Quantum-PEFT: Ultra parameter-efficient fine-tuning

Toshiaki Koike-Akino(1,2), Francesco Tonin(2), Yongtao Wu(2), Leyla Naz Candogan(2), Volkan Cevher(2)

(1)Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA

(2)Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

Introduction

- Parameter-efficient fine-tuning (**PEFT**) is a cost-effective framework for downstream task to specialize pre-trained large foundation models.
- Low-rank adaptation (**LoRA**)[1] variants achieve excellent performance.
- We propose a novel framework, named **Quantum-PEFT**, that leverages quantum unitary parameterizations, achieving orders-of-magnitudes higher compression rates over state-of-the-art PEFT methods.

Quantum-Inspired Machine Learning

- Quantum machine learning (QML) is an emerging framework leveraging quantum processing units (QPUs) for AI tasks.
- QML realizes ultra-efficient operations due to exponential expressivity.
- We introduce generalized QML framework based on alternating RY/CZ simplified two-design ansatz[2].

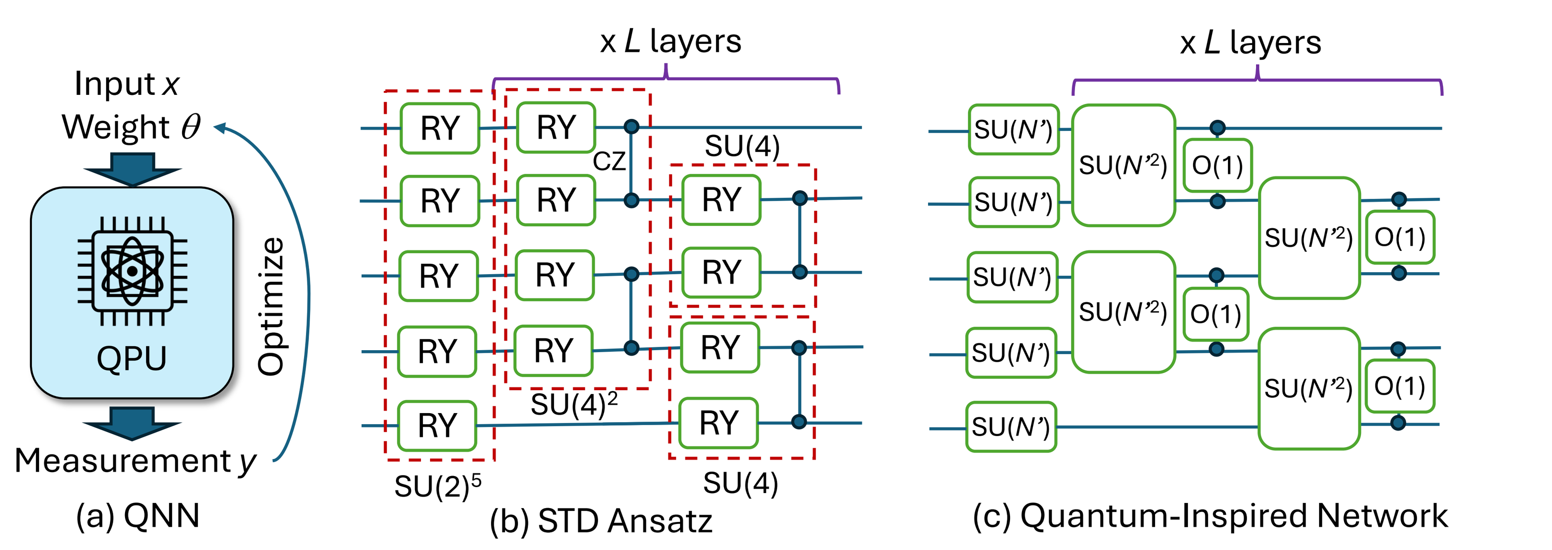

Figure: QML: (a) General pipeline for quantum neural network (QNN), embedding classical data x and variational parameters θ to control measurement y. (b) Simplified two-design ansatz. (c) Generalized quantum-inspired network.

