RWKV，引领大模型架构变更的新型RNN

林玥煜 元始智能

演讲嘉宾

林玥煜

元始智能算法工程VP

原始智能算法工程VP，曾任大数医达科技有限公司算法总监，阿里巴巴数据事业部系统架构师，多年来深耕大数据、人工智能在工业界应用和开发管理。对大语言模型在严肃医疗场景的应用、开发拥有丰富的实战经验。

目录
CONTENTS

RWKV

新一代模型架构/超越Transformer

计算效率高

推理速度、内存恒定

无限上下文

适合长文本处理、多轮对话等

对芯片友好

只做矩阵乘矢量，无 KV Cache

全球开源开放

Apache 2.0 协议

项目历史

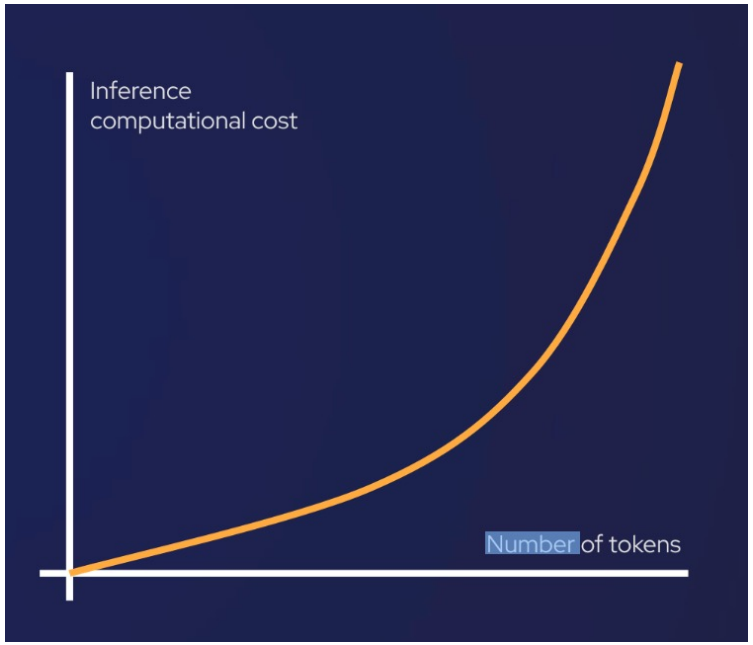

2020年初
一个人的
开源项目
2022.11
RWKV-4 7B
模型开源
2023.03
RWKV-4 14B
模型开源
2023.05
RWKV-4
论文公开
2023.06
2023.07
成立
元始智能
2023.08
全球开发者
超过2万
2023.10
第一个
商业客户
2023.10
RWKV-6
代码开源
2023.10
高通
全球合作
2023.12
种子轮
奇绩创坛
2024.01
RWKV-5 7B
模型开源
2024.2
第一个
toC 端侧应用
2024.04
RWKV-5
和RWKV-6
论文公开
2024.05
RWKV-6 7B
模型开源

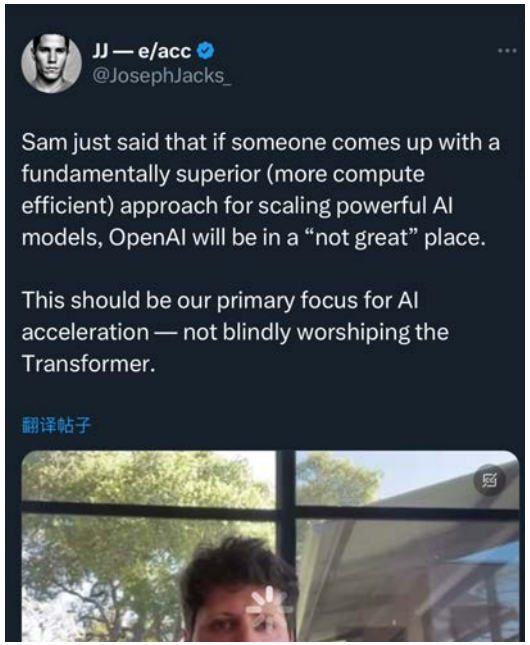
RWKV要解决的问题

⚠️ Transformer 是死胡同

算力需求巨大，Scaling-law 失效

OpenAI





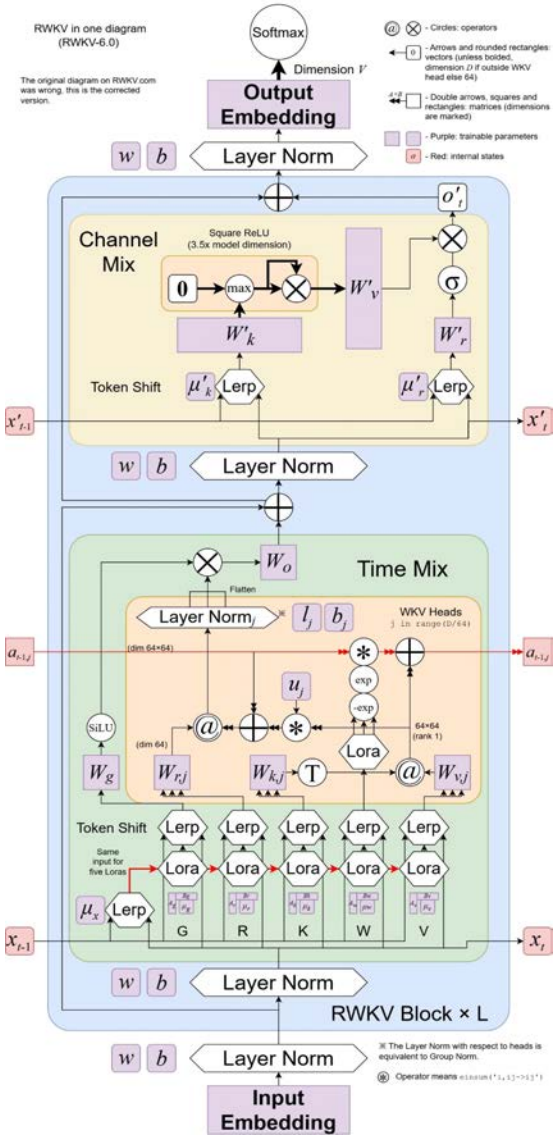
TESLA

RWKV 开始于 2020 年初，正在研发 RWKV-7

RWKV 正引领大模型的架构迁移

架构名称	作者和论文地址	架构版本	阶段	算法复杂度	最大模型参数	最大训练TOKEN
RWKV	Bo PENG https://arxiv.org/abs/2305.13048	RWKV-6	商用	O(N)	14 B	2.5 T (SlimPajama+pile+全球语言+代码)
Mamba	CMU，Princeton https://arxiv.org/abs/2312.00752	接近 RWKV-6	发展	O(N)	6.7 B	0.627 T (SlimPajama)
Gated Linear Attention	MIT https://arxiv.org/abs/2312.06635	接近 RWKV-6	研究	O(N)	1.3 B	0.1 T
Striped Hyena	Together, Stanford https://arxiv.org/abs/2302.10866	接近 RWKV-4.5 与 Llama2 的混合	发展	O(N logN) 与 O(N^2) 之间	7 B	1 T+
xLSTM	LSTM 作者 https://arxiv.org/abs/2405.04517	接近 RWKV-6	研究	O(N)	1.3 B	0.3 T
RetNet	微软亚洲研究院，清华大学 https://arxiv.org/abs/2307.08621	接近 RWKV-5	研究	O(N)	6.7 B	0.1 T
TransnormerLLM	上海人工智能实验室，OpenNLPLab https://arxiv.org/abs/2307.14995	接近 RWKV-5	发展	O(N)	6.8B	1.4T

RWKV
兼具
Transformer
高效训练和
RNN高效推理
的新架构



我们是怎么做的?

RNN和Transformer各自的局限性

- RNN 在训练长序列时容易出现梯度消失问题。
- RNN 在训练过程中无法在时间维度上进行并行化，限制了其可扩展性。
- Transformer 具有二次复杂度, 长序列任务中计算成本高和占用内存多。

时间和空间复杂度比较

Architecture	Inference			Training	
	Time	Memory	Parallel	Time	Memory
LSTM/LMU	$O(1)$	$O(1)$	✗	$O(N)$	$O(N)$
Transformer	$O(N)$	$O(N)^a$	✓	$O(N^2)$	$O(N)^b$
Linear Transformer	$O(1)$	$O(1)$	✓	$O(N)$	$O(N)$
H3/S4	$O(1)$	$O(1)$	✓	$O(N \log N)$	$O(N)$
Hyena	$O(N)$	$O(N)$	✓	$O(N \log N)$	$O(N)$
RWKV/Mamba/RetNet	$O(1)$	$O(1)$	✓	$O(N)$	$O(N)$

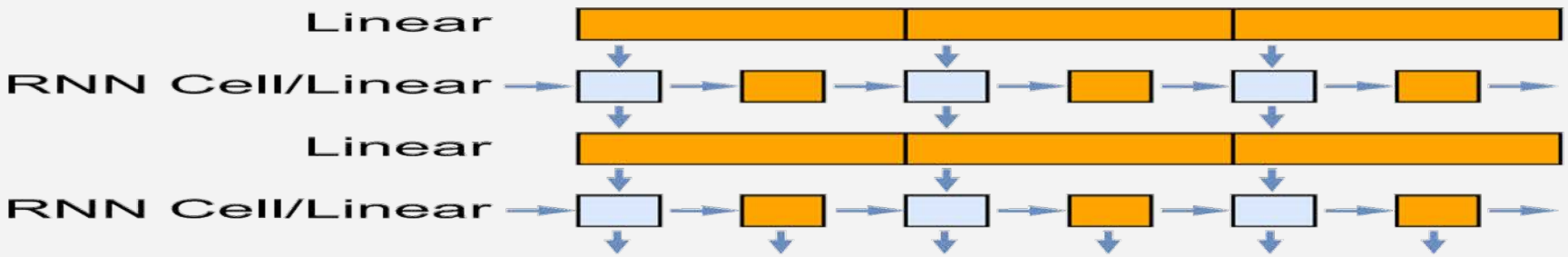
我们是怎么做的?

