YOU ONLY SCAN ONCE: EFFICIENT MULTI-DIMENSION SEQUENTIAL MODELING WITH LIGHTNET

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear attention mechanisms have gained prominence in causal language models due to their linear computational complexity and enhanced speed. However, the inherent decay mechanism in linear attention presents challenges when applied to multi-dimensional sequence modeling tasks, such as image processing and multi-modal learning. In these scenarios, the utilization of sequential scanning to establish a global receptive field necessitates multiple scans for multi-dimensional data, thereby leading to inefficiencies. This paper identifies the inefficiency caused by a "multiplicative" linear recurrence and proposes an efficient alternative "additive" linear recurrence to avoid the issue, as it can handle multi-dimensional data within a single scan. We further develop an efficient multi-dimensional sequential modeling framework called LightNet based on the new recurrence. Moreover, we present two new multi-dimensional linear relative positional encoding methods, MD-TPE and MD-LRPE to enhance the model's ability to discern positional information in multi-dimensional scenarios. Our empirical evaluations across various tasks, including image classification, image generation, bidirectional language modeling, and autoregressive language modeling, demonstrate the efficacy of LightNet, showcasing its potential as a versatile and efficient solution for multi-dimensional sequential modeling.

A APPENDIX

A.1 PROOF OF EQ 4

Note that

$$A_t y_t = A_t a_t y_{t-1} + A_t x_t = A_{t-1} y_{t-1} + A_t x_t,$$
$$A_t y_t - A_{t-1} y_{t-1} = A_t x_t,$$
$$\dots,$$
$$A_2 y_2 - A_1 y_1 = A_2 x_2.$$

By summing up, we can obtain:

$$A_t y_t - A_1 y_1 = \sum_{s=2}^{t} A_s c x_s, y_t A_t = \sum_{s=1}^{t} A_s x_s, y_t = \sum_{s=1}^{t} \frac{A_s}{A_t} x_s.$$

A.2 MORE EXPERIMENTS

In this section, we provide additional experimental results. In Table 1, we show the performance of LightNet under the Commonsense Reasoning Tasks. In Table 3, we present the effects of LightNet on image generation tasks across various sizes.

Table 1: **Performance Comparison on Commonsense Reasoning Tasks.** PS, T, HS, WG stand for parameter size (billion), tokens (billion), HellaSwag, and WinoGrande, respectively.

Model	P	T	PIQA	HS	WG	ARC-e	ARC-c	OBQA	AVG
OPT	2.7	300	73.83	60.60	61.01	60.77	31.31	35.2.0	53.79
Pythia	2.8	300	74.10	59.31	59.91	64.14	33.02	35.60	54.35
BLOOM	3.0	350	70.57	54.53	58.48	59.43	30.38	32.20	50.93
RWKV-4	3.0	-	72.42	58.75	57.30	62.92	35.15	36.20	53.79
LightNet	2.9	100	73.56	55.47	57.77	61.57	32.94	33.60	52.49
LightNet	3	300	75.14	60.00	59.75	65.99	33.87	35.80	55.09

Table 2: Performance comparison for image generation task on ImageNet1k, where LightNet use 1 scan, Tnl/RetNet and Hgrn2 use 2 scan.

Model	50K	100K	150K	200K	250K	300K	350K	400K
LightNet-B/8	170.79	146.43	134.63	127.31	122.18	118.50	115.40	113.02
Tnl/RetNet-S/8	178.96	150.09	136.36	127.92	122.77	118.92	115.64	113.36
Hgrn2-S/8	182.75	152.13	140.94	133.95	129.14	125.78	123.27	121.08

A.3 CONFIGURATIONS

In this section, we provide training configurations for all experiments. The configuration for Bidirectional Language Modeling is the same as (Geiping & Goldstein, 2022), while the configurations for the other experiments are as shown in Table 4, 5, 6, 7. We use Pytorch (Paszke et al., 2019) and A100 for training.

