Making Pretrained Language Models Good Long-tailed Learners

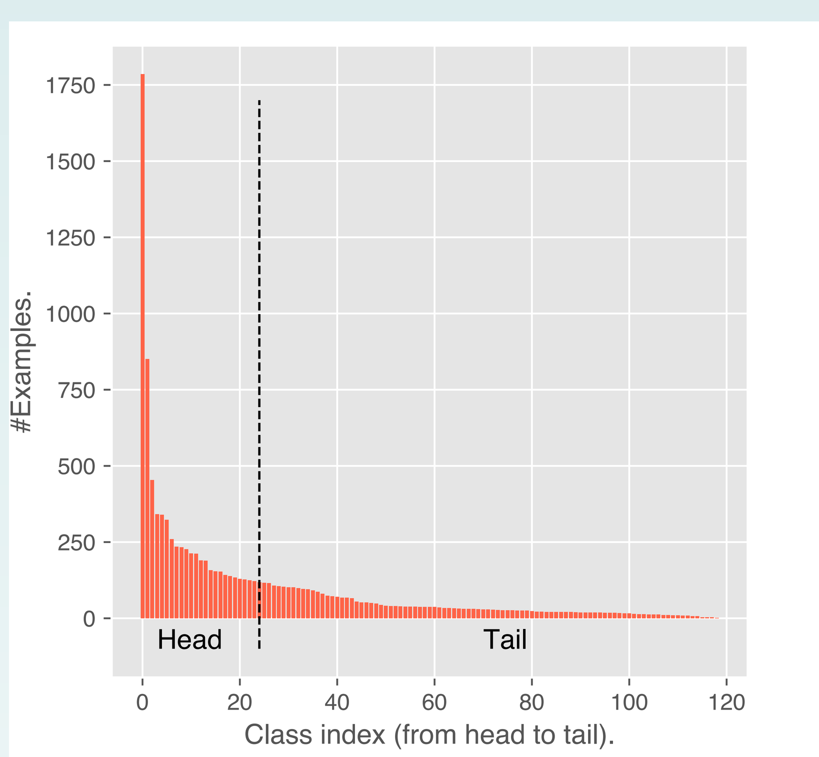
Abbreviated as 🎉 Glee

Chen Zhang[1], Lei Ren[2], Jingang Wang[2*], Wei Wu[2], Dawei Song[1*]

