NEURAL INFORMATION PROCESSING SYSTEMS

MONASH University

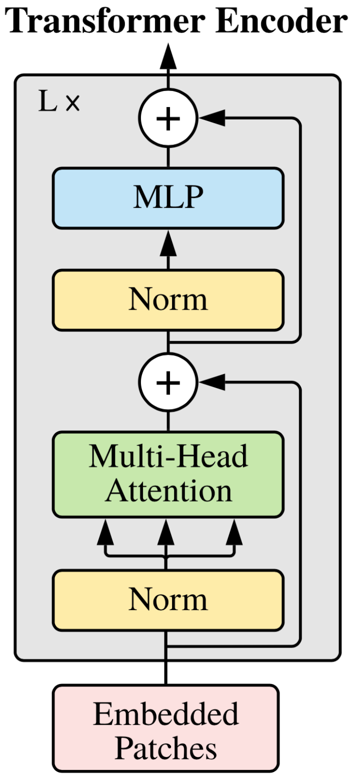
EcoFormer: Energy-Saving Attention with Linear Complexity

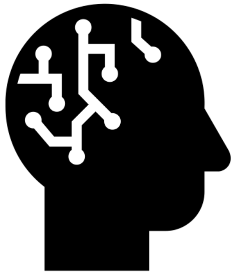
Jing Liu*, Zizheng Pan*, Haoyu He, Jianfei Cai, Bohan Zhuang†

Department of Data Science & AI, Monash University, Australia

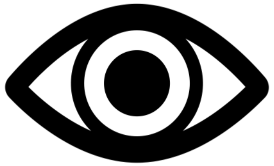
bohan.zhuang@monash.edu

Background: Transformers





Natural Language Processing

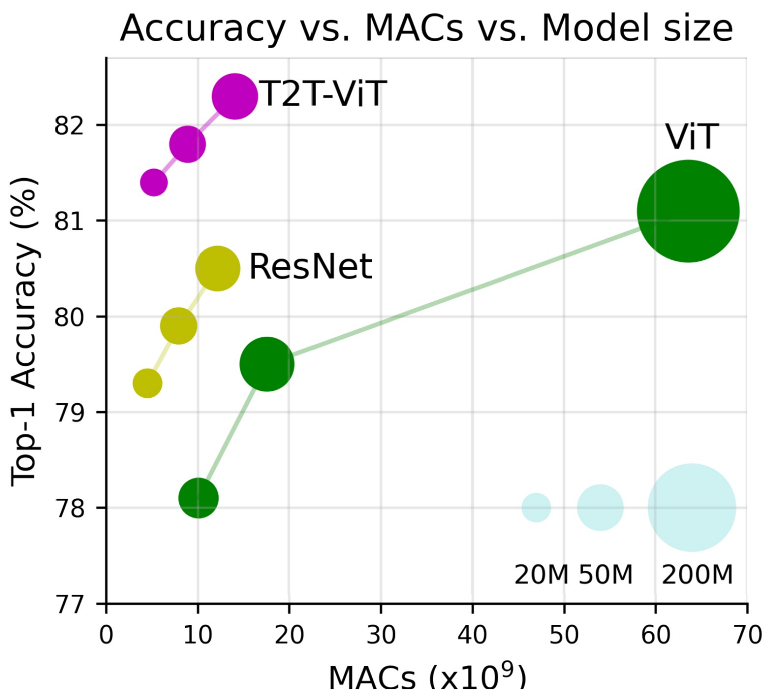


Computer Vision

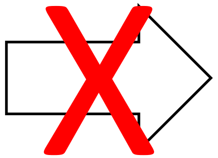
Transformers treat input as a sequence of patches and processes with a Transformer encoder.

