

Transformer Architecture 대안

(보편적 LLM 성능 강화: Transformer & Other)

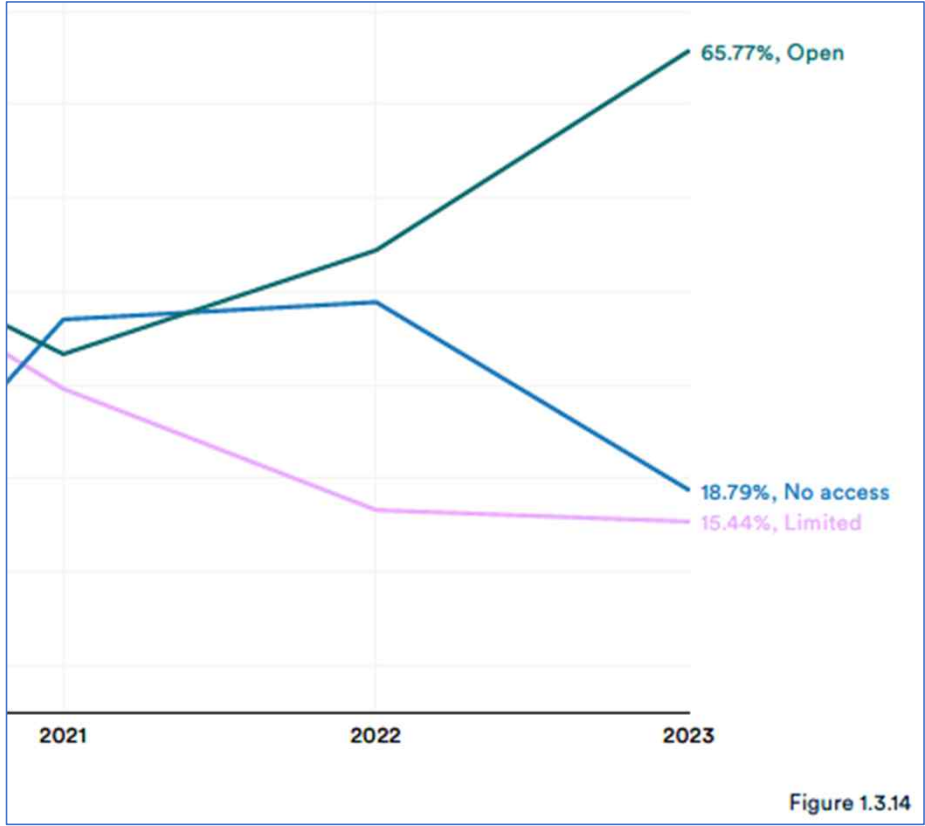
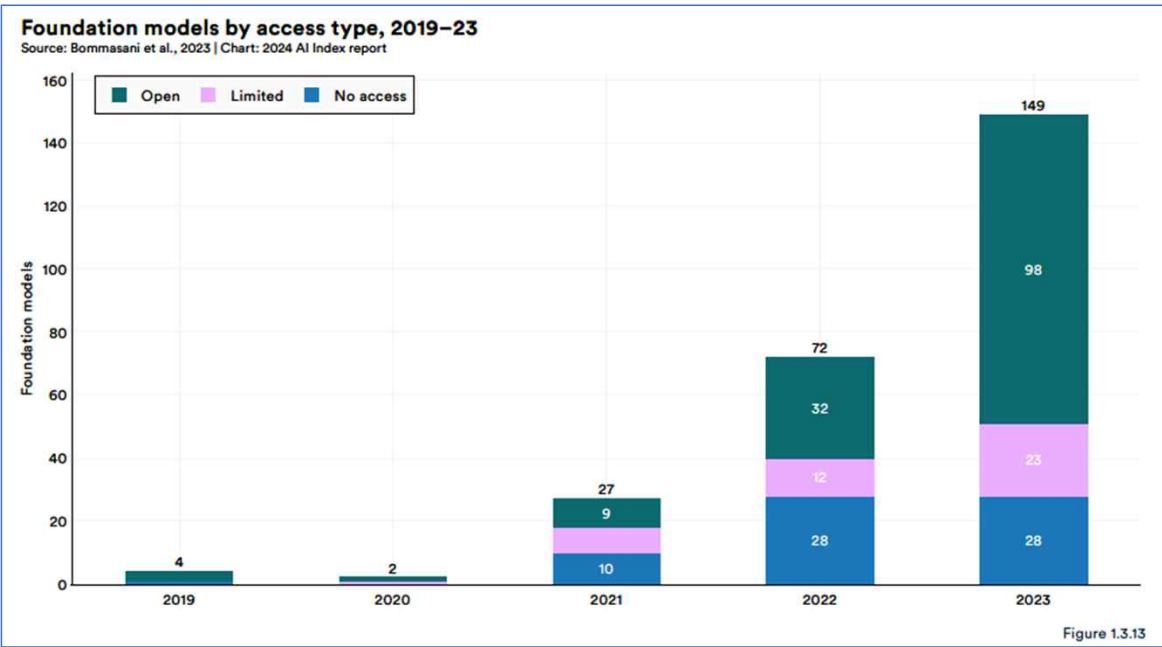
GEN-AIA 개발을 위한 기초 연구



