Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures

Yuchen Duan*, Weiyun Wang*, Zhe Chen*, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, Wenhai Wang ✉
*Equal contribution. ✉Corresponding authors: wangwenhai@pjlab.org.cn

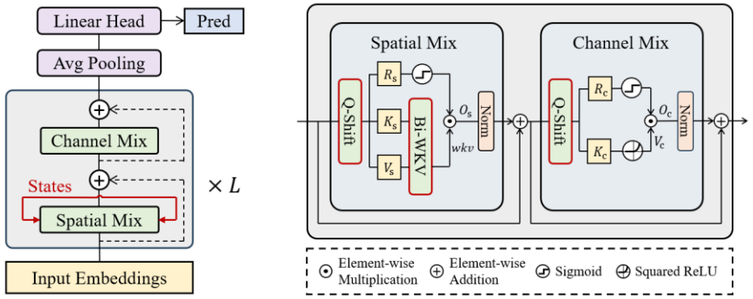
香港中文大學 The Chinese University of Hong Kong
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory
FUDAN UNIVERSITY 南京大學 NANJING UNIVERSITY Tsinghua University

Challenges in RWKV

- Missing Local bias in Vision;
- Monodirectional attention;
- Scalable ability in Vision.

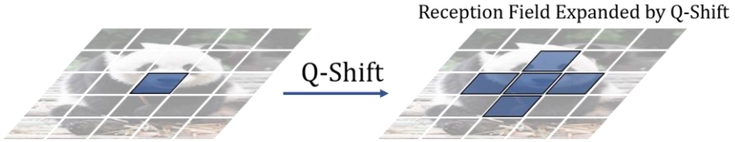
- **Locality:** Q-Shift
- **Bi-directional attention**: Bi-WKV
- **Scale up stability**: Bi-WKV & Extra Norm

Overall Architecture



(a) Vision-RWKV Architecture (b) Vision-RWKV Encoder Layer

Token Shift adapts Vision: Q-Shift



$$\text{Q-Shift}_{(*)}(X) = X + (1 - \mu_{(*)})X^{\dagger},\ (*) \in \{R, K, V\}$$
$$\text{where } X^{\dagger}[h, w] = \text{Concat}(X[h-1, w, 0:C/4],\ X[h+1, w, C/4:C/2],$$
$$X[h, w-1, C/2:3C/4],\ X[h, w+1, 3C/4:C]).$$

- **Motivation:** Enlarge Reception Field (by one patch width each shift)
  RWKV: Words → Phrases
  VRWKV: Image Tokens → Larger Image Tokens
- **Complexity:** Only a few FLOPs: $O(Td)$.

Code & Models are available at:
https://github.com/OpenGVLab/Vision-RWKV

Linear Complexity Global Attn: Bi-WKV

$$wkv = \text{Bi-WKV}(K, V, w, u),\ \{wkv, K, V\} \in \mathbb{R}^{T\times d}, \{w, u\} \in \mathbb{R}^{d}$$
w : Learnable decay
u : Learnable bonus

- **Summation form:**

$$wkv_t = \frac{\sum_{i=0, i\neq t}^{T-1} e^{-(|t-i|-1)/T\cdot w + k_i} v_i + e^{u+k_t} v_t}{\sum_{i=0, i\neq t}^{T-1} e^{-(|t-i|-1)/T\cdot w + k_i} + e^{u+k_t}}$$

wkv_t contains information of all tokens (**Bi−directional**);
Relative bias (**Bounded and Adapted to Vision**).

- **RNN form:** (Bi−directional RNN Cell)

$$wkv_t = \frac{a_{t-1} + b_{t-1} + e^{k_t+u} v_t}{c_{t-1} + d_{t-1} + e^{k_t+u}}, \quad \begin{cases} a_t = e^{-w/T} a_{t-1} + e^{k_t} v_t \\ b_t = e^{w/T}\left(b_{t-1} - e^{k_{t+1}} v_{t+1}\right) \\ c_t = e^{-w/T} c_{t-1} + e^{k_t} \\ d_t = e^{w/T}\left(d_{t-1} - e^{k_{t+1}}\right) \end{cases}$$

FLOPs each step: $O(Td)$.

Scale up Stability

Risks

1. All metrics/vectors in Exponent (May cause vanish/exceed);
2. Squared ReLU.

Solutions

1. Bounded Exponential: $\exp(*/T)$;
2. Extra Layer Normalization: $\text{norm}(wkv), \text{norm}(\text{ReLU}^2(K_c))$;
3. Layer Scale.

Model Details

Model	Emb Dim	Hidden Dim	Depth	Extra Norm	#Param
VRWKV-T	192	768	12	×	6.2M
VRWKV-S	384	1536	12	×	23.8M
VRWKV-B	768	3072	12	×	93.7M
VRWKV-L	1024	4096	24	✓	334.9M

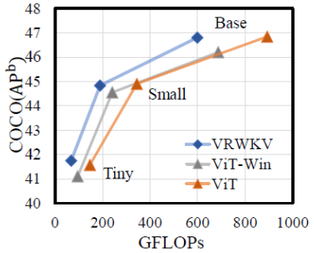
Design follows the rules of ViT-T/S/B/L

Experiments

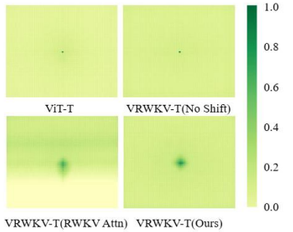
Classification (ImageNet)

	VRWKV−T *vs.*	DeiT−T
Acc	75.1 *vs.*	72.2 (+2.9)
#Param	5.7M *vs.*	6.2M
FLOPs	1.2G *vs.*	1.3G
	VRWKV−L *vs.* ViT−L	
Acc	86.0 *vs.*	85.2 (+2.9)
#Param	334.9M *vs.*	309.5M
FLOPs	189.5G *vs.*	191.1G

Detection (Coco)



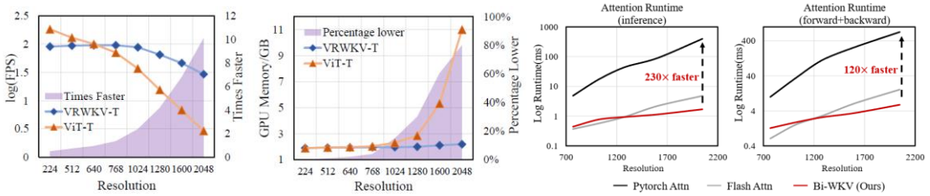
Ablation Study: Shift and Attn Type



Method	Token Shift	Bidirectional Attention	Top-1 Acc
RWKV	original	×	71.1 (-4.0)
Variant 1	none	✓	71.5 (-3.6)
Variant 2	original	✓	74.4 (-0.7)
Variant 3	Q-Shift	×	72.8 (-2.3)
VRWKV-T	Q-Shift	✓	75.1

- Bidirectional Attention vs. Original RWKV attention:
  Acc + 2.3%, More comprehensive ERF;
- Q−Shift *vs.* No Shift *vs.* Original Shift
  Acc 75.1 *vs.* 71.5 *vs.* 74.7
  Q−Shift brings inductive bias for vision tasks.

Efficiency Analysis



- **Linear Complexity**: Higher speed at high resolution.
- **RNN form computation**: Save GPU memory.