Approaches

- **Prefix Tuning vs. Prompt Tuning:**
  - Prefix tuning adds parameters at every transformer layer, whereas prompt tuning only prepends to the input.
- **Other Methods:**
  - WARP, P-tuning, soft words, and adapters modify the network in various ways, often adding comple xity or extra layers.
- **Advantage:**
  - Prompt tuning is highly parameter-efficient (less than 0.01% task-specific parameters in large mode ls) and maintains the original model's architecture.

# Resilience to Domain Shift

- **Frozen Model Advantage:**
  - Freezing the core language model prevents overfitting to specific datasets.
- **Experimental Evidence:**
  - In QA and paraphrase detection tasks, prompt tuning shows improved zero-shot transfer performa nce.
  - Particularly effective when the domain shift is large.

# Prompt Ensemble

- **Concept:**
  - Train multiple prompts for the same task and combine them via ensembling (e.g., majority voting).
- **Benefits:**
  - Shares the frozen core model across all ensembles, drastically reducing storage and inference costs compared to full model ensembles.

# Interpretability & Conclusion

- **Interpretability:**
  - Learned soft prompts form semantically coherent clusters.
  - Class-label initialization helps preserve meaningful token representations.
- **Conclusion:**
  - Prompt tuning is an effective, parameter-efficient method for adapting frozen language models.
  - It scales well, improves domain robustness, and enables efficient multi-task serving and ensemble.