[1]Beijing Institute of Technology, [2]Meituan NLP, *Corresponding Author

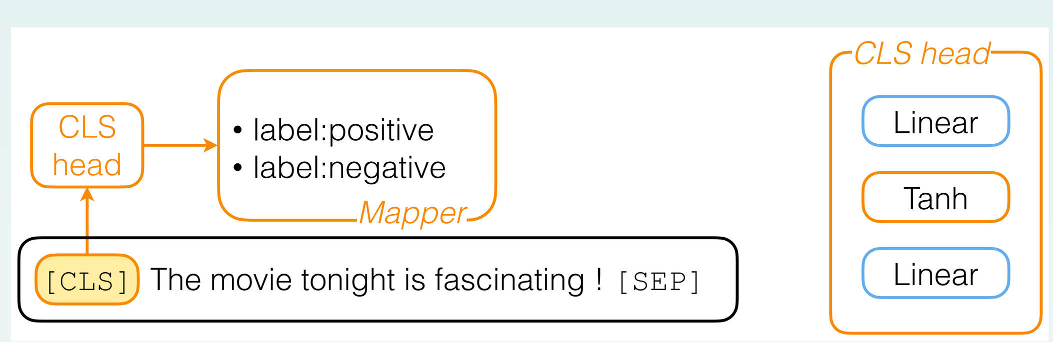
Motivation

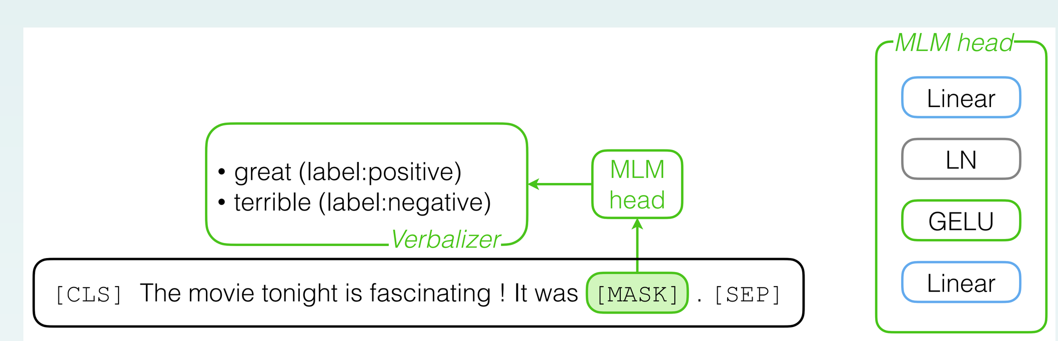
- Tail bottleneck
  - Tail classes, which are with very few examples, in a long-tailed class distribution prevents PLMs from achieving good performance.
- Head-to-tail transfer
  - Tail classes are intuitively few-shot ones. However, long-tailed classification allows the possibility to transfer knowledge from head classes to tail ones.
- Prompt-tuning makes PLMs better few-shot learners
  - It motivates us to hypothesize that Prompt-tuning can relieve the tail bottleneck and thus make PLMs at least good long-tailed learners.



Background

- Long-tailed classification
  - Dataset $\mathscr{D} = \{(x_i, y_i)\}_i$, where $(x, y) \sim P(\mathscr{X}, \mathscr{Y})$.
  - $P(\mathscr{Y})$ is a long-tailed one.
  - PLM $\mathscr{M}$ is hard-to-optimize on $\mathscr{D}$.
- CLS-tuning
  - Input: x; Output: y.
  - Backbone $\mathscr{E}$: [CLS] vector.
  - Classifier $\mathscr{C}$: a Tanh-activated MLP, CLS head.
  - Objective $\mathbb{L}_{\text{CLS}} = \mathbb{E}_{\mathscr{D}} - \log P(y \mid x; \mathscr{M})$.
- Prompt-tuning
  - Input: $\mathscr{T}(x)$; Output: $\mathscr{V}(x)$.
  - Backbone $\mathscr{E}$: [MASK] vector.
  - Classifier $\mathscr{C}$: pretrained MLM head.
  - Objective: $\mathbb{L}_{\text{CLS}} = \mathbb{E}_{\mathscr{D}} - \log P(\mathscr{V}(y) \mid \mathscr{T}(x); \mathscr{M})$.



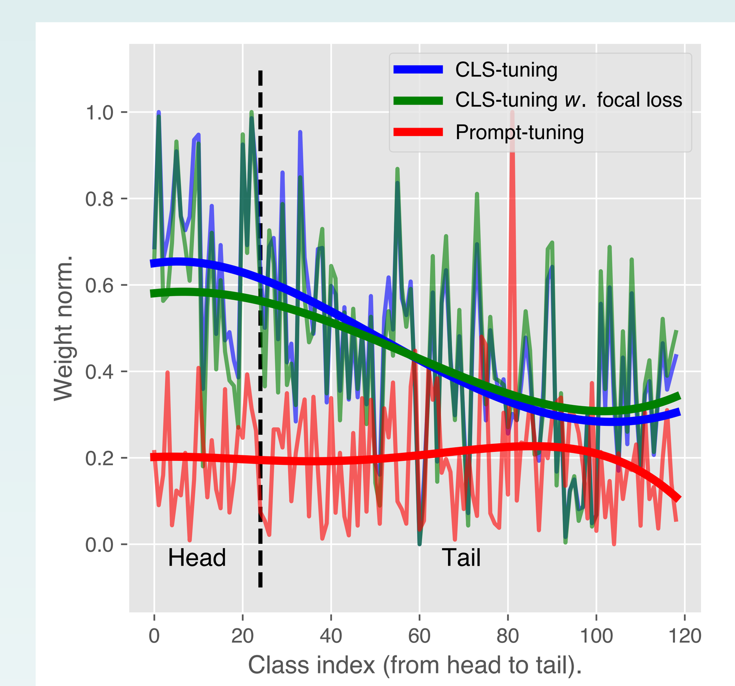


Setup

- Long-tailed datasets
  - Medical Question Intent (Cmid), Application Category (Iflyteck), Clinical Trial Criterion (Ctc), Entity Typing (Msra), Document Topic (R52).
- Baselines
  - CLS-tuning, w/ η-norm, w/ focal loss, w/ prompt, w/ LN, w/ pt. (pretrained) LN.
  - Prompt-tuning, w/ focal loss, w/ ed. (Embedding decoupling).
- Metrics
  - Accuracy scores: tail insensitive, F1 scores: tail sensitive, Head F1 scores for head classes, Tail F1 scores for tail classes.