RNN最简单有效的基本形式

- $ht = \alpha_t \odot h_{t-1} + (1 - \alpha_t) \odot x_t$
- RNN 一步一步执行，每次仅处理一个字或一个词
- 内存占用小，计算量小
- 对前一步结果的依赖，使得 RNN 无法并行化训练，极大限制了 RNN 的可扩展性

相比较，Transformer 一次处理一整句话，或一整段话，可以并行训练

RNN 结构示意图



我们是怎么做的?

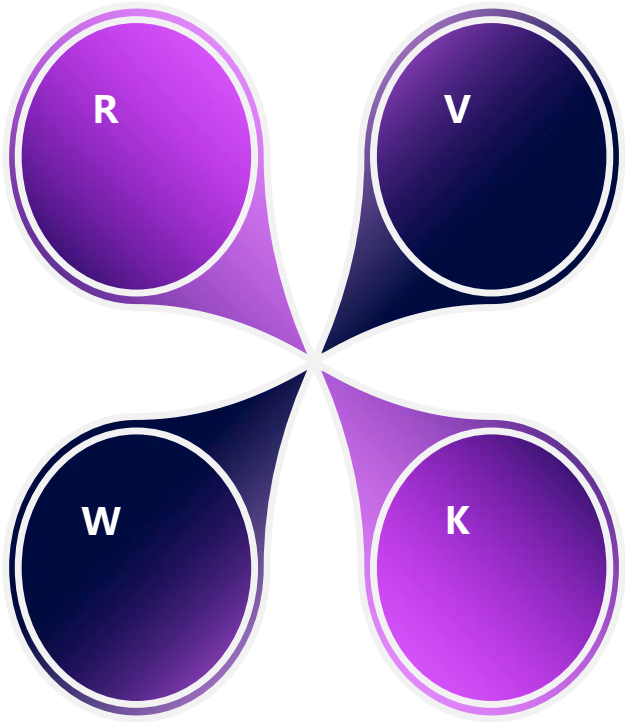
RWKV的由来

Receptance

作为过去信息的接受程度的接受向量

值（Value）

类似于传统注意力中 V 的向量

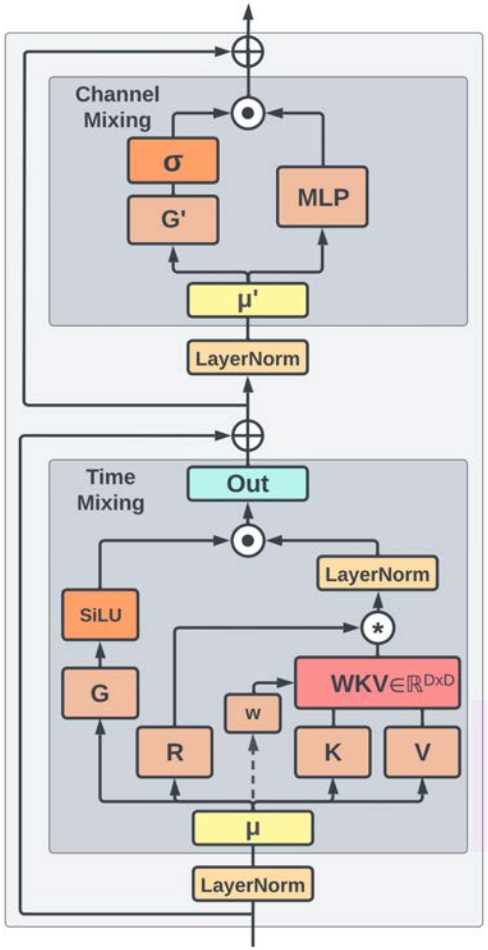


Weight

位置权重衰减向量，可训练的模型参数

键（Key）

类似于传统注意力中 K 的向量



虚线代表RWKV-6中有
RWKV-5中没有

RWKV与QKV相对，贯穿整个RWKV系列

时间混合模块的演进

RWKV-5 time-mixing时间混合模块

$$\square_t = \text{lerp}_\square(x_t, x_{t-1})W_\square, \quad \square \in \{r, k, v, g\}$$
$$w = \exp(-\exp(\omega))$$
$$wkv_t = \text{diag}(u) \cdot k_t^{\mathrm{T}} \cdot v_t + \sum_{i=1}^{t-1} \text{diag}(w)^{t-1-i} \cdot k_i^{\mathrm{T}} \cdot v_i \in \mathbb{R}^{(D/h)\times(D/h)}$$
$$o_t = \text{concat}\left(\text{SiLU}(g_t) \odot \text{LayerNorm}(r_t \cdot wkv_t)\right) W_o \in \mathbb{R}^D$$

- RWKV-5中消除了归一化项（分母）
- RWKV-5引入了矩阵值状态, k,v的维度从D-> (D/H, D/H)

RWKV-6 time-mixing时间混合模块

$$\square_t = \text{ddlerp}_\square(x_t, x_{t-1})W_\square, \quad \square \in \{r, k, v, g\}$$
$$d_t = \text{lora}_d(\text{ddlerp}_d(x_t, x_{t-1}))$$
$$w_t = \exp(-\exp(d_t))$$
$$wkv_t = \text{diag}(u) \cdot k_t^{\mathrm{T}} \cdot v_t + \sum_{i=1}^{t-1} \text{diag}\left(\bigodot_{j=1}^{i-1} w_j\right) \cdot k_i^{\mathrm{T}} \cdot v_i \in \mathbb{R}^{(D/h)\times(D/h)}$$
$$o_t = \text{concat}\left(\text{SiLU}(g_t) \odot \text{LayerNorm}(r_t \cdot wkv_t)\right) W_o \in \mathbb{R}^D$$

- RWKV-6引入了channel-wise的衰减率w_t

time-mixing时间混合模块的演进

t	RWKV-4 $u, w, k_t, v_t \in \mathbb{R}^D$, head size 1
0	$\sigma(r_0) \odot \left(\frac{u \odot k_0 \odot v_0}{u \odot k_0}\right)$
1	$\sigma(r_1) \odot \left(\frac{u \odot k_1 \odot v_1 + k_0 \odot v_0}{u \odot k_1 + k_0}\right)$
2	$\sigma(r_2) \odot \left(\frac{u \odot k_2 \odot v_2 + k_1 \odot v_1 + w \odot k_0 \odot v_0}{u \odot k_2 + k_1 + w \odot k_0}\right)$
3	$\sigma(r_3) \odot \left(\frac{u \odot k_3 \odot v_3 + k_2 \odot v_2 + w \odot k_1 \odot v_1 + w^2 \odot k_0 \odot v_0}{u \odot k_3 + k_2 + w \odot k_1 + w^2 \odot k_0}\right)$

t	Eagle (RWKV-5) $\text{diag}(u)$, $\text{diag}(w)$, $k_t, v_t \in \mathbb{R}^{64\times64}$ for each head, head size 64
0	$r_0 \cdot \left(\text{diag}(u) \cdot k_0^{\mathrm{T}} \cdot v_0\right)$
1	$r_1 \cdot \left(\text{diag}(u) \cdot k_1^{\mathrm{T}} \cdot v_1 + k_0^{\mathrm{T}} \cdot v_0\right)$
2	$r_2 \cdot \left(\text{diag}(u) \cdot k_2^{\mathrm{T}} \cdot v_2 + k_1^{\mathrm{T}} \cdot v_1 + \text{diag}(w) \cdot k_0^{\mathrm{T}} \cdot v_0\right)$
3	$r_3 \cdot \left(\text{diag}(u) \cdot k_3^{\mathrm{T}} \cdot v_3 + k_2^{\mathrm{T}} \cdot v_2 + \text{diag}(w) \cdot k_1^{\mathrm{T}} \cdot v_1 + \text{diag}(w^2) \cdot k_0^{\mathrm{T}} \cdot v_0\right)$

t	Finch (RWKV-6) $\text{diag}(u)$, $\text{diag}(w_t)$, $k_t, v_t \in \mathbb{R}^{64\times64}$ for each head, head size 64
0	$r_0 \cdot \left(\text{diag}(u) \cdot k_0^{\mathrm{T}} \cdot v_0\right)$
1	$r_1 \cdot \left(\text{diag}(u) \cdot k_1^{\mathrm{T}} \cdot v_1 + k_0^{\mathrm{T}} \cdot v_0\right)$
2	$r_2 \cdot \left(\text{diag}(u) \cdot k_2^{\mathrm{T}} \cdot v_2 + k_1^{\mathrm{T}} \cdot v_1 + \text{diag}(w_1) \cdot k_0^{\mathrm{T}} \cdot v_0\right)$
3	$r_3 \cdot \left(\text{diag}(u) \cdot k_3^{\mathrm{T}} \cdot v_3 + k_2^{\mathrm{T}} \cdot v_2 + \text{diag}(w_2) \cdot k_1^{\mathrm{T}} \cdot v_1 + \text{diag}(w_2 \odot w_1) \cdot k_0^{\mathrm{T}} \cdot v_0\right)$

RWKV的RNN视角

RWKV5/6写成递归形式:

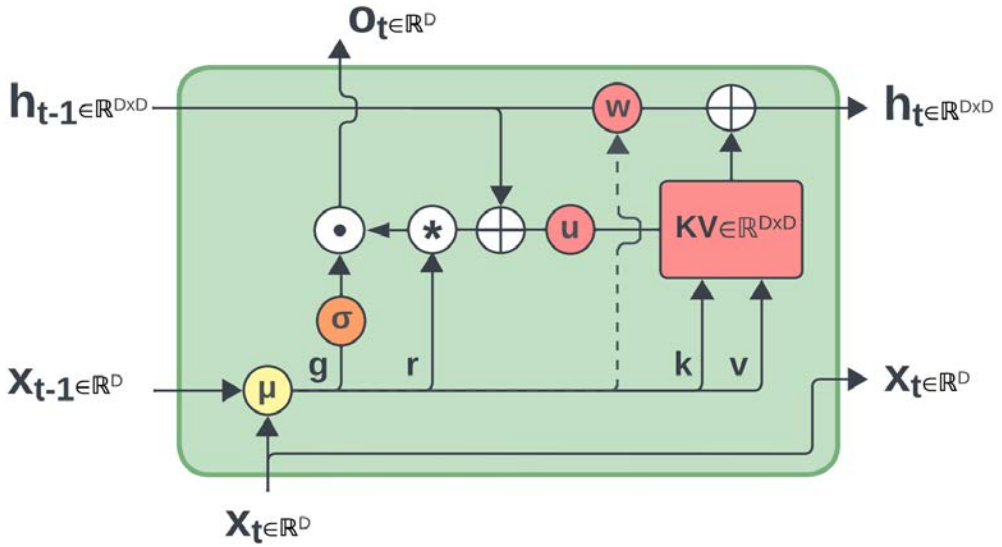
$$wkv_t = S_{t-1} + diag(u) \cdot k^T \cdot v$$

$$S_t = diag(w) \cdot S_{t-1} + k^T \cdot v$$

虽然递归形式一样，但是RWKV-5中的w是data-independent 的， 而 RWKV-6 中 的 w 是 data-dependent的w_t

代码形式:

```
# r, k, v parameter shape (B,H,1,D//H)
# w parameter of shape (1,H,1,D//H) for Eagle (RWKV-5),
#                      (B,H,1,D//H) for Finch (RWKV-6)
# u parameter of shape (1,H,1,D//H)
def rwkv_5_or_6_recurrent(r, k, v, w, u, wkv_state):
    kv = k.mT @ v # x.mT is equivalent to x.transpose(-2, -
1)
    out = r @ (wkv_state + u.mT * kv)
    wkv_state = w.mT * wkv_state + kv # (B,H,D//H,D//H)
    return out, wkv_state
```



虚线代表RWKV-6中有，RWKV-5中没有

WKV模块的改进:

RWKV-5: 通过学习得到的通道衰减率来替代RetNet中的静态衰减率。

RWKV-6: 通过动态生成依赖于数据的token-shift量和衰减率。

RWKV架构相对应传统RNN的改造

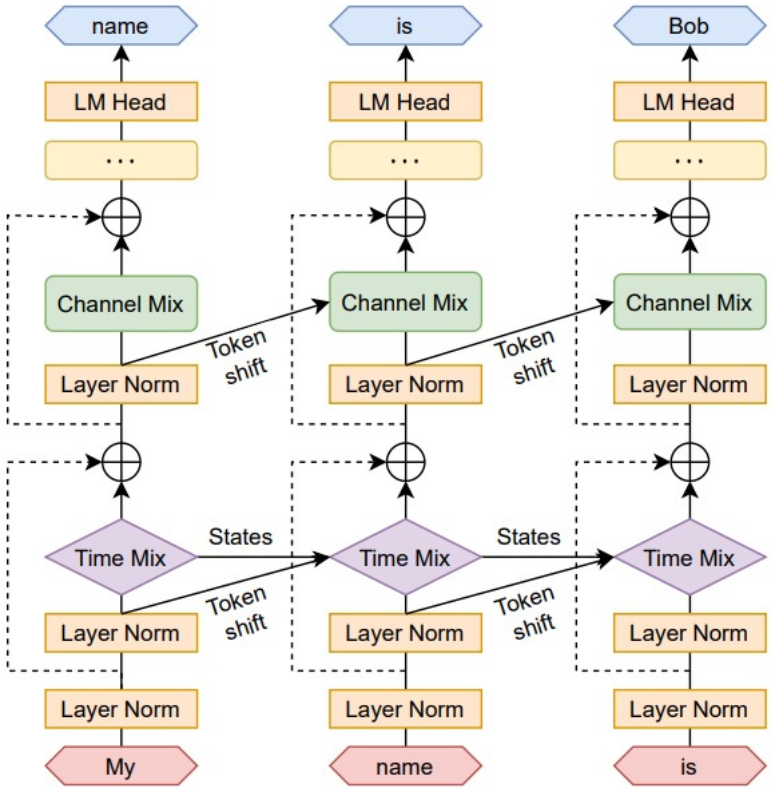


Figure 3: RWKV architecture for language modelling.

1.把每一个Block拆成若干个部分，在训练/预测的时候，不互相依赖的模块可以相互并行计算。

2.在需要状态传递的Time Mixer模块，通过CUDA/FLA扩展，在Channel Wise+Head Wise并行处理。由于Channel和Head的数目很多，通常都超过了一个GPU所拥有的Tensor core的数目，我们在Time Mixer模块也能充分利用GPU的并行计算能力。

```
template <typename F>
__global__ void kernel_forward(const int B, const int T, const int C, const int H,
                               const F *__restrict__ const _r, const F *__restrict__ const _k, const F
                               *__restrict__ const _v, const float *__restrict__ _w, const F
                               *__restrict__ _u,
                               F *__restrict__ const _y)
{
    const int b = blockIdx.x / H;
    const int h = blockIdx.x % H;
    const int i = threadIdx.x;
    _u += h*_N_;

    __shared__ float r[_N_], k[_N_], u[_N_], w[_N_];
    float state[_N_] = {0};

    __syncthreads();
    u[i] = float(_u[i]);
    __syncthreads();

    for (int t = b*T*C + h*_N_ + i; t < (b+1)*T*C + h*_N_ + i; t += C)
    {
```

高效并行训练不逊色于Transformer

RWKV的训练效率高，长上下文时更快

- Transformer Baseline: karpathy的nanoGPT，优化程度相当高的训练代码:
- 混合精度 + compile + flash_attn

- 当上下文长度较短时，RWKV训练速度略慢于GPT+ flash_attn
- 当上下文长度较长时(8k)，RWKV训练速度比GPT+flash_attn 更快。

model	block_size	batch_size	n_layer	n_head	n_embd	parameters	time(ms)
GPT	256	16	6	32	2048	302.15M	**60.13**
GPT	512	16	6	32	2048	302.15M	**108.12**
GPT	1024	16	6	32	2048	302.15M	**218.81**
GPT	2048	16	6	32	2048	302.15M	**494.59**
GPT	4096	16	6	32	2048	302.15M	**1011.19**
GPT	8192	16	6	32	2048	302.15M	**2590.22**
RWKV6	256	16	6	32	2048	333M	**104.27**
RWKV6	512	16	6	32	2048	333M	**175.13**
RWKV6	1024	16	6	32	2048	333M	**317.46**
RWKV6	2048	16	6	32	2048	333M	**595.23**
RWKV6	4096	16	6	32	2048	333M	**1186.23**
RWKV6	8192	16	6	32	2048	333M	**2384.32**

RWKV推理性能全球最佳

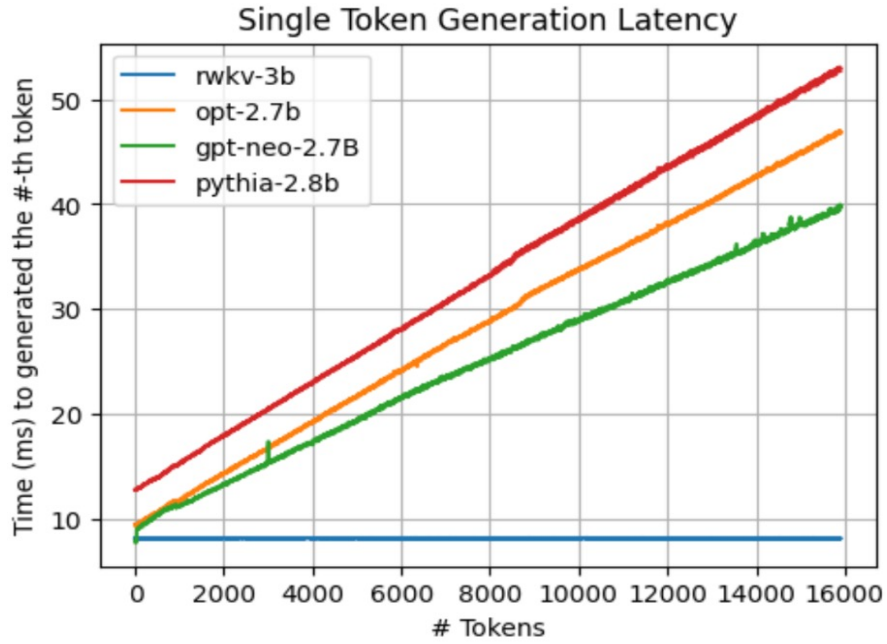
- RWKV 内存占用恒定，时间复杂度线性增加

VS

- Transformer 内存占用越来越大，时间复杂度指数增加

Model	Time	Space
Transformer	$O(T^2 d)$	$O(T^2 + Td)$
Reformer	$O(T \log T d)$	$O(T \log T + Td)$
Linear Transformers	$O(Td^2)$	$O(Td + d^2)$
Performer	$O(Td^2 \log d)$	$O(Td \log d + d^2 \log d)$
AFT-full	$O(T^2 d)$	$O(Td)$
MEGA	$O(cTd)$	$O(cd)$
RWKV (ours)	$O(\mathbf{Td})$	$O(\mathbf{d})$

推理复杂度比较



RWKV 架构与模型

⭐ 模型架构	语言模型	多模态模型	其他模型
RWKV-4 Dove 鸽	RWKV-6-World 1.6B	RWKV-Visual 图片识别	RWKV-TS 时序预测
RWKV-5 Eagle 鹰	RWKV-6-World 3B	RWKV-ASR 语音识别	RWKV-Math 数学模型
RWKV-6 Finch 雀	RWKV-6-World 7B	RWKV-Vision 视觉模型	RWKV-NLM 类脑模型
RWKV-7 Goose 雁	RWKV-6-World 14B	RWKV-Diffusion 扩散模型	
RWKV-8 Heron 鹭	RWKV-6-World 30B	RWKV-TTS 文生语音	
RWKV-9 Ibis 鹮	RWKV-6-World 70B	RWKV- SAM 分割一切模型	
		RWKV-Music 音乐模型	

