Quaternion Self-Attention with Shared Scores

Shogo Yamauchi,[1] Tohru Nitta,[2] Hideaki Tamori[1]
[1] The Asahi Shimbun Company, [2] Tokyo Woman's Christian University

ICML International Conference On Machine Learning

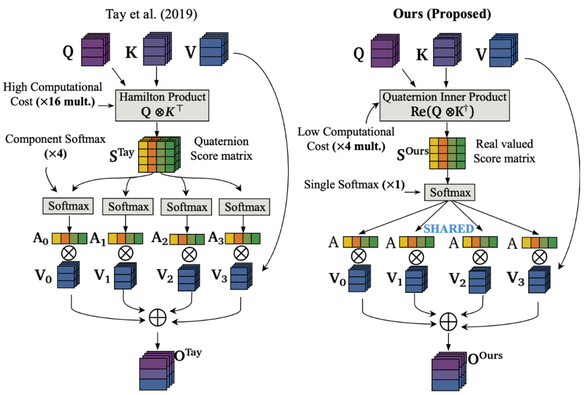
朝日新聞 Media R&D Center

Motivation

Quaternion: $q = q_0 + q_1\mathrm{i} + q_2\mathrm{j} + q_3\mathrm{k}$.

- Quaternion Neural Networks (QNNs) are attractive for parameter efficiency.
- Quaternion self-attention computes 4 separate scores (r, i, j, k), each with its own softmax (Tay et al.) [1] — increasing computation and slowing inference.
- Question: are the four scores really independent? → They are not.

Key Idea / Proposed Method



Key Idea

- Tay et al. computes 4 scores from (Q, K), each with its own softmax.
- We replace them with one real-valued score via the quaternion inner product.
- This single attention map is shared across all four V components.

Tay et al. Attention

$$S^{\text{Tay}} = \frac{1}{\sqrt{d_h}} Q \otimes K^\top = S_0 + S_1 i + S_2 j + S_3 k$$

$$A_\alpha = \mathrm{softmax}(S_\alpha^{\text{Tay}}) \quad \alpha \in \{0,1,2,3\}$$

$$O^{\text{Tay}} = A_0V_0 + A_1V_1 i + A_2V_2 j + A_3V_3 k$$

Hamilton Product $Q \otimes K^\top$

$$S_0 = Q_0K_0^\top - Q_1K_1^\top - Q_2K_2^\top - Q_3K_3^\top$$
$$S_1 = Q_0K_1^\top + Q_1K_0^\top + Q_2K_3^\top - Q_3K_2^\top$$
$$S_2 = Q_0K_2^\top - Q_1K_3^\top + Q_2K_0^\top + Q_3K_1^\top$$
$$S_3 = Q_0K_3^\top + Q_1K_2^\top - Q_2K_1^\top + Q_3K_0^\top$$

Ours Attention

$$S = \frac{1}{\sqrt{4d_h}} \mathrm{Re}(Q \otimes K^\dagger), \quad K^\dagger := K^{*\top}$$

$$A = \mathrm{softmax}(S)$$

$$O = AV_0 + AV_1 i + AV_2 j + AV_3 k$$

Quaternion Inner Product

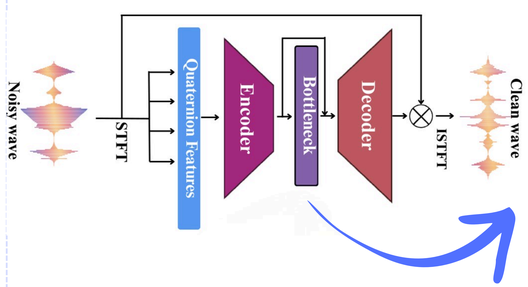
$$\mathrm{Re}(Q \otimes K^*) = Q_0K_0 + Q_1K_1 + Q_2K_2 + Q_3K_3$$

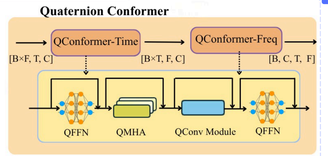
- Reduce score-computation multiplications by 75%
- Reduce softmax operations from 4 to 1

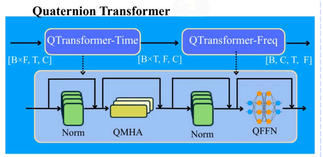
T	Ours		Tay et al. (2019)		Speedup
	MACs	Time [ms]	MACs	Time [ms]	
512	134M	1.210	336M	1.536	1.27×
1024	537M	1.842	1.34G	3.038	1.65×
2048	2.15G	5.251	5.37G	9.780	1.86×
4096	8.59G	15.500	21.5G	32.231	2.08×

Speedup grows with sequence length — 1.27× at T=512 → 2.08× at T=4096.