Stanford HAI Releases 2024 Artificial Intelligence Index Report

Source: <https://aiindex.stanford.edu/report/>

2. More foundation models and more open foundation models. In 2023, a total of 149 foundation models were released, more than double the amount released in 2022. Of these newly released models, 65.7% were open-source, compared to only 44.4% in 2022 and 33.3% in 2021.



<https://velog.io/@tobigs-nlp/Attention-is-All-You-Need-Transformer>

[Transformer Survey] #1 Transformer_Basic



트랜스포머 공동 저자들: 7명 (사진:구글)

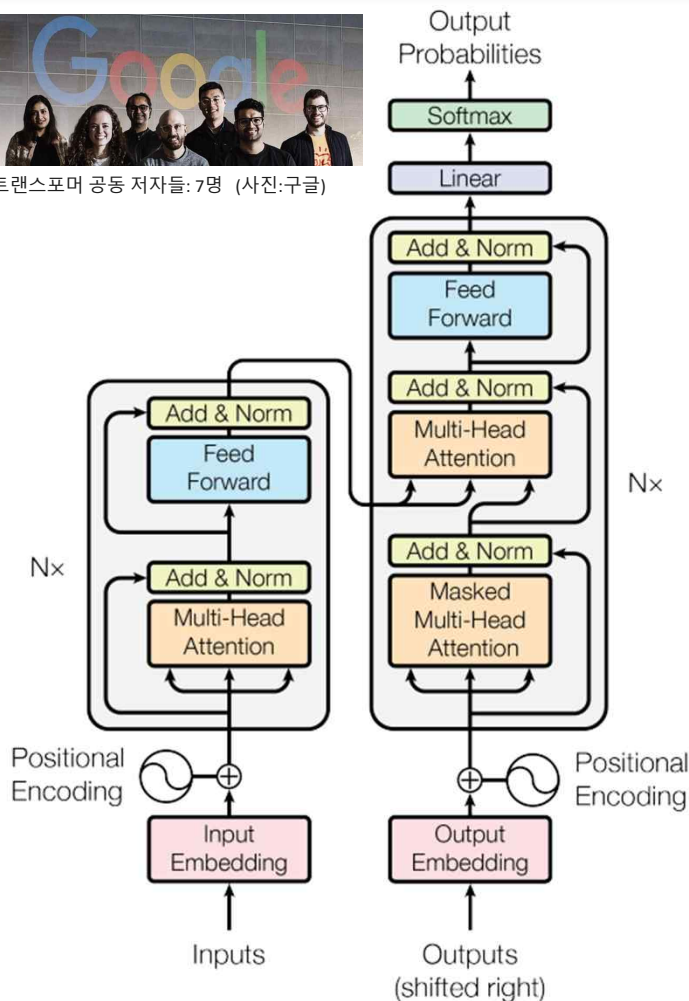


Figure 1: The Transformer - model architecture.

Attention/Transformer 시각화로 설명

Transformer는 2017년 구글이 발표한 논문인 "Attention is All You Need"에서 나온 모델로 기존의 seq2seq의 구조인 인코더, 디코더를 따르면서도, RNN을 사용하지 않고 Attention만으로 구현한 모델입니다. **Transformer가 출현함으로써 BERT, GPT, T5, Llama 등과 같은 아키텍처가 발전하는 기반이 마련**되었습니다.

1. Abstract

• Sequence transduction models는 인코더와 디코더를 포함하는 복잡한 recurrent 또는convolutional neural networks에 기초합니다. 본 논문에서는 오로지 attention mechanism에만 기초한 새로운 간단한 network 구조를 제안합니다.