已完成
开发中
计划中

RWKV 的落地场景

11000+

RWKV-LM
在 Github 的星

370+

Github 上
RWKV 项目数量

9000+

海外社群
开发者人数

10000+

国内社群
开发者和用户人数

RWKV ASR 语音识别模型

https://arxiv.org/pdf/2309.14758
https://github.com/alibaba-damo-academy/FunASR

AiDD AI+研发数字峰会 AI+ Development Digital summit

Alibaba Group 阿里巴巴集团

達摩院

Table 2. The latency, left context, and accuracy of different steaming models on AISHELL-1 (CER) and Librispeech (WER).

model	encoder	latency (ms)	left context (#frames)	AISHELL-1 test	LibriSpeech test clean	LibriSpeech test other
CTC + Att rescoring	chunk conformer [7]	640 + Δ	all history	5.05	3.80	10.38
Transducer	chunk conformer [20]	400	40	6.15	-	-
Transducer	streaming transformer [3]	0	10	-	4.2	11.3
Transducer	streaming transformer [3]	0	2	-	4.5	14.5
Transducer	causal conformer [6]	0	all history	-	4.6	9.9
Transducer	causal conformer + distill [6]	0	all history	-	3.7	9.2
Transducer	conv augmented LSTM [21]	0	1	-	5.11	13.82
Transducer	chunk conformer	640	16	**6.04**	**3.58**	**9.27**
Transducer	chunk conformer	320	8	6.32	4.19	10.84
Transducer	RWKV(S)	0	1	6.11	3.83	9.63
BAT	RWKV(S)	0	1	6.11	3.90	9.56

Visual-RWKV 视觉语言模型

光明实验室
GUANGMING LABORATORY

https://arxiv.org/pdf/2406.13362
https://github.com/howard-hou/VisualRWKV

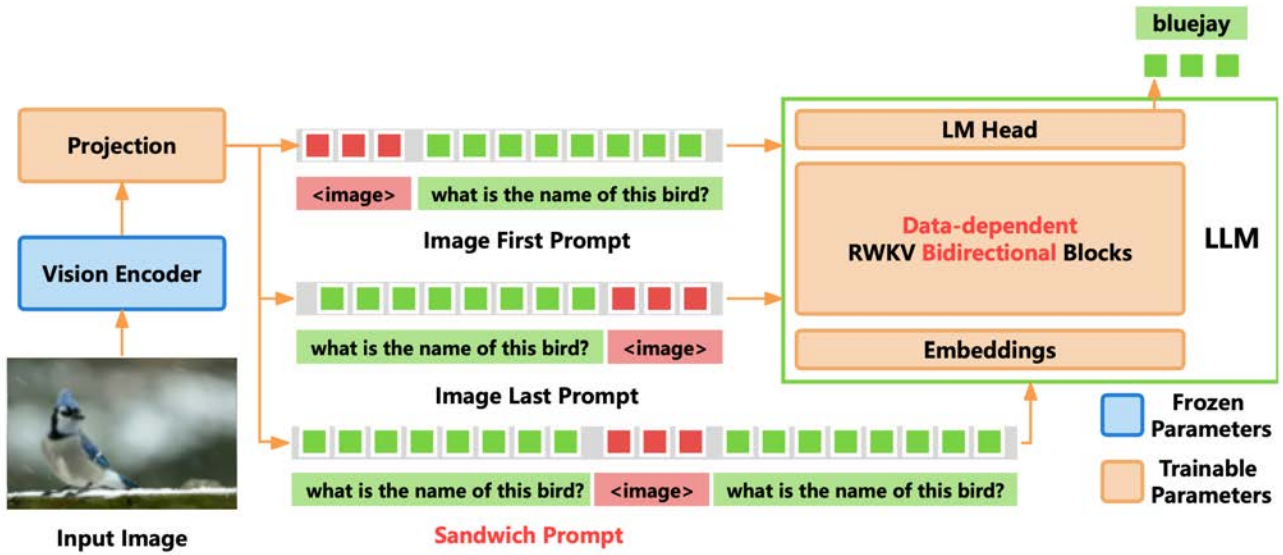


Figure 2: VisualRWKV architecture overview and three prompting method. **Image First Prompt:** place image tokens before instruction tokens; **Image Last Prompt:** place image tokens after instruction tokens; **Sandwich Prompt:** place image tokens in the middle of instruction tokens. Red words indicate the key contributions.

Method	LLM	Res.	PT/IT	VQA	GQA	SQA	TQA	POPE	MME	MMB	MMB-cn
BLIP-2 (Li et al., 2023c)	Vicuna-13B	224	129M/ -	41.0	41.0	61.0	42.5	85.3	1293.8	–	22.4
MiniGPT-4 (Zhu et al., 2024a)	Vicuna-7B	224	5M/5K	-	32.2	-	-	-	581.7	23.0	-
InstructBLIP (Dai et al., 2023)	Vicuna-7B	224	129M/1.2M	–	49.2	60.5	50.1	–	–	36	26.2
InstructBLIP (Dai et al., 2023)	Vicuna-13B	224	129M/1.2M	–	49.5	63.1	50.7	78.9	1212.8	–	25.6
Shikra (Chen et al., 2023b)	Vicuna-13B	224	600K/5.5M	77.4	–	–	–	–	–	58.8	–
Otter (Li et al., 2023a)	LLaMA-7B	224	-	-	-	-	-	-	1292.3	48.3	24.6
mPLUG-Owl (Ye et al., 2023)	LLaMA-7B	224	2.1M/102K	-	-	-	-	-	967.3	49.4	-
IDEFICS-9B (IDEFICS, 2023)	LLaMA-7B	224	353M/1M	50.9	38.4	–	25.9	–	–	48.2	–
IDEFICS-80B (IDEFICS, 2023)	LLaMA-65B	224	353M/1M	60.0	45.2	–	30.9	–	–	54.5	–
Qwen-VL (Bai et al., 2023)	Qwen-7B	448	1.4B/50M	**78.8**	59.3	67.1	**63.8**	–	–	38.2	–
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	448	1.4B/50M	78.2	57.5	68.2	61.5	–	1487.5	60.6	–
LLaVA-1.5 (Liu et al., 2023a)	Vicuna-7B	336	558K/665K	78.5	62.0	66.8	58.2	**85.9**	**1510.7**	64.3	30.5
LLaVA-Phi (Zhu et al., 2024b)	Phi2-2.7B	336	558K/665K	71.4	-	**68.4**	48.6	85.0	1335.1	59.8	28.9
MobileVLM-3B (Chu et al., 2023)	LLaMA-2.7B	336	558K/665K	-	59.0	61.2	47.5	84.9	1288.9	59.6	-
VL-Mamba (Qiao et al., 2024)	Mamba-2.8B	224	558K/665K	76.6	56.2	65.4	48.9	84.4	1369.6	57.0	32.6
VisualRWKV	RWKV6-1.6B	336	558K/665K	69.4	55.2	59.1	43.6	83.2	1204.9	55.8	53.2
VisualRWKV	RWKV6-3B	336	558K/665K	71.5	59.6	65.3	48.7	83.1	1369.2	59.5	56.3
VisualRWKV	RWKV6-7B	336	558K/665K	75.8	**64.3**	68.2	51.0	84.7	1387.8	**65.8**	**63.7**

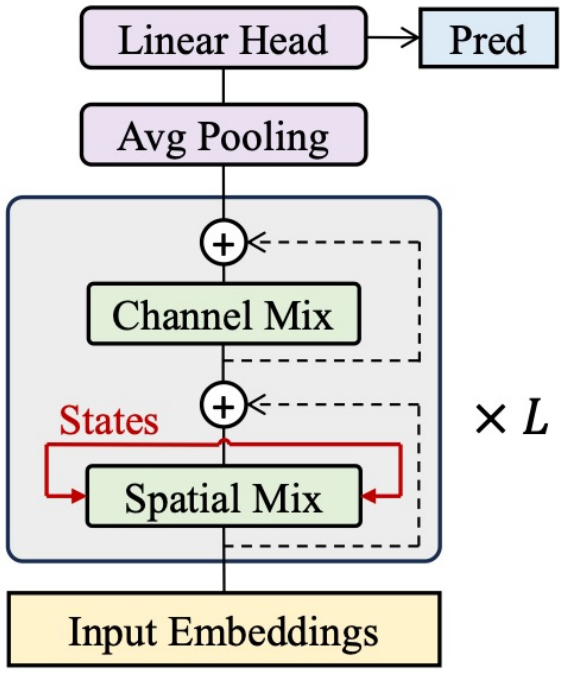
Table 2: **Comparison with SoTA methods on 8 benchmarks.** Due to space constraints, benchmark names are abbreviated. VQA (Goyal et al., 2017); GQA (Hudson and Manning, 2019); SQA: ScienceQA-IMG (Lu et al., 2022); TQA: TextVQA (Singh et al., 2019); POPE (Li et al., 2023d); MME (Fu et al., 2023); MMB: MMBench (Liu et al., 2023d); MMB-cn: MMBench-CN (Liu et al., 2023d). PT and IT denote the quantity of samples involved in the pre-training and instruction-tuning phases. "Res." stands for "Resolution.