Experiments / Speech Enhancement



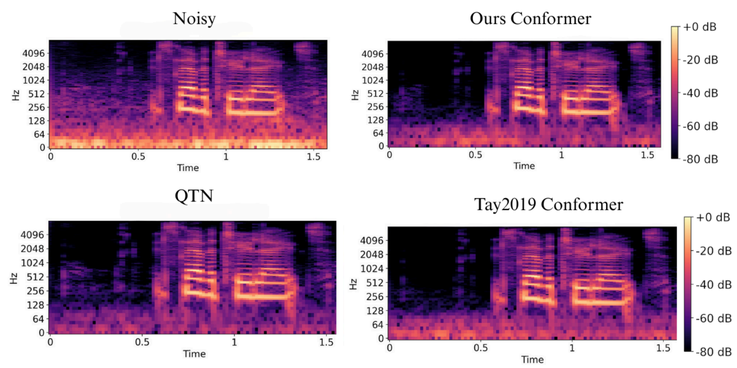




Result on VoiceBank+DEMAND

Method	Attention Type	Params (M)	PESQ	SI-SDR
Real-valued Conformer	Standard	3.19	3.25	19.09
Quaternion				
QCNN	–	0.74	2.80	18.42
QTN (Yang et al., 2023)	QConv-based attention	0.63	2.76	19.55
QTransformer (Tay et al., 2019)	Hamilton	0.62	3.07	19.62
QTransformer (Ours)	Shared Score	0.62	3.07	**19.69**
QConformer (Tay et al., 2019)	Hamilton	0.80	3.11	19.64
QConformer (Ours)	Shared Score	0.80	**3.18**	19.36

- Our QConformer reaches PESQ 3.18 — best among all quaternion models.



Paper

Code

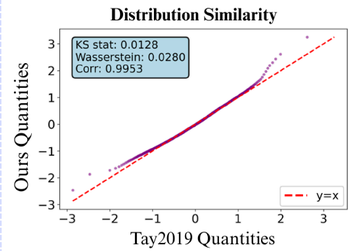
Our Team

Result on DNS-Challenge

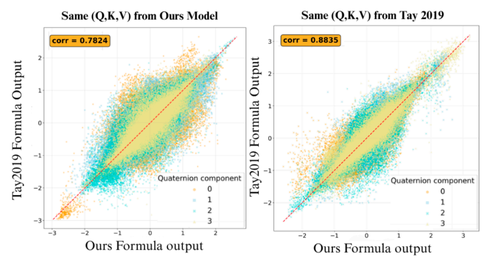
Method	Attention Type	Params (M)	OVRL	P.808	GPU RTF	CPU RTF
Noisy	–	–	2.11	2.90	–	–
Real-valued Conformer	Standard	3.19	2.77	3.41	0.0071	–
Quaternion Models (Ours vs. Tay et al. (2019) Attention)						
QTransformer (Tay et al., 2019)	Hamilton	0.62	2.61	3.30	0.0192	0.594
QTransformer (Ours)	Shared Score	0.62	2.67	**3.36**	**0.0107**	**0.249**
QConformer (Tay et al., 2019)	Hamilton	0.80	2.67	3.33	0.0202	0.610
QConformer (Ours)	Shared Score	0.80	**2.69**	3.32	**0.0157**	**0.259**

- GPU RTF −44.3% (1.79× faster), CPU RTF −58.1% (2.39× faster).

Why Does a Shared Score Suffice?



The shared-score (ours) and Tay et al. (2019) attentions produce near-identical attention output distributions (quantile correlation: 0.995).



When the same (Q, K, V) extracted from independently trained models is fed into both formulas, the outputs are highly correlated (corr = 0.78–0.88).

Cross-Domain Validation

Domain	Metric	Quality		Efficiency Gain	KS Stat.
		Tay et al.	Ours		
Speech (DNS-3, QTransformer)	OVRL	2.61	2.67	GPU RTF 44.3% ↓ (1.79×)	0.0128
Vision (CIFAR-100)	Acc. (%)	71.09	70.94	Training time 1.40×	0.0151
NLP (SST-2)	Acc. (%)	81.12	81.04	Latency 26.8% ↓ (1.37×)	0.0407

The same trend holds in vision and NLP — quality preserved, output distributions nearly identical (corr > 0.995), and consistent efficiency gains.

One shared score is enough — four were redundant.

[1] Y. Tay, A. Zhang, A. T. Luu, J. Rao, S. Zhang, S. Wang, J. Fu, and S. C. Hui. Lightweight and Efficient Neural Natural Language Processing with Quaternion Networks. ACL 2019, pp. 1494–1503.