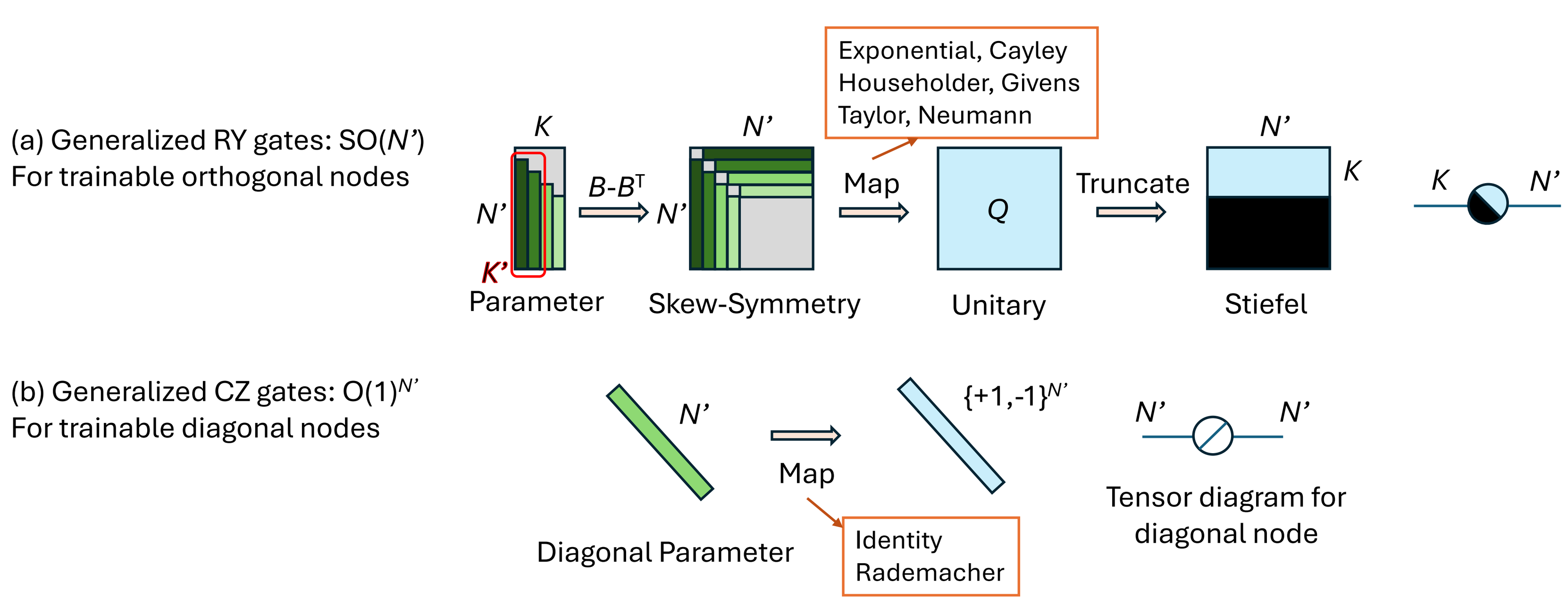

Figure: Proposed modules with corresponding tensor diagrams: (a) generalized RY modules for orthogonal nodes on Stiefel manifold $\mathcal{V}_K(N')$; (b) generalized CZ modules for diagonal nodes on either $\mathrm{O}(1)^{N'}$ or $\mathbb{R}^{N'}$. Top K' columns are trainable parameters in B as intrinsic rank.

Quantum-PEFT

- Pauli parameterization enables logarithmically fewer number of trainable parameters.