Vision-RWKV 视觉感知模型

AI+研发数字峰会 AI+ Development Digital summit

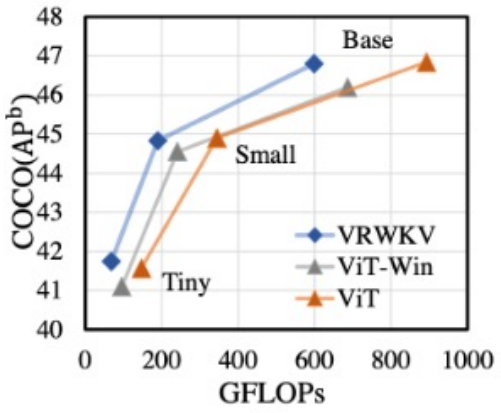
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

https://arxiv.org/pdf/2403.02308
https://github.com/OpenGVLab/Vision-RWKV

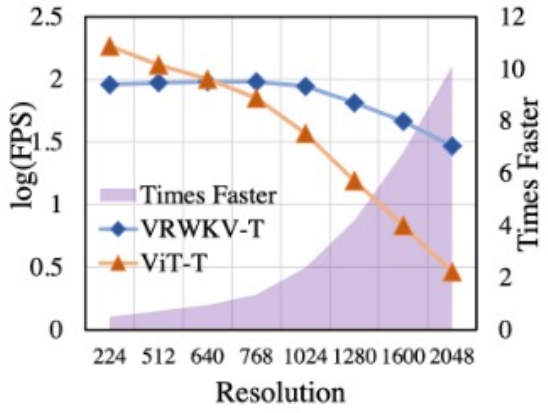


(a) Vision-RWKV Architecture

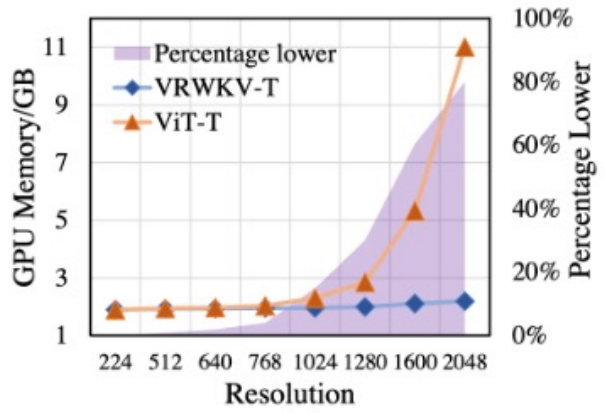
- **High-Resolution Efficiency**: Processed high-resolution images smoothly with a global receptive field.
- **Scalability**: Pre-trained with large-scale datasets and posses scale up stablity.
- **Superior Performance**: Achieved a better performance in classfication tasks than ViTs. Surpassed window-based ViTs and comparabled to global attention ViTs with lower flops and higher speed in dense prediction tasks.
- **Efficient Alternative**: Capability to be an alternative backbone to ViT in comprehensive vision tasks.



(a)



(b)



(c)

Diffusion-RWKV 扩散模型

https://arxiv.org/abs/2404.04478
https://github.com/feizc/Diffusion-RWKV

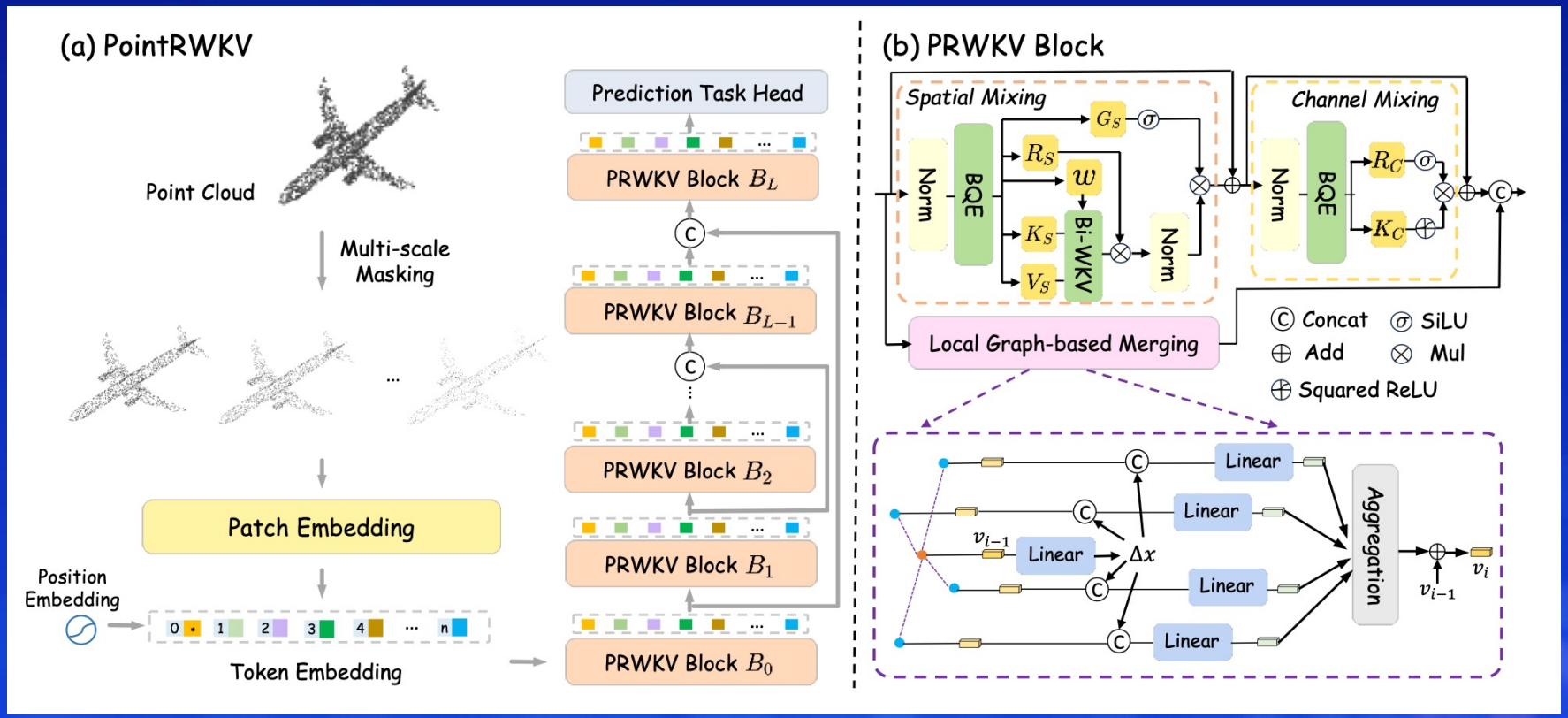
Figure 1. **Diffusion models with RWKV-like backbones achieve comparable image quality.** Selected samples generated by class-conditional Diffusion-RWKV trained on the ImageNet with resolutions of 256×256 and 512×512, respectivly.

PointRWKV 3D点云学习框架

AI+研发数字峰会 AI+ Development Digital summit

腾讯 Tencent

https://arxiv.org/pdf/2405.15214
https://hithqd.github.io/projects/PointRWKV



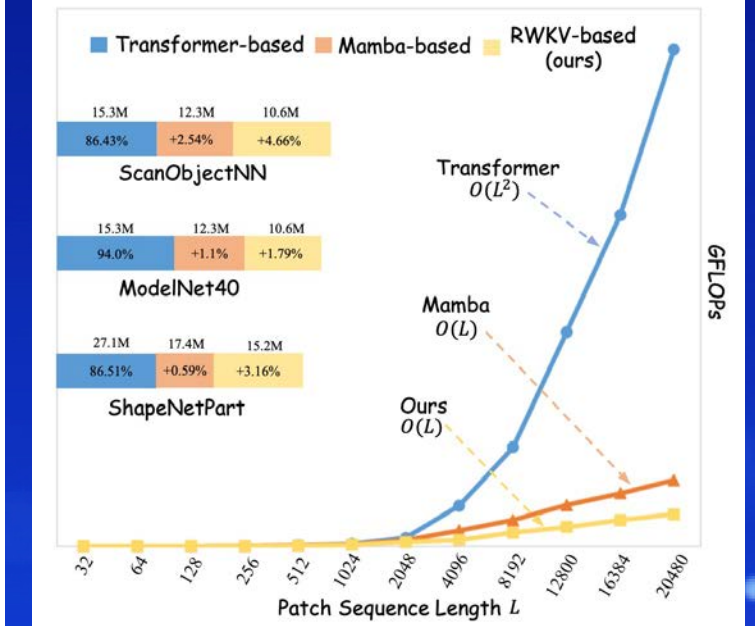


Figure 2: **Accuracy-speed tradeoff.** (Left) Overall accuracy acquired by different methods with relative parameters, (Right) FLOPs increase with sequence length.

