TiTok: *Transfer Token-level Knowledge via Contrastive Excess To Transplant LoRA*

Chanjoo Jung and Jaehyung Kim
chanjoo0427@yonsei.ac.kr / jaehyungk@yonsei.ac.kr
Yonsei University

ICLR · YONSEI UNIVERSITY

Project Page · Code

Overview

Keywords:

 # LoRA knowledge transfer
 # Contrastive excess
 # Token-level transfer

Background & Motivation

- LoRA *cannot* be directly transferred to different models
- Existing methods depend on:
  1) task-specific training data
  2) costly synthetic data filtering pipelines

Our Approach: TiTok

- **Core idea**: Transplanting token-level knowledge from the source model's LoRA by ***selectively training on informative tokens*** in synthetic data.
- Works without any external models
- Robust transfer across models, sizes, and versions

Methodology

Step 1: Synthetic Data Generation

- k seed prompts → source model generates queries ($\mathbf{q}$) & response tokens (y_i)
- Diversity enforced via ROUGE-L < 0.7

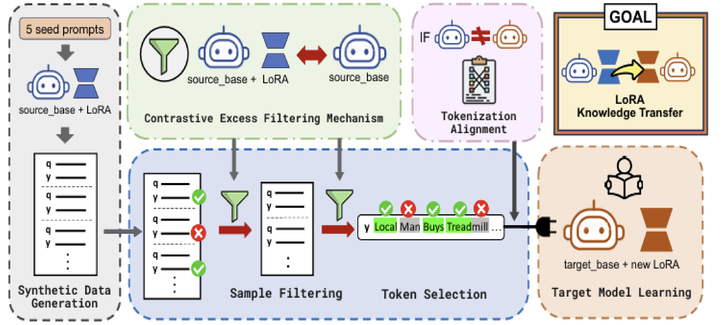
Step 2: Compute Contrastive Excess

- Source model contains two roles:
  1) "amateur" ($\mathcal{M}_s$: base model only)
  2) "expert" ($\mathcal{M}_s + A_s$: base model + LoRA)
- Identify informative tokens via **contrastive excess**, defined as:

$$S(y_i) = L_e(y_i) - L_a(y_i)$$

where:

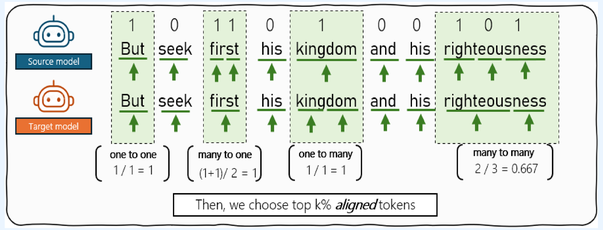
$$L_e(y_i) = \log P_{\mathcal{M}_s+A_s}(y_i \mid \mathbf{q}, \mathbf{y}_{<i}),$$
$$L_a(y_i) = \log P_{\mathcal{M}_s}(y_i \mid \mathbf{q}, \mathbf{y}_{<i})$$



Step 3: Target Model LoRA Training with Filtering

- Filter synthetic samples with top average contrastive excess scores.
- Filter top k% highest contrastive excess scoring tokens in each sample.
- Train target LoRA on the filtered tokens.

Q) How about cases where source and target models have different tokenizers?



- Use dual-pointer approach (0: mask, 1:tokens to train)
- If source model's tokens match without being split:
  → copy the score
- If they don't match (e.g., source tokens are split):
  → compute average score over corresponding source token span & assign to all aligned target tokens

Experiments

Main Results

Table 1: **Main results.** Experiments on BBH, MMLU, News Headline and Scholarly Title Generation tasks under four transfer settings. Reasoning tasks (BBH, MMLU) are evaluated via LM-Eval Harness, and personalization tasks (News Headline, Scholarly Title Generation) via ROUGE-1/L. Zero-shot evaluations are reported as the mean across 3 random seeds. Best scores are in **bold**.

Transfer	Method	BBH	MMLU	News Headline		Scholarly Title	
		Acc.	Acc.	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L
Mistral 7B → Mistral 7B	Vanilla	0.397	0.557	0.117	0.101	0.381	0.311
	KD (+MinED)	0.417	0.560	0.117	0.104	0.385	0.310
	TransLoRA	0.416	0.534	0.156	0.137	0.447	0.382
	TiTok (ours)	**0.424**	**0.561**	**0.161**	**0.143**	**0.473**	**0.413**
Mistral 7B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378
	KD (+MinED)	0.475	0.482	0.127	0.112	0.454	0.387
	TransLoRA	0.473	0.473	0.126	0.110	0.461	0.397
	TiTok (ours)	**0.484**	**0.485**	**0.139**	**0.123**	**0.464**	**0.403**
Llama3 3B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378
	KD (+MinED)	0.474	0.477	0.125	0.110	0.449	0.383
	TransLoRA	0.471	0.467	0.122	0.108	0.454	0.387
	TiTok (ours)	**0.496**	**0.478**	**0.127**	**0.113**	**0.456**	**0.392**
Llama2 7B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378
	KD (+MinED)	0.473	0.476	0.125	0.110	0.449	0.382
	TransLoRA	0.472	0.468	0.123	0.109	0.457	0.394
	TiTok (ours)	**0.488**	**0.477**	**0.138**	**0.120**	**0.461**	**0.403**

Ablation Study

Settings		BBH	MMLU	News Headline		Scholarly Title	
Sample filtering	Token selection	Acc.	Acc.	R-1	R-L	R-1	R-L
✗	✗	0.458	0.485	0.133	0.117	0.456	0.393
✗	✓	0.463	0.496	0.137	0.121	0.460	0.397
✓	✗	0.470	0.500	0.139	0.122	0.460	0.397
✓	✓	**0.483**	**0.501**	**0.142**	**0.125**	**0.464**	**0.403**

Strong Generalization to External Data

Data Setting	Method	News Headline		Scholarly Title	
		R-1	R-L	R-1	R-L
Other-user	Vanilla	0.117	0.101	0.381	0.311
	KD	0.127	0.113	0.405	0.333
	TiTok (ours)	**0.151**	**0.136**	**0.481**	**0.425**
Mixed-user	Vanilla	0.117	0.101	0.381	0.311
	MinED	0.124	0.110	0.409	0.337
	TiTok (ours)	**0.151**	**0.137**	**0.480**	**0.422**
Cross-task	Vanilla	0.117	0.101	0.381	0.311
	MinED	0.118	0.106	0.403	0.331
	TiTok (ours)	**0.133**	**0.120**	**0.450**	**0.383**