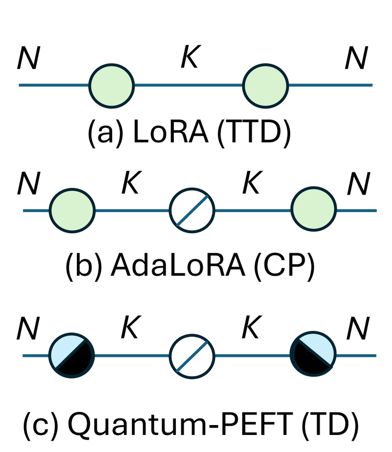

Figure: Tensor diagram of LoRA variants.

Table: Comparison of different PEFT methods and their computational requirements.

Method	# Trainable Parameters
LoRA (TTD)	$2NK$
AdaLoRA (CP)	$2NK+K$
Quantum-PEFT (TD: Q_T)	$2NK-K^2$
Quantum-PEFT (TD: Q_P)	$2(2L+1)\log_2(N)+K$

Mixed-Precision Tensor Network

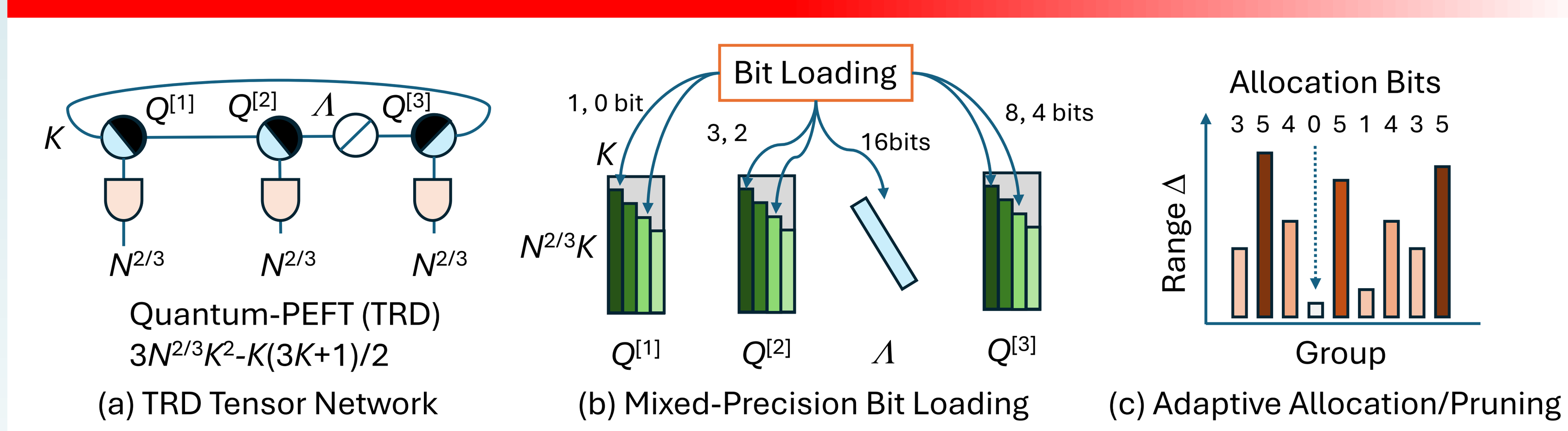

Figure: Mixed-precision Quantum-PEFT in 3-dimensional TRD tensor network. Each tensor node and tensor parameter can have non-uniform bit assignments. Adaptive bit loading depends on group range Δ. Assignment of 0 bit corresponds to adaptive structural pruning.

Experiments

- **3 transfer learning tasks**: LLM GLUE benchmark[4]; E2E challenge[5]; ImageNet to CIFAR10 classification.
- **3 foundation models**: DeBERTaV3[6]; GPT2 Medium[7]; ViT[8].
- **5 baseline methods**: LoRA[1]; AdaLoRA[3]; BitFit[9]; HAdapter[10]; PAdapter[11].

Results

- Quantum-PEFT shows competitive performance with extremely fewer number of trainable parameters.
- Quantization and mixed-precision Quantum-PEFT keep good performance over full-precision PEFT.