RWKV-CLIP 视觉语言表示学习

https://arxiv.org/pdf/2406.06973
https://github.com/deepglint/RWKV-CLIP

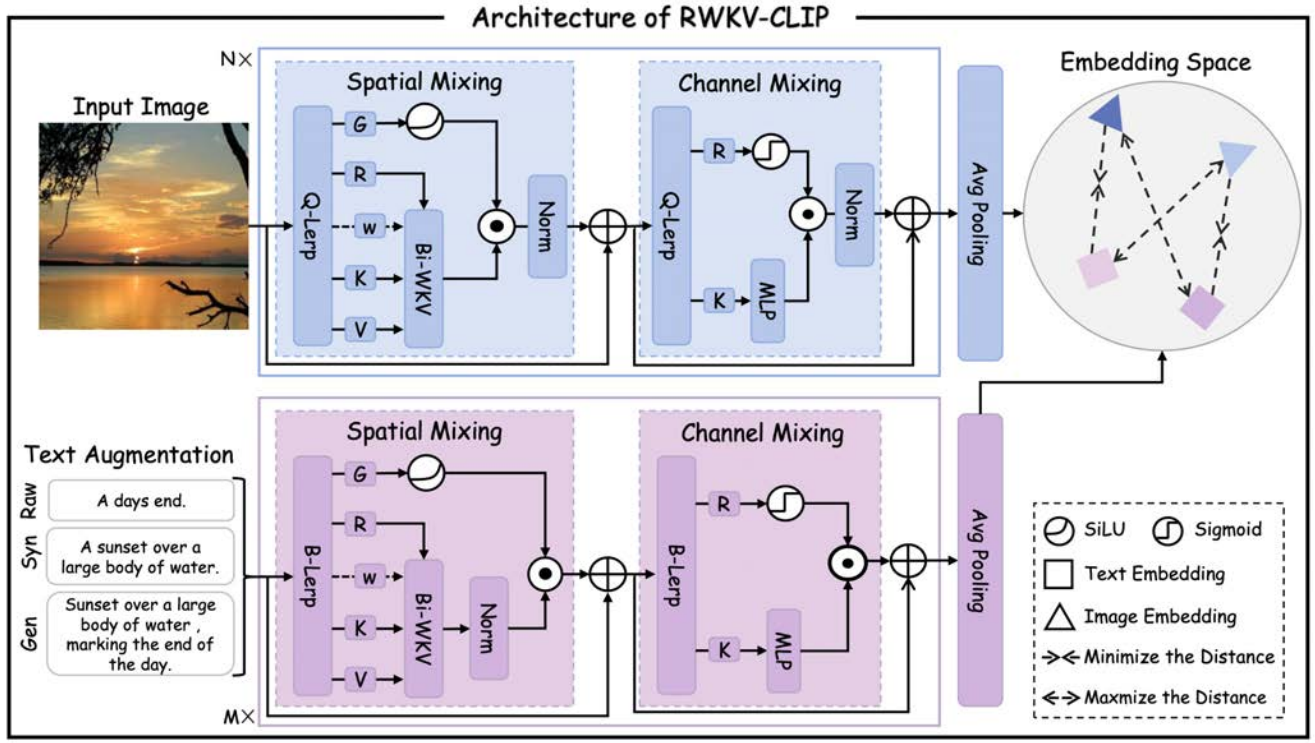
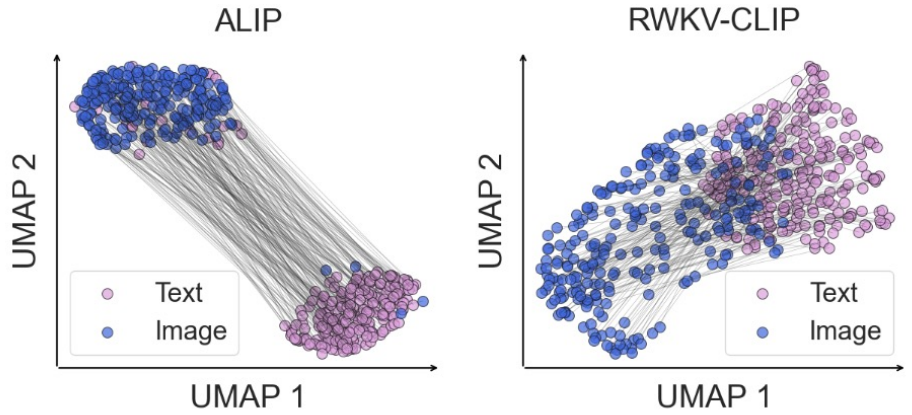


Figure 3: The architecture of RWKV-CLIP, which consists of M× and N× RKWV-driven blocks followed by an average pooling layer.

et al., 2018). As shown in Fig. 7, we found that the representations learned by RWKV-CLIP exhibit clearer discriminability within the same modality. Additionally, compared to ALIP, RWKV-CLIP demonstrates closer distances in the image-text modality space, indicating superior cross-modal alignment performance.



RWKV-SAM 分割一切模型

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

https://arxiv.org/pdf/2406.19369
https://github.com/HarborYuan/ovsam

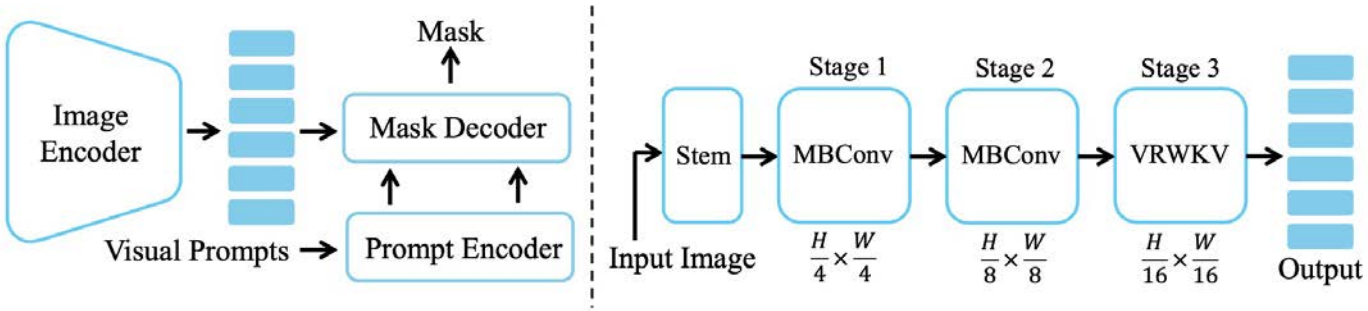
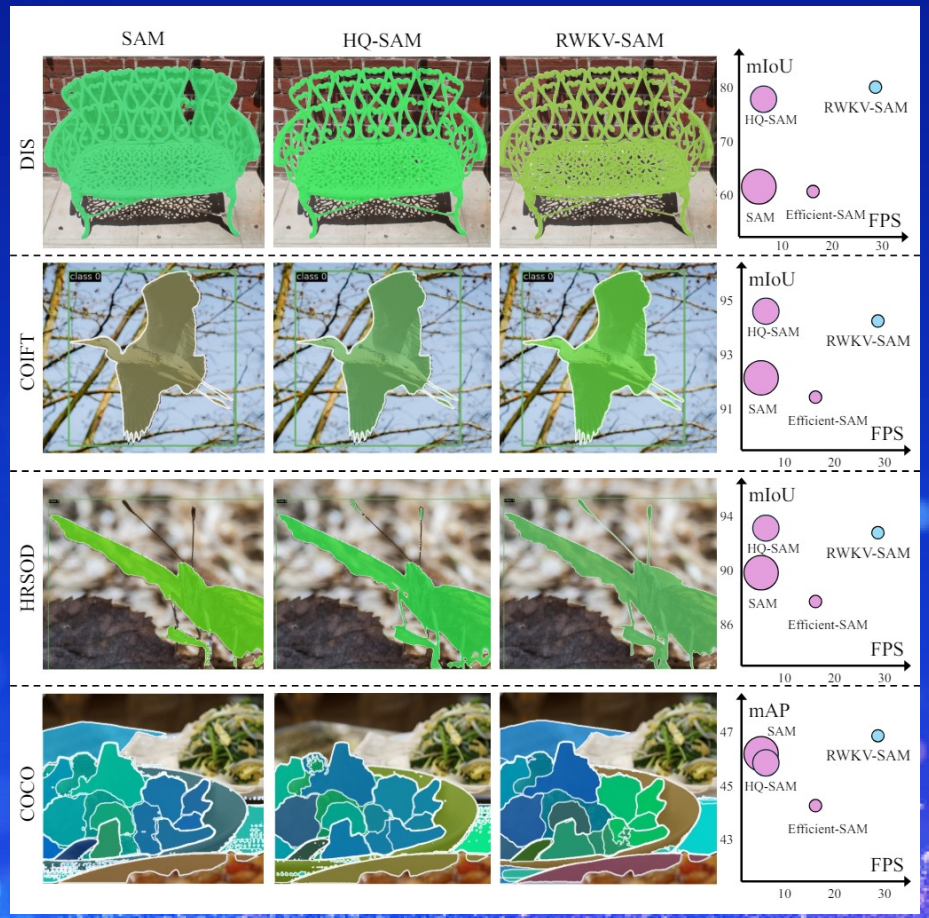


Figure 2: (Left) Overview of our RWKV-SAM. RWKV-SAM contains an image encoder, a prompt encoder, and a mask decoder. (Right) The efficient segmentation backbone architecture. The first two stages use the MBConv blocks, and the third uses the VRWKV blocks.

find that under the efficient segmentation setting of high-resolution image inputs, RWKV runs faster than Mamba. Thus, we aim to explore RWKV architecture as our backbone.



RWKV-TS 时序预测

https://github.com/howard-hou/RWKV-TS
https://arxiv.org/pdf/2401.09093

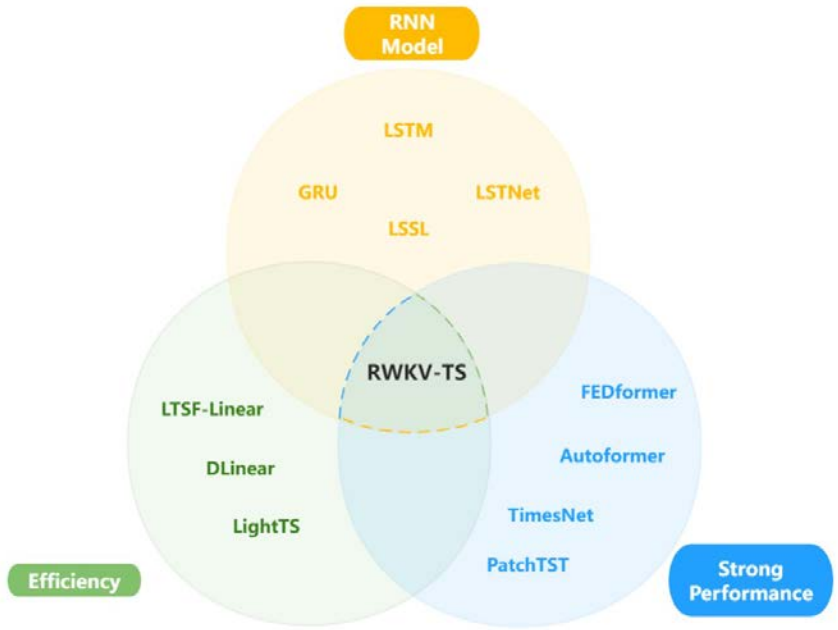


Figure 1: RWKV-TS is a time-series RNN-based model that achieves both strong performance and efficiency simultaneously. In contrast, other RNN models are considered to perform poorly in both aspects for time-series tasks.