Results & Analyses

- Results
  - Bottom left: Prompt-tuning largely outperforms CLS-tuning and calibrated CLS-tuning (e.g. CLS-tuning w/ focal loss) mainly due to the improved tail performance.
  - Bottom right: Weight norm visualization indicates that Prompt-tuning learns a better balance between head and tail classes.
- Research questions (bottom right)
  - RQ1: Does the shared embedding contribute to Prompt-tuning?
    - Prompt-tuning w/ ed. decreases the performance. => negative response.
  - RQ2: Does the input structure (i.e., MLM input) contribute to Prompt-tuning?
    - CLS-tuning w/ prompt decreases the performance. => negative response.
  - RQ3: Does the classifier structure and parameterization (e.g., layer normalization used in MLM head) contribute to Prompt-tuning?
    - CLS-tuning w/ LN and w/ pt. LN increases the performance. => positive response.



Dataset	Cmid		Iflytek		Ctc		Msra		R52		Avg	
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CLS-tuning	$51.1_{0.4}$	$37.3_{2.3}$	$58.7_{0.4}$	$33.7_{1.6}$	$84.6_{0.3}$	$77.2_{2.9}$	$99.0_{0.1}$	$97.5_{1.0}$	$95.3_{0.2}$	$67.3_{1.3}$	77.7	62.6
w/ η-norm	$51.1_{0.5}$	$37.4_{2.0}$	$59.1_{0.3}$	$35.7_{1.6}$	$84.7_{0.2}$	$77.3_{3.1}$	$99.0_{0.1}$	$97.4_{0.9}$	$95.4_{0.3}$	$68.9_{1.9}$	**77.9**	63.3
w/ focal loss	$51.0_{0.7}$	$42.1_{1.3}$	$58.8_{0.3}$	$36.0_{1.6}$	$84.3_{0.4}$	$78.5_{2.4}$	$99.0_{0.1}$	$96.8_{1.2}$	$95.7_{0.2}$	$72.8_{2.3}$	77.8	65.2
Prompt-tuning	$49.3_{0.7}$	$43.4_{0.7}$	$61.2_{0.6}$	$44.4_{1.0}$	$84.2_{0.1}$	$80.9_{0.1}$	$99.1_{0.0}$	$97.8_{0.3}$	$95.7_{0.1}$	$85.3_{0.6}$	**77.9**	**70.4**
w/ focal loss	48.6_{06}	$42.5_{0.6}$	$59.7_{0.6}$	$43.9_{0.7}$	$83.5_{0.6}$	$80.2_{0.7}$	$99.0_{0.1}$	$97.2_{0.7}$	$95.5_{0.3}$	$82.6_{2.4}$	77.3	69.3
Metric	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail
CLS-tuning	$50.3_{1.0}$	$34.1_{3.0}$	$61.8_{0.6}$	$27.4_{1.9}$	$87.7_{0.2}$	$74.1_{3.7}$	$99.2_{0.1}$	$97.4_{1.1}$	$99.0_{0.1}$	$66.6_{1.3}$	79.6	59.9
w/ η-norm	$50.3_{0.9}$	$34.3_{2.7}$	$62.1_{0.4}$	$29.7_{2.0}$	$87.8_{0.2}$	$74.3_{4.0}$	$99.2_{0.1}$	$97.3_{1.0}$	$99.0_{0.1}$	$68.3_{1.9}$	**79.7**	60.8
w/ focal loss	$49.8_{0.8}$	$40.2_{1.5}$	$62.0_{0.4}$	$30.2_{1.9}$	$87.5_{0.3}$	$75.9_{3.1}$	$99.3_{0.0}$	$96.7_{1.3}$	$99.0_{0.0}$	$72.3_{2.4}$	79.5	63.1
Prompt-tuning	$48.4_{1.0}$	$42.2_{0.7}$	$63.6_{0.4}$	$40.1_{1.2}$	$87.4_{0.2}$	$79.0_{0.2}$	$99.2_{0.1}$	$97.7_{0.3}$	$98.6_{0.1}$	$85.0_{0.6}$	79.4	**68.8**
w/ focal loss	$47.1_{0.8}$	$41.4_{0.8}$	$62.3_{0.6}$	$39.8_{0.8}$	$86.7_{0.5}$	$78.3_{0.7}$	$99.3_{0.1}$	$97.1_{0.7}$	$98.8_{0.1}$	$82.3_{2.4}$	78.8	67.8