REFERENCES

Jonas Geiping and Tom Goldstein. Cramming: Training a language model on a single gpu in one day. *arXiv preprint arXiv:2212.14034*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Table 3: Performance Metrics Across Different LightNet Configurations

Model	50K	100K	150K	200K	250K	300K	350K	400K
LightNet-S/8	192.79	172.23	161.23	154.34	150.25	147.40	145.27	143.31
LightNet-S/4	167.33	132.89	118.77	110.88	105.15	101.25	97.56	94.90
LightNet-S/2	145.66	119.20	104.90	94.45	87.18	82.41	78.63	75.61
DiT-S/2	-	-	-	-	-	-	-	67.16
LightNet-B/8	170.79	146.43	134.63	127.31	122.18	118.50	115.40	113.02
LightNet-B/4	126.37	93.86	81.44	74.11	68.80	65.09	62.34	59.81
LightNet-B/2	104.19	74.27	59.60	51.22	45.70	41.65	38.60	36.45
DiT-B/2	-	-	-	-	-	-	-	42.76
LightNet-L/8	157.76	130.29	116.06	107.50	101.10	96.47	92.79	89.51
LightNet-L/4	104.18	77.02	64.55	56.16	49.99	45.58	41.91	37.54
LightNet-L/2	84.38	48.98	35.32	28.05	23.75	21.06	18.94	17.42
DiT-L/2	-	-	-	-	-	-	-	24.37
LightNet-XL/8	158.75	129.23	114.72	105.75	99.35	94.53	90.66	87.22
LightNet-XL/4	101.39	70.84	56.75	48.04	42.04	37.43	34.16	31.51
LightNet-XL/2	79.22	45.46	31.61	25.55	21.37	18.74	16.84	15.52
DiT-XL/2	-	-	-	-	-	-	-	19.20

Table 4: **Comprehensive Configurations of the Model and Training Procedures for LightNet Experiments** "Total batch size" means $\mathrm{batch\_per\_gpu} \times \mathrm{update\_freq} \times \mathrm{num\_gpus}$; "ALM" stands for Autoregressive Language Model; "IM" stands for Image Modeling, "IG" stands for image generation.

	ALM	IM	IG
Dataset	WikiText-103	ImageNet-1k	ImageNet-1k
Tokenizer method	BPE	-	-
Src Vocab size	50265	-	-
Sequence length	512	-	-
Total batch size	128	2048	256
Number of updates/epochs	50k updates	300 epochs	80 epochs
Warmup steps/epochs	4k steps	20 epochs	-
Peak learning rate	5e-4	5e-4	1e-4
Learning rate scheduler	Inverse sqrt	Cosine	-
Optimizer	Adam	Adamw	Adamw
Adam ϵ	1e-8	1e-8	1e-8
Adam (β_1, β_2)	(0.9, 0.999)	(0.9, 0.98)	(0.9, 0.98)
Weight decay	0.1	0.1 for Base, else 0.05	0
Gradient clipping	-	5.0	-
GPUS	4	8	8

Table 5: **Configurations for LLM**

Params(B)	Layers	Hidden Dim	L.R.	Batch Size Per GPU	SeqLen	GPUs
0.15	15	768	3.00E-04	26	2048	8
0.385	26	1024	3.00E-04	15	2048	8
1.0	18	2048	3.00E-04	10	2048	16
2.9	36	2560	3.00E-04	36	2048	48

Table 6: **Model Configurations for Image Generation task.**

Model	Layers	Hidden Dim	Heads	Params
LightNet-S	18	384	6	33M
LightNet-B	18	768	6	131M
LightNet-L	36	1024	16	470M
LightNet-XL	42	1152	16	680M

Table 7: **Model Configurations for Image Classification task.**

Model	Layers	Hidden size	Heads	Params
LightNet-T	12	192	6	6.0M
LightNet-S	12	384	16	22.6M
LightNet-B	12	768	16	87.7M