Methods	RWKV-TS		TimesNet		ETSformer		LightTS		DLinear	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	**0.231**	**0.266**	0.259	0.287	0.271	0.334	0.261	0.312	0.249	0.300
ETTh1	0.433	0.445	0.458	0.450	0.542	0.510	0.491	0.479	**0.423**	**0.437**
ETTh2	**0.375**	**0.412**	0.414	0.427	0.439	0.452	0.602	0.543	0.431	0.447
ETTm1	0.376	0.401	0.400	0.406	0.429	0.425	0.435	0.437	**0.357**	**0.378**
ETTm2	0.287	0.338	0.291	0.333	0.293	0.342	0.409	0.436	**0.267**	**0.334**
ILI	**1.910**	**0.925**	2.139	0.931	2.497	1.004	7.382	2.003	2.169	1.041
ECL	**0.159**	**0.253**	0.192	0.295	0.208	0.323	0.229	0.329	0.166	0.263
Traffic	**0.398**	**0.276**	0.620	0.336	0.621	0.396	0.622	0.392	0.434	0.295
Average	**0.521**	**0.414**	0.596	0.433	0.662	0.473	1.303	0.616	0.562	0.436

Methods	FEDformer		PatchTST		Stationary		Autoformer		Informer	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	0.309	0.360	**0.225**	**0.264**	0.288	0.314	0.338	0.382	0.634	0.548
ETTh1	0.440	0.460	**0.413**	**0.430**	0.570	0.537	0.496	0.487	1.040	0.795
ETTh2	0.437	0.449	**0.330**	**0.379**	0.526	0.516	0.450	0.459	4.431	1.729
ETTm1	0.448	0.452	**0.351**	**0.387**	0.481	0.456	0.588	0.517	0.961	0.734
ETTm2	0.305	0.349	**0.255**	**0.315**	0.306	0.347	0.327	0.371	1.410	0.810
ILI	2.847	1.144	**1.443**	**0.798**	2.077	0.914	3.006	1.161	5.137	1.544
ECL	0.214	0.327	0.161	0.253	0.193	0.296	0.227	0.338	0.311	0.397
Traffic	0.610	0.376	**0.390**	**0.264**	0.624	0.340	0.628	0.379	0.764	0.416
Average	0.701	0.489	**0.446**	**0.386**	0.633	0.465	0.757	0.511	1.836	0.871

Table 3: Long-term forecasting task. All the results are averaged from 4 different prediction lengths, that is {24, 36, 48, 60} for ILI and {96, 192, 336, 720} for the others. Bold black is the best, red is the second best.

SDiT 基于 RWKV 的类脑模型

https://arxiv.org/pdf/2402.11588

SDiT: Spiking Diffusion Model with Transformer

Shu Yang, Hanzhi Ma, *Member, IEEE*, Chengting Yu, Aili Wang, *Member, IEEE*, Er-Ping Li, *Fellow, IEEE*

***Abstract*—Spiking neural networks (SNNs) have low power consumption and bio-interpretable characteristics, and are considered to have tremendous potential for energy-efficient computing. However, the exploration of SNNs on image generation tasks remains very limited, and a unified and effective structure for SNN-based generative models has yet to be proposed. In this paper, we explore a novel diffusion model architecture within spiking neural networks. We utilize transformer to replace the commonly used U-net structure in mainstream diffusion models. It can generate higher quality images with relatively lower computational cost and shorter sampling time. It aims to provide an empirical baseline for research of generative models based on SNNs. Experiments on MNIST, Fashion-MNIST, and CIFAR-10 datasets demonstrate that our work is highly competitive compared to existing SNN generative models.**

***Index Terms*—Image generation, deep learning, spiking neural network**

I. Introduction

SPIKING neural networks (SNNs) are considered to be the third generation of neural networks with higher biological interpretability, event-driven properties, and lower power consumption, and thus have the potential to become competitive alternatives to Artificial Neural Networks (ANNs) in the future. In SNNs, all information is encoded in spike sequences, enabling SNNs to perform accumulative operations at lower power budgets for energy efficiency.

SNNs trained with deep learning techniques, especially surrogate gradient learning methods [1], have shown promising results on basic tasks like image classification and segmentation [2]. However, the application of SNNs on more complex computer vision tasks, especially generative models, has been limited.

Recently, diffusion models have achieved significant suc-

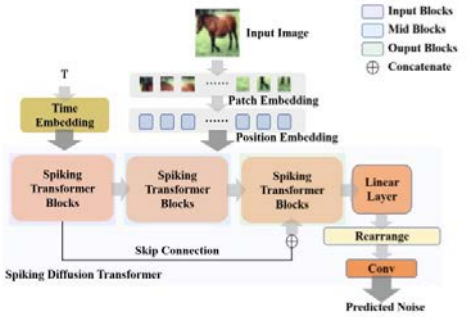


Fig. 1. Diagram of SDiT architecture, illustrating the flow from input time and patch embeddings through multiple spiking transformer blocks with skip connections, culminating in a final processing stage with linear and convolutional layers for predicted noise generation.

In this work, we propose Spiking Diffusion Transformer (SDiT), a novel SNN diffusion model architecture based on transformer. It demonstrates superior image generation potential for SNNs. We employ an efficient self-attention : RWKV[11], and introduce the Reconstruction Module, a specially designed module aimed at supplementing information lost after the firing of spiking neurons, thereby enhancing the quality of the reconstructed image. Comprehensive experiments on MNIST [12], Fashion-MNIST [13] and CIFAR-10 [14] show that SDiT has great competitiveness among the existing image generation models based on SNNs.

TABLE I
RESULTS ON DIFFERENT DATASETS.

Dataset	Model	Time Steps	FID	IS
MNIST	SGAD	16	69.64	-
	FSVAE	16	97.06	6.209
	Spiking-Diffusion	16	37.50	-
	SDDPM	4	29.48	-
	SDiT	4	**5.54**	2.452
Fashion-MNIST	SGAD	16	165.42	-
	FSVAE	16	90.12	4.551
	Spiking-Diffusion	16	91.98	-
	SDDPM	4	21.38	-
	SDiT	4	**5.49**	4.549
CIFAR-10	SGAD	16	181.50	-
	FSVAE	16	175.50	2.945
	Spiking-Diffusion	16	120.50	-
	SDDPM	4	16.89	7.655
	SDiT	4	**22.17**	4.080

TABLE II
ABLATION EXPERIMENTS EVALUATING FID ON DIFFERENT DATASETS.

Model	Reconstruction Module	MNIST	Fashion-MNIST
SDiT	✗	224.66	111.52
SDiT	✓	5.54	5.49

RWKV 生态 / 知名使用者

Fabrice Bellard

法国计算机科学家，传奇程序员

FFmpeg QEMU tinyg amarisoft

压缩软件 ts_zip 使用 RWKV-4

https://bellard.org/ts_zip/ts_zip-2024-03-02-win64.zip

实现了速度与压缩率之间的良好平衡。

文件	初始大小 (bytes)	XZ Utils (bytes)	(bpb)	ts_zip (bytes)	(bpb)
alice29.txt	152089	48492	2.551	21713	1.142
book1	768771	261116	2.717	137477	1.431
enwik8	100000000	24865244	1.989	13825741	1.106
enwik9	1000000000	213370900	1.707	135443237	1.084
linux-1.2.13.tar	9379840	1689468	1.441	1196859	1.021

RWKV 生态 / 海外创业者

Recursal

离线版 AI 小镇，120 个 Agent

Andrej Karpathy and Julian Bilcke liked

martin_casado
@martin_casado

Incredible work by @picocreator and team. They have a fine tuned RWKV v5 3B model working on a local macbook pro in AI town (video of packed level using it below :)

The model is here: huggingface.co/recursal/rwkv-...

Writeup to come from that team. Can't wait to try!

MidReal

MIT 的创业团队

/Start
/Featured World: Harry Potter
/Scenario: Harry refuses to enter Hogwarts and becomes a Muggle scientist.