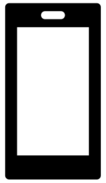
Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

Background: Transformers



Deploy





Mobile Phone

unmanned
aerial vehicles

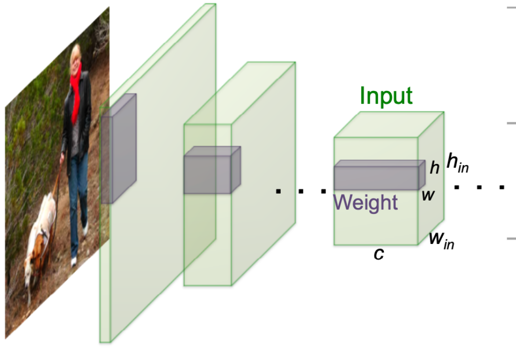
The efficiency bottlenecks greatly hamper the massive deployment to resource-constrained edge devices.

Yuan et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet. CVPR 2021.

Background: Binary Quantization

Table. Energy cost for different operations (on 45nm CMOS technology.

Operation	16-bit FP Add	16-bit FP Mult	32-bit FP Add	32-bit FP Mult
Energy (pJ)	0.4	1.1	0.9	3.7
Area (μm^2)	1,360	1,640	4,184	7,700



	Network Variations	Operations used in Convolution	Memory Saving (Inference)	Computation Saving (Inference)	Accuracy on ImageNet (AlexNet)
Standard Convolution	Real-Value Inputs: 0.11 -0.21 ... -0.34 / -0.25 0.61 ... 0.52; Real-Value Weights: 0.12 -1.2 ... 0.41 / -0.2 0.5 ... 0.68	+ , − , ×	1x	1x	%56.7
Binary Weight	Real-Value Inputs: 0.11 -0.21 ... -0.34 / -0.25 0.61 ... 0.52; Binary Weights: 1 -1 ... 1 / -1 1 ... 1	+ , −	~32x	~2x	%56.8
BinaryWeight Binary Input (**XNOR-Net**)	Binary Inputs: 1 -1 ... -1 / -1 1 ... 1; Binary Weights: 1 -1 ... 1 / -1 1 ... 1	XNOR , bitcount	~32x	~58x	%44.2

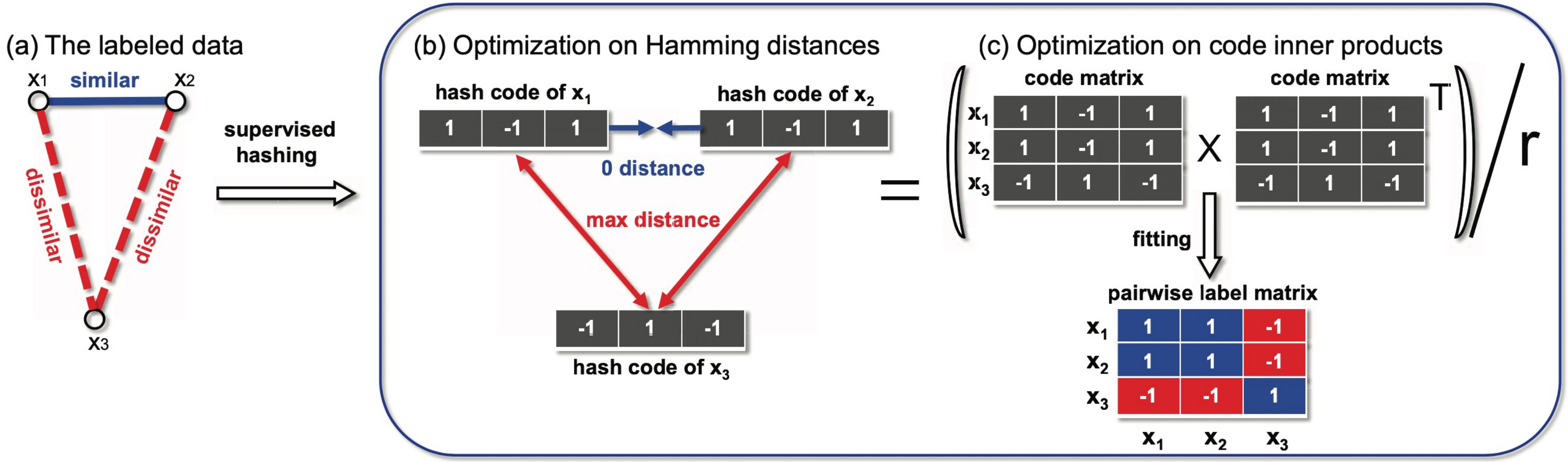
Binarize weights enable to replace the multiply-accumulate operations with energy-efficient accumulations.

