PETapter: Leveraging PET-style classification heads for modular few-shot parameter-efficient fine-tuning

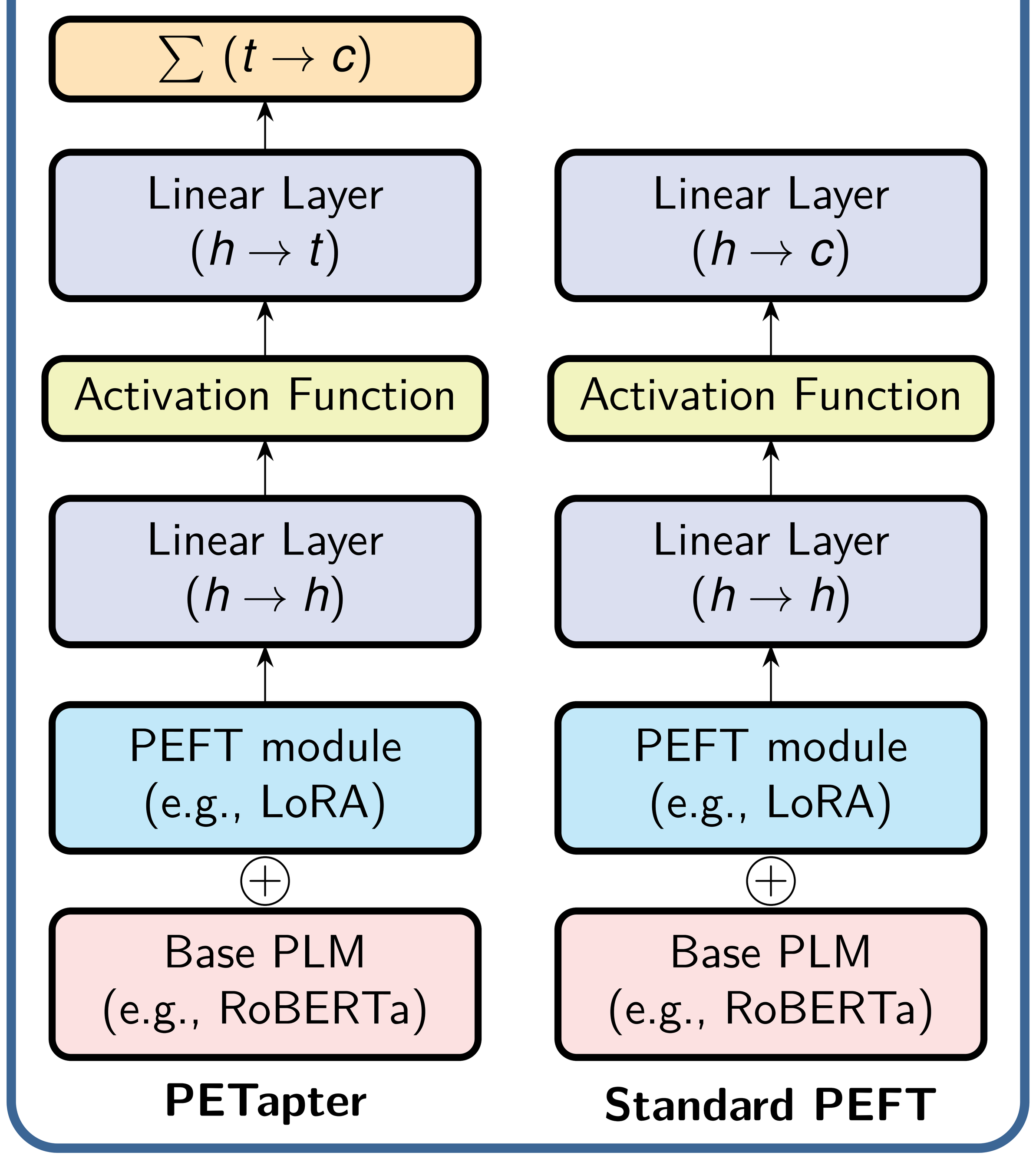
Jonas Rieger, Mattes Ruckdeschel, Gregor Wiedemann

Parameter Efficient Fine-Tuning (PEFT) | Pattern-Exploiting Training (PET) | Adapters | Few-Shot Learning | Efficiency | Modularity | Sharibility | Reliability

Summary

- combination of advantages from few-shot learning and PEFT
- performance on par with PET
- no catastrophic forgetting
- fast training in low resource settings
- less disk space needed for sharing (16 MB vs. 2.1 GB)
- Ukraine: German PVPs perform worse
- Ukraine: zero-shot GPT-4 performs better than using in-context learning (max. 10 obs.)
- Ukraine: label unbalancedness is represented in class performances
- Bagging + majority vote is useful

PETapter Architecture



PETapter

- $\mathcal{L}$ set of target labels with $|\mathcal{L}| = c$
- $P^m(x)$ be a pattern function inserting m [MASK] tokens to an input x
- $v^m(\ell)$ injective function mapping $\ell \in \mathcal{L}$ to m vocabulary tokens
- $T = \bigcup_{\ell \in \mathcal{L}} \bigcup_{i=1}^{m} v^m(\ell)_i$ relevant vocabulary
- $v^m : \mathcal{L} \to T^m$
- logits for the m [MASK] tokens: $M(v^m(\ell) \mid P^m(x)) \in \mathbb{R}^m$
- final score per label ℓ for an input x: $s(\ell \mid x) = \sum_{i=1}^{m} M(v^m(\ell) \mid P^m(x))_i$
- corresponding pseudo-probability: $q(\ell \mid x) = \frac{\exp(s(\ell \mid x))}{\sum_{\ell' \in \mathcal{L}} \exp(s(\ell' \mid x))}$
- cross-entropy loss over all observations:

$$L_{\mathrm{CE}} = - \sum_{(x,\ell^*)} \sum_{\ell \in \mathcal{L}} \mathbb{1}_{\{\ell = \ell^*\}}(x, \ell^*) \log[q(\ell \mid x)]$$
$$= - \sum_{(x,\ell^*)} \log[q(\ell^* \mid x)]$$

Pattern-Verbalizer-Pair (PVP)

Yelp Q&A PVP:

Pattern [text]* [SEP] Question: What does the customer think of this restaurant? Answer: [MASK].