MidReal

pygmalion

端侧模型的 Agent

RWKV性能评测结果

英文效果仅次于最好的LLaMA-8B和Mistral-7B，多语言能力最佳

lm-evaluation-harness	params	LAMBADA	English	LAMBADA	PIQA	StoryCloze16	Hellaswag	WinoGrande	arc_challeng	arc_easy	headQA_en	openbookQA	sciq	ReCoRD	COPA	MultiLang	xLBD	xSC	xWG	xCOPA
model	B	ppl	avg%	acc	acc	acc	acc_norm	acc	acc_norm	acc	acc_norm	acc_norm	acc	em	acc	avg%	acc	acc	acc	acc
pythia-6.9b-v0	6.86	4.30	65.1%	67.9%	74.5%	73.0%	63.9%	61.4%	35.3%	67.0%	38.3%	38.2%	90.3%	86.5%	85.0%	50.8%	35.2%	52.1%	61.6%	54.4%
RWKV-4 "Dove" World v1	7.52	3.93	65.8%	70.2%	75.3%	75.6%	65.3%	62.0%	36.6%	68.0%	35.8%	39.8%	91.2%	84.1%	86.0%	57.4%	41.6%	59.5%	68.7%	60.0%
RWKV-5 "Eagle" World v2	7.52	3.36	70.2%	74.2%	77.3%	79.7%	70.8%	68.4%	46.1%	74.9%	41.3%	41.2%	95.1%	88.5%	85.0%	61.4%	47.8%	62.1%	73.5%	62.4%
RWKV-6.0 "Finch" World v2.1	7.63	3.22	71.8%	75.2%	78.7%	79.7%	75.1%	70.0%	46.3%	76.8%	41.4%	44.8%	95.2%	88.9%	89.0%	62.9%	49.1%	62.8%	76.5%	63.4%
Mistral-7B-v0.1	7.24	3.18	74.7%	75.7%	80.6%	80.8%	81.0%	74.1%	53.7%	81.1%	46.5%	44.2%	95.9%	91.4%	91.0%	58.2%	46.0%	57.2%	73.8%	55.8%
Llama-3-8b	8.03	3.09	73.8%	75.6%	79.5%	79.6%	79.1%	72.8%	53.4%	80.0%	43.8%	45.0%	96.1%	91.7%	89.0%	60.6%	44.5%	61.9%	74.8%	61.4%
Llama-2-7b	6.74	3.40	71.0%	73.9%	78.1%	78.4%	76.0%	69.1%	46.3%	76.3%	40.3%	44.2%	93.7%	90.4%	86.0%	56.6%	45.0%	55.6%	69.4%	56.7%
falcon-7b	6.92	3.37	70.7%	74.5%	79.4%	78.8%	76.4%	67.3%	43.6%	74.7%	40.0%	44.2%	94.0%	89.2%	86.0%	55.8%	45.5%	53.8%	68.0%	56.0%
mpt-7b-8k	6.65	3.29	70.5%	72.9%	78.7%	77.7%	75.0%	68.8%	43.1%	75.9%	40.0%	43.8%	94.4%	88.9%	87.0%	55.6%	44.3%	55.3%	67.9%	55.1%
OLMo-7B	6.89	4.13	69.6%	68.9%	78.8%	78.4%	75.5%	66.5%	40.1%	73.4%	38.8%	42.8%	92.7%	89.0%	90.0%	51.2%	35.2%	53.1%	62.7%	53.8%
open_llama_7b_v2	6.74	3.82	69.4%	71.6%	78.7%	77.6%	74.5%	65.9%	41.4%	72.4%	37.9%	40.8%	93.8%	88.7%	89.0%	54.0%	40.9%	53.9%	66.6%	54.5%
Qwen1.5-7B	7.72	3.78	68.9%	70.8%	78.5%	76.8%	76.9%	66.2%	42.8%	71.1%	38.0%	41.6%	92.0%	88.3%	84.0%	56.8%	38.9%	58.1%	70.7%	59.5%
RedPajama-INCITE-7B-Base	6.86	3.92	68.3%	71.4%	77.2%	75.5%	70.4%	64.3%	39.4%	72.3%	38.1%	40.8%	92.8%	89.5%	88.0%	51.2%	37.2%	53.3%	61.7%	52.6%
lm-evaluation-harness	params	LAMBADA	English	LAMBADA	PIQA	StoryCloze16	Hellaswag	WinoGrande	arc_challeng	arc_easy	headQA_en	openbookQA	sciq	ReCoRD	COPA	MultiLang	xLBD	xSC	xWG	xCOPA

RWKV性能评测结果

语言建模能力就是压缩能力，用新数据衡量模型的泛化能力

https://github.com/jellyfish042/uncheatable_eval

Name	Parameters Count (B)	Average (lower=better)	ao3 english	bbc news	wikipedia english	arxiv computer science	arxiv physics	github cpp	github python
Meta-Llama-3-8B	8.030	7.286	10.605	8.221	7.981	7.742	7.676	4.287	4.489
Mistral-7B-v0.1	7.242	7.581	10.653	8.251	8.229	7.982	8.040	4.854	5.057
RWKV-x060-World-7B-v2.1-20240507-ctx4096	7.636	7.839	10.513	8.686	8.566	8.235	8.208	5.133	5.529
OLMo-1.7-7B-hf	6.888	7.870	11.113	8.627	8.713	8.131	8.163	5.015	5.330
Qwen1.5-7B	7.721	7.911	11.218	9.090	9.170	8.078	8.055	4.864	4.904
RWKV-5-World-7B-v2-20240128-ctx4096	7.518	7.922	10.598	8.816	8.659	8.312	8.267	5.203	5.600
mpt-7b	6.649	7.953	11.286	8.678	8.533	8.289	8.403	5.016	5.463
Llama-2-7b-hf	6.738	7.995	10.955	8.487	8.458	8.464	8.575	5.266	5.757
falcon-7b	6.922	8.298	10.861	8.690	8.888	8.689	9.005	5.805	6.148
aya-23-8B	8.028	8.492	11.804	8.959	9.255	8.837	9.205	5.593	5.791
pythia-6.9b-v0	6.857	8.500	11.607	9.349	9.283	8.797	8.524	5.582	6.361
mamba-7b-rw	6.947	9.765	10.909	8.542	8.768	8.755	9.073	11.092	11.217

RWKV性能测试结果

能耗只有 LLaMa 的一半

model	gpu	task	joule_per_token	token_per_joule
RWKV/rwkv-raven-7b	A100	chat	3.43	0.29
FreedomIntelligence/phoenix-inst-ch	A100	chat	4.29	0.23
H2OAI/H2OGPT-oasst1-7B	A100	chat	4.48	0.22
Salesforce/xgen-7b-8k-inst	A100	chat	4.53	0.22
StabilityAI/stablelm-tuned-alpha-7b	A100	chat	4.61	0.22
project-baize/baize-v2-7B	A100	chat	4.87	0.21
tatsu-lab/alpaca-7B	A100	chat	5.15	0.19
BAIR/koala-7b	A100	chat	5.15	0.19
togethercomputer/RedPajama-INCITE-7	A100	chat	5.31	0.19
LMSys/vicuna-7B	A100	chat	5.38	0.19
Neutralzz/BiLLa-7B-SFT	A100	chat	5.45	0.18
MetaAI/Llama-7B	A100	chat	5.53	0.18

RWKV性能测试

上下文长度

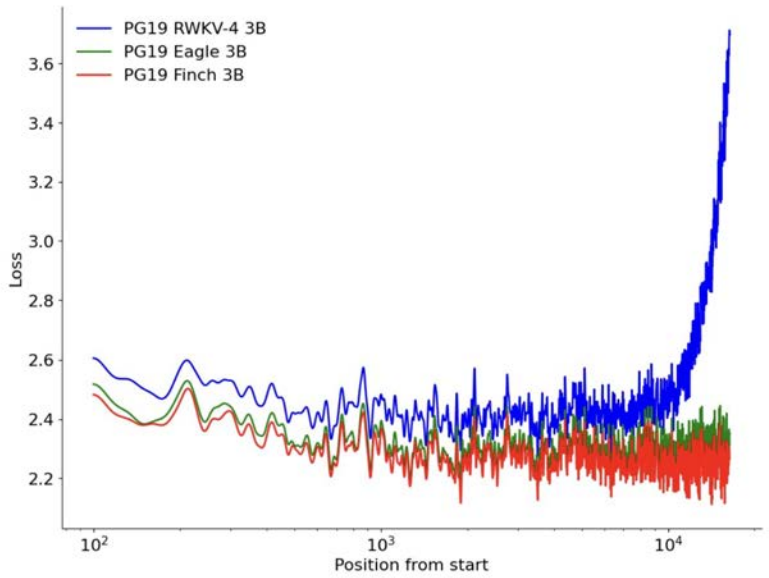
ctx4k 训练的 RWKV-6
可良好适应到 ctx20k 以上

内存占用低

RWKV-6 内存占用
比 Flash Attention 少 40%

MQAR 优

RWKV-6 在 MQAR 测试中
有显著优势



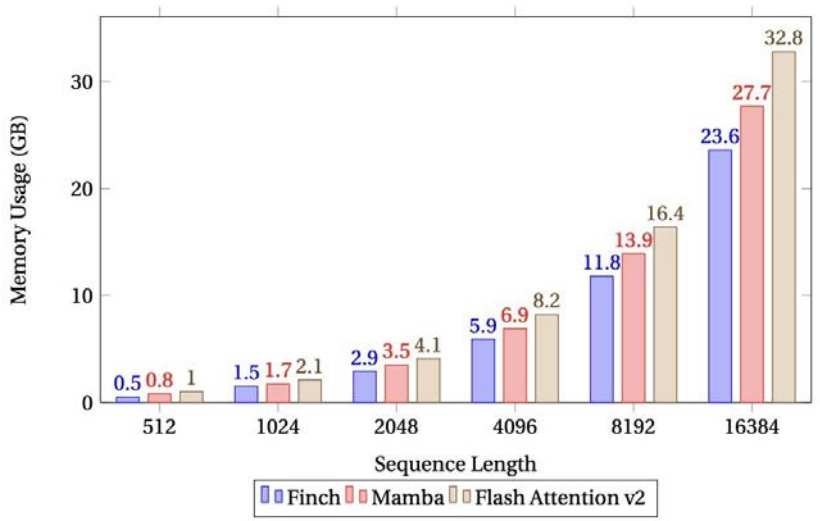
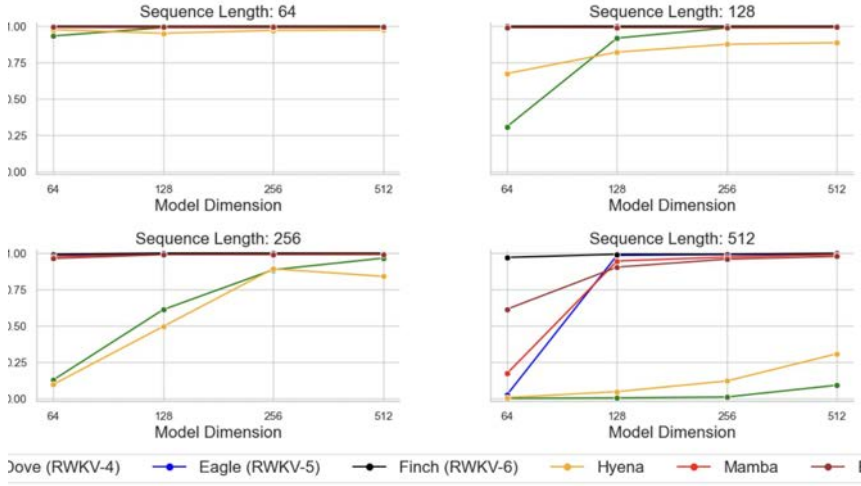


Figure 6: Memory Usage vs. Sequence Length (A100 80GB)



e 4: MQAR tasks. An increase in sequence length correlates with increased task diffic

THANKS