(a) Rastegari et al. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. ECCV 2016.
(b) Han et al. EIE: Efficient Inference Engine on Compressed Deep Neural Network. ISCA 2016.
(c) Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). ISSCC 2014.
(d) Lian et al. High-performance FPGA-based CNN accelerator with block-floating-point arithmetic. VLSI 2019.

Background: Limitation of Binary Quantization

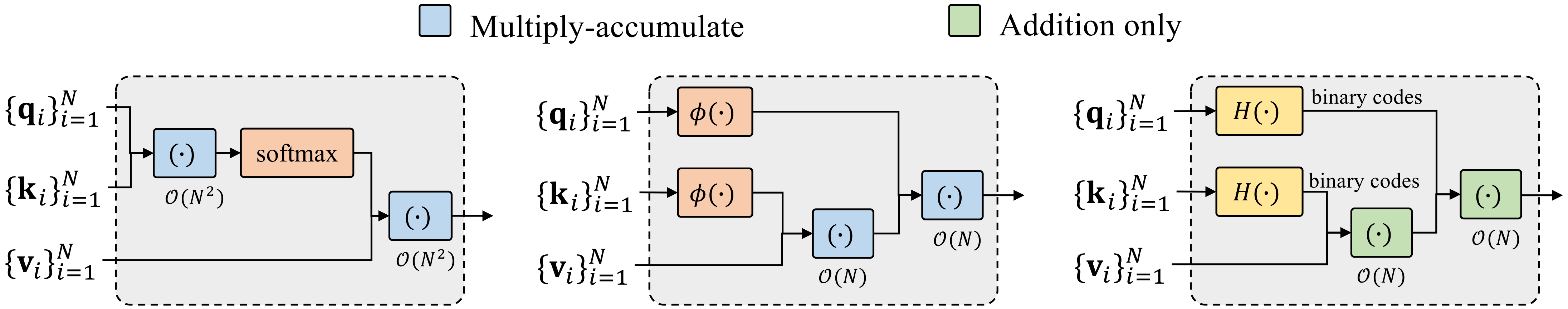
Limitation: Binary quantization can not well preserve the similarity relations among tokens in attention.

Solution: Learn kernelized hashing to map the queries and keys into low-dimensional binary codes while preserving the similarity relations in Hamming space.



Liu et al. Supervised hashing with kernels. CVPR 2012.

EcoFormer



(a) Attention (b) Kernel-based Linear Attention (c) EcoFormer

- Utilize kernel-based linear attention reduces the time complexity from $\mathcal{O}\left(N^2\right)$ to $\mathcal{O}(N)$.
- Use self-supervised kernelized hashing to map the queries and keys to compact binary codes.
- Replace most of the energy-hungry floating-point multiplications with energy-efficient additions.

EcoFormer

Hash function: $h(\mathbf{Q}) = \operatorname{sign}\left(\sum_{j=1}^{m}\left(\kappa\left(\mathbf{q}_{(j)}, \mathbf{Q}\right) - \mu_j\right) a_j\right) = \operatorname{sign}\left(\mathbf{g}(\mathbf{Q})\mathbf{a}\right)$

$$H(\mathbf{Q}) = [h_1(\mathbf{Q}), \cdots, h_b(\mathbf{Q})] = \begin{bmatrix} h_1(\mathbf{q}_1), \cdots, h_b(\mathbf{q}_1) \\ \cdots\cdots \\ h_1(\mathbf{q}_N), \cdots, h_b(\mathbf{q}_N) \end{bmatrix} = \operatorname{sign}\left(\mathbf{g}(\mathbf{Q})\mathbf{A}\right)$$

Objective for hash function learning:

$$\min_{\mathbf{A}} \left\|H(\mathbf{Q})H(\mathbf{Q})^\top - b\mathbf{Y}\right\|_F^2 = \min_{\mathbf{A}} \left\|\sum_{r=1}^{b} h_r(\mathbf{Q})h_r(\mathbf{Q})^\top - b\mathbf{Y}\right\|_F^2$$

where $\mathbf{Y}_{ij} = \begin{cases} 1, & (\mathbf{q}_i, \mathbf{q}_j) \in \mathcal{S} \quad \text{Similar pairs of tokens} \\ -1, & (\mathbf{q}_i, \mathbf{q}_j) \in \mathcal{U} \quad \text{Dissimilar pairs of tokens} \\ 0, & \text{otherwise.} \end{cases}$

Results on ImageNet-1k

Table. Main results on ImageNet-1K.