Dataset	Cmid		Iflytek		Ctc		Msra		R52		Avg	
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CLS-tuning ∘ T	$51.1_{0.4}$	$37.3_{2.3}$	$58.7_{0.4}$	$33.7_{1.6}$	$84.6_{0.3}$	$77.2_{2.9}$	$99.0_{0.1}$	$97.5_{1.0}$	$95.3_{0.2}$	$67.3_{1.3}$	77.7	62.6
w/ η-norm	$51.1_{0.5}$	$37.4_{2.0}$	$59.1_{0.3}$	$35.7_{1.6}$	$84.7_{0.2}$	$77.3_{3.1}$	$99.0_{0.1}$	$97.4_{0.9}$	$95.4_{0.3}$	$68.9_{1.9}$	77.9	63.3
w/ focal loss	$51.0_{0.7}$	$42.1_{1.3}$	$58.8_{0.3}$	$36.0_{1.6}$	$84.3_{0.4}$	$78.5_{2.4}$	$99.0_{0.1}$	$96.8_{1.2}$	$95.7_{0.2}$	$72.8_{2.3}$	77.8	65.2
CLS-tuning ∘ R	$50.9_{0.4}$	$34.5_{1.4}$	$58.7_{0.3}$	$33.3_{1.1}$	$84.4_{0.4}$	$77.1_{1.0}$	$99.0_{0.1}$	$97.7_{0.5}$	$94.2_{0.4}$	$56.2_{2.5}$	77.4	59.8
w/ η-norm	$50.9_{0.5}$	$34.8_{1.8}$	$58.4_{0.3}$	$33.3_{1.0}$	$84.6_{0.4}$	$78.0_{1.5}$	$99.1_{0.0}$	$97.8_{0.5}$	$94.3_{0.3}$	$56.3_{1.9}$	77.5	60.0
w/ focal loss	$51.0_{0.5}$	$40.1_{1.5}$	$58.8_{0.4}$	$34.6_{0.1}$	$84.6_{0.3}$	$76.9_{0.6}$	$99.0_{0.1}$	$97.0_{1.5}$	$95.1_{0.3}$	$66.0_{2.7}$	77.7	62.9
CLS-tuning ∘ R												
w/ prompt	$49.7_{0.5}$	$33.1_{0.4}$	$58.4_{0.3}$	$32.8_{1.0}$	$84.6_{0.1}$	$77.2_{3.0}$	$99.0_{0.1}$	$96.9_{0.3}$	$94.1_{0.3}$	$54.5_{3.1}$	77.2	58.9
w/ LN	$51.3_{0.6}$	$42.0_{1.4}$	$59.7_{0.6}$	$39.1_{0.8}$	$84.6_{0.5}$	$79.4_{2.2}$	$99.1_{0.1}$	$97.1_{0.8}$	$96.1_{0.2}$	$77.7_{3.5}$	**78.2**	67.1
w/ pt. LN	$50.8_{0.6}$	$42.5_{1.2}$	$59.4_{0.4}$	$41.4_{0.9}$	$84.4_{0.5}$	$79.7_{1.5}$	$99.1_{0.1}$	$97.7_{0.5}$	$96.2_{0.2}$	$82.0_{1.8}$	78.0	68.7
Prompt-tuning	$49.3_{0.7}$	$43.4_{0.7}$	$61.2_{0.6}$	$44.4_{1.0}$	$84.2_{0.1}$	$80.9_{0.1}$	$99.1_{0.0}$	$97.8_{0.3}$	$95.7_{0.1}$	$85.3_{0.6}$	77.9	**70.4**
w/ ed.	$49.4_{0.7}$	$43.6_{0.7}$	$61.0_{0.7}$	$44.4_{1.0}$	$84.2_{0.4}$	$80.5_{0.9}$	$99.0_{0.2}$	$96.9_{1.4}$	$95.7_{0.2}$	$84.9_{1.0}$	77.9	70.1

Conclusions

- Prompt-tuning essentially makes pretrained language models good long-tailed learners.
- Through in-depth analyses, we uncover that the structure and parameterization are the key to enhancing long-tailed performance of pretrained language models.
- The finding may shed light on the design of Prompt-tuning.