Verbalizers

class label	verbalizer
1 star	terrible
2 stars	bad
3 stars	okay
4 stars	good
5 stars	great

Ukraine PVP:

Pattern This sentence contains [MASK] [MASK] arms deliveries to Ukraine: {[target_sentence] [SEP] [context_before] [SEP] [context_after]}*

Verbalizers argument against, argument for, claim against, claim for

Benchmark Study (Accuracy ± Std.)

RoBERTa Large		Prompt Pattern			Q&A Pattern			Linear Layer	
		PETapter			PETapter				
n	Data	LoRA	Pfeif.	PET	LoRA	Pfeif.	PET	LoRA	Pfeif.
10	AG	0.714 ±.070	0.702 ±.081	**0.842 ±.025**	0.746 ±.054	0.738 ±.060	0.836 ±.032	0.373 ±.049	0.443 ±.104
10	Yahoo	0.331 ±.040	0.290 ±.056	**0.574 ±.030**	0.365 ±.049	0.346 ±.054	0.550 ±.040	0.150 ±.027	0.169 ±.041
10	Yelp	0.470 ±.041	0.479 ±.035	0.475 ±.026	0.472 ±.049	**0.490 ±.046**	0.486 ±.041	0.221 ±.012	0.216 ±.014
100	AG	0.873 ±.010	0.875 ±.010	**0.877 ±.009**	0.870 ±.010	0.873 ±.010	0.874 ±.009	0.875 ±.008	0.875 ±.008
100	Yahoo	0.662 ±.014	0.661 ±.017	**0.680 ±.013**	0.654 ±.008	0.656 ±.012	0.675 ±.013	0.648 ±.016	0.647 ±.015
100	Yelp	0.613 ±.014	0.614 ±.010	0.593 ±.014	0.622 ±.013	0.620 ±.013	0.595 ±.016	0.551 ±.019	0.512 ±.043

Training times for $n = 100$ observations

RoBERTa	Architecture	AG	Yahoo	Yelp
Base	PETapter	0.33	0.33	0.32
Base	PET	0.38	0.39	0.39
Large	PETapter	0.65	0.64	0.65
Large	PET	1.00	1.00	1.00
Large	PET	6.1s	6.2s	6.2s

Ukraine Study (Macro-F1 ± Std.)

		PETapter			Linear Layer	
n	Samp.	LoRA	Pfeiffer	PET	LoRA	Pfeiffer
10	Equal	0.31 ±.043	**0.33 ±.057**	0.33 ±.080	0.13 ±.042	0.15 ±.041
10	Strat.	0.27 ±.039	0.33 ±.027	**0.40 ±.055**	0.16 ±.001	0.17 ±.021
100	Equal	0.57 ±.020	0.57 ±.028	**0.59 ±.027**	0.26 ±.029	0.29 ±.030
100	Rand.	**0.56 ±.036**	0.55 ±.036	0.56 ±.053	0.20 ±.041	0.26 ±.037
100	Strat.	0.58 ±.042	0.57 ±.035	**0.59 ±.054**	0.20 ±.030	0.26 ±.035
250	Equal	0.67 ±.014	0.68 ±.018	**0.70 ±.025**	0.46 ±.050	0.49 ±.075
250	Rand.	**0.67 ±.021**	0.67 ±.024	0.67 ±.109	0.38 ±.031	0.45 ±.086
250	Strat.	0.67 ±.019	**0.67 ±.018**	0.67 ±.109	0.37 ±.040	0.46 ±.082

PVP Study LoRA-PETapter (Macro-F1 ± Std.)

		No Pattern			Pattern		
n	Samp.	Alpha	Normal	Shuffle	Alpha	Normal	Shuffle
10	Equal	0.22 ±.039	0.25 ±.040	0.22 ±.039	0.23 ±.039	0.31 ±.043	0.28 ±.043
10	Strat.	0.20 ±.025	0.22 ±.031	0.22 ±.033	0.23 ±.067	0.27 ±.039	0.27 ±.043
100	Equal	0.47 ±.026	0.43 ±.041	0.41 ±.037	0.57 ±.041	0.57 ±.020	0.56 ±.033
100	Rand.	0.43 ±.046	0.39 ±.040	0.39 ±.043	0.53 ±.035	0.56 ±.036	0.54 ±.044
100	Strat.	0.40 ±.027	0.38 ±.035	0.37 ±.027	0.52 ±.051	0.58 ±.042	0.54 ±.048
250	Equal	0.62 ±.022	0.61 ±.024	0.60 ±.022	0.67 ±.021	0.67 ±.014	0.68 ±.019
250	Rand.	0.58 ±.035	0.57 ±.054	0.56 ±.054	0.65 ±.029	0.67 ±.021	0.66 ±.020
250	Strat.	0.60 ±.025	0.59 ±.030	0.58 ±.032	0.65 ±.019	0.67 ±.019	0.66 ±.017

LEIBNIZ INSTITUTE FOR MEDIA RESEARCH HANS-BREDOW-INSTITUT

tu dortmund university

FLACA (BMBF project no. 16DKWN064B)
FAME (DFG project no. 406289255)
Trustworthy performance evaluation of large language models (RC Trust)
GitHub Repository: `https://github.com/JonasRieger/petapter`