Model	Method	#Mul. (B)	#Add. (B)	Energy (B pJ)	Throughput (images/s)	Top-1 Acc. (%)
PVTv2-B0 [62]	MSA	2.02	1.99	9.25	850	70.77
	Ours	**0.54**	**0.56**	**2.49**	**1379**	**70.44**
PVTv2-B1	MSA	5.02	5.00	23.07	621	78.83
	Ours	**2.03**	**2.09**	**9.39**	**874**	**78.38**
PVTv2-B2	MSA	8.64	8.60	39.71	404	81.82
	Ours	**3.85**	**3.97**	**17.82**	**483**	**81.28**
PVTv2-B3	MSA	11.86	11.82	54.56	310	82.26
	Ours	**6.54**	**6.72**	**30.25**	**325**	**81.96**
PVTv2-B4	MSA	15.97	15.93	73.43	247	82.42
	Ours	**9.57**	**9.82**	**44.25**	**249**	**81.90**
Twins-SVT-S [10]	MSA	5.96	5.91	27.36	426	81.66
	Ours	**2.72**	**2.81**	**12.59**	**576**	**80.22**

- Our EcoFormer achieves **lower computational complexity**, **less energy consumption** and **higher throughput** with comparable performance.

Results on LRA

Table. Comparisons of different methods on Long Range Arena (LRA).

Method	#Mul. (B)	#Add. (B)	Energy (B pJ)	Text (4K)	Retrieval (4K)	Average
Transformer	4.63	4.57	21.25	64.87	79.62	72.25
Performer [9]	0.83	0.84	3.83	64.82	79.08	71.95
Linformer [60]	0.81	0.81	3.74	57.03	78.11	67.57
Reformer [29]	0.54	0.54	2.49	65.19	79.46	72.33
Combiner* [50]	0.51	0.51	2.34	64.36	56.10	60.23
EcoFormer	**0.25**	**0.29**	**1.17**	64.79	78.67	71.73

- Our EcoFormer saves around **94.6%** multiplications, **93.7%** additions and **94.5%** on-chip energy consumption compared with standard multi-head self-attention.

(a) Choromanski et al. Rethinking attention with performers. ICLR 2021.
(b) Wang et al. Linformer: Self-attention with linear complexity. ArXiv 2020.
(c) Kitaev et al. Reformer: The efficient transformer. ICLR 2020.
(d) Ren et al. Combiner: Full attention transformer with sparse computation cost.. NeurIPS 2021.

Quantization vs. hashing

Table. Performance comparisons with different binarization methods on CIFAR-100.

Model	Method	#Mul. (B)	#Add. (B)	Energy (B pJ)	Top-1 Acc. (%)
PVTv2-B0	FP-EcoFormer	0.94	0.94	4.33	70.78
	Bi-EcoFormer	0.63	0.83	3.09	70.06
	EcoFormer	**0.54**	**0.56**	**2.49**	**71.23**
Twins-SVT-S	FP-EcoFormer	5.96	5.91	27.36	80.04
	Bi-EcoFormer	3.01	3.59	14.38	80.04
	EcoFormer	**2.72**	**2.81**	**12.58**	**80.31**

- ❑ Our EcoFormer with lower energy cost **consistently outperforms** Bi-EcoFormer on different frameworks.
- ❑ Our proposed self-supervised hash functions **preserve the pairwise similarity** of attention.

Hubara et al. Quantized neural networks: Training neural networks with low precision weights and activations. JMLR 2017.

Thanks for Watching

Please refer to our paper and code for more details

Paper

Code