Table: Results with DeBERTaV3 base on GLUE benchmark. We present the Matthew's correlation for CoLA, the average correlation for STS-B, and the accuracy for other tasks. In each column, the best-performing PEFT approach is highlighted in **bold** and the second best is <u>underlined</u>.

Method	# Trainable Parameters	SST-2	CoLA	RTE	MRPC	STS-B
FT	184M	95.63	69.19	83.75	89.46	91.60
BitFit	0.1M	94.84	66.96	78.70	87.75	91.35
HAdapter	0.61M	95.30	67.87	85.56	89.22	91.30
PAdapter	0.60M	95.53	<u>69.48</u>	84.12	89.22	91.52
HAdapter	0.31M	95.41	67.65	83.39	89.25	91.31
PAdapter	0.30M	94.72	69.06	84.48	89.71	91.38
LoRA	0.33M	94.95	68.71	85.56	89.71	**91.68**
AdaLoRA	0.32M	<u>95.80</u>	**70.04**	**87.36**	<u>90.44</u>	<u>91.63</u>
Quantum-PEFT	**0.013M**	**95.85**	67.85	<u>86.57</u>	**90.78**	91.06

Table: Results for different adaptation methods on the E2E benchmark and GPT2 Medium model. Quantum-PEFT achieves similar performance as LoRA with 4 times less trainable parameters.

Method	# Trainable Parameters	BLEU	NIST	METEOR	ROUGE-L	CIDEr
FT	354.92M	68.2	8.62	46.2	71.0	2.47
AdaLoRA	0.38M	64.64	8.38	43.49	65.90	2.18
LoRA	0.39M	66.88	8.55	**45.48**	**68.40**	**2.31**
Quantum-PEFT	**0.098M**	**67.46**	**8.58**	<u>45.02</u>	<u>67.36</u>	**2.31**

Table: Results for ViT transfer learning from ImageNet-21k to CIFAR10. Base ViT is quantized with 3 bits.

Method	Original	FT	LoRA$_{K=1}$	LoRA$_{K=2}$	LoRA$_{K=4}$	Quantum-PEFT
# Parameters	—	85.81M	0.037M	0.074M	0.147M	**0.007M**
Accuracy	76.21%	98.05%	98.14%	98.30%	98.39%	**98.46%**

Table: Quantization impact on Lie parameters with Taylor parameterization for ViT transfer learning from ImageNet-21k to CIFAR10. Base ViT is not quantized.

Quantization	FP32	INT8	INT4	INT3	INT2	INT1
# Bits per parameter	32	8.25	4.25	3.25	2.25	1.25
Accuracy (Uniform Bit Loading)	98.81%	**98.79%**	98.78%	98.75%	98.67%	97.96%
Accuracy (Adaptive Bit Loading)	98.81%	98.78%	**98.87%**	**98.80%**	**98.77%**	**98.64%**

References

[1] E.J. Hu, et al., "LoRA: Low-rank adaptation of large language models," ICLR, 2021.

[2] M. Cerezo, et al., "Cost function dependent barren plateaus in shallow parametrized quantum circuits," Nature communications, 12(1):1791, 2021.

[3] Q. Zhang, et al., "Adaptive budget allocation for parameter-efficient fine-tuning," ICLR, 2023.

[4] A. Wang, et al., "GLUE: A multi-task benchmark and analysis platform for natural language understanding," ICLR, 2019.

[5] J. Novikova, et al., "The E2E dataset: New challenges for end-to-end generation."

[6] P. He, et al., "DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing."

[7] A. Radford, et al., "Language models are unsupervised multitask learners." OpenAI blog, 1(8):9, 2019.

[8] A. Dosovitskiy, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2020.

[9] E.B. Zaken, et al., "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," Annual Meeting of the Association for Computational Linguistics, 2022.

[10] N. Houlsby, et al., "Parameter-efficient transfer learning for NLP," ICML, 2019.

[11] J. Pfeiffer, et al., "AdapterFusion: Non-destructive task composition for transfer learning," EACL 2021.