• Machine translation 작업에 대한 실험 결과에 따르면, 이 모델은 병렬 처리가 가능하고 학습시간이 훨씬 덜 소요됩니다. WMT English-to-German 번역 대회에서 28.4 BLEU를 달성하면서 기존 최고 모델보다 2 BLEU 이상 향상하며 SOTA를 달성했습니다. 또한 다른 task에서도 잘 일반화됨을 보여줍니다

•

1. Introduction

• RNN, LSTM, GRU는 language modeling, machine translation 등과 같은 시퀀스 모델링 문제에서 확고히 자리를 잡았습니다. 하지만 Recurrent 모델은 필연적으로 이전 결과를 입력으로 받는 순차적인 특성 때문에 병렬 처리를 배제합니다. 최근의 연구는 factorization tricks와 conditional computation을 통해 연산 효율의 대폭적인 향상을 달성했지만 아직까지 순차적 계산의 근본적인 제약은 여전히 남아있다고 합니다.

• 본 논문은 recurrent 모델의 제약 사항들은 피하고 입력과 출력 사이의 dependency를 이끌어내기 위해 transformer를 제안합니다.

LLM 발전의 근간이 BERT 개념과 연관이 있나요 ?

ChatGPT 4o

네, LLM(Large Language Models)의 개발은 BERT(Bidirectional Encoder Representations from Transformers)에서 도입한 개념과 밀접한 관련이 있습니다. Google에서 개발한 BERT는 자연어 처리(NLP) 분야에 상당한 영향을 미쳤으며 최신 LLM을 포함한 많은 후속 모델에 영향을 미쳤습니다.

BERT의 핵심 개념이 LLM 개발에 영향을 미침

1. Transformer 아키텍처 :

• BERT는 Vaswani 등이 2017년에 처음 도입한 트랜스포머 아키텍처의 사용을 대중화했습니다. 이 아키텍처는 셀프 어텐션 메커니즘을 사용하여 입력 시퀀스를 병렬로 처리하는데, 이는 다양한 NLP 작업에 매우 효율적이고 효과적인 것으로 입증되었습니다.

• Transformer 아키텍처가 대량의 데이터를 처리하고 복잡한 패턴을 학습할 수 있는 능력은 현대 LLM 개발에 필수적입니다.

2. 양방향 맥락적 이해 :

• 텍스트를 단방향(좌우 또는 우좌)으로 처리하던 이전 모델과 달리 BERT는 양방향 학습을 도입했습니다. 즉, BERT는 양쪽 방향에서 맥락을 살펴보고 주변 맥락에 따라 단어의 의미를 보다 포괄적으로 이해할 수 있습니다.

• 양방향 맥락의 개념은 질의 응답, 감정 분석 등 다양한 업무에서 LLM의 성과를 개선하는 데 중요한 역할을 했습니다.

3. 사전 훈련 및 미세 조정 패러다임 :

• BERT는 대규모 코퍼스에서 모델을 사전 학습한 다음 특정 작업에 대해 미세 조정하는 방식으로 NLP에서 표준 관행이 되었습니다. 이 패러다임은 모델이 사전 학습 중에 일반 언어 표현을 학습한 다음 미세 조정 중에 특정 작업에 특화할 수 있도록 합니다.

• GPT(Generative Pre-trained Transformer) 시리즈, Llama 등을 비롯한 대부분의 최신 LLM은 이 접근 방식을 채택하여 대규모 데이터 세트를 활용하여 사전 학습을 수행한 다음 특정 응용 프로그램에 맞게 미세 조정합니다.

4. 마스크드 언어 모델링(MLM) :

• BERT는 문장의 특정 단어가 가려지고 모델이 이러한 가려진 단어를 예측하도록 훈련되는 마스크 언어 모델링을 도입했습니다. 이를 통해 모델은 단어 간의 깊은 맥락적 관계를 학습할 수 있습니다.

• 이 기술의 변형은 다른 모델에서도 맥락에 대한 이해를 높이고 언어 생성 능력을 향상시키기 위해 사용되었습니다.

후속 모델에 미치는 영향

• **GPT 시리즈** : GPT 모델은 자기 회귀 언어 모델링에 단방향 접근 방식을 사용하지 않, BERT가 도입한 트랜스포머 아키텍처의 이점을 여전히 누리고 있습니다. 또한, 대규모 사전 학습에 이어 미세 조정이라는 개념이 직접적인 영향을 미칩니다.

• **RoBERTa** : BERT의 최적화된 버전으로, Robustly Optimized BERT Approach의 약자로, 학습 방법론을 더욱 개선하여 BERT 개념의 영향을 보여줍니다.

• **T5(Text-To-Text Transfer Transformer)** : Google에서 개발한 T5는 모든 NLP 문제를 텍스트-텍스트 작업으로 처리하며, BERT에서 영감을 받은 변형기 아키텍처와 대규모 사전 학습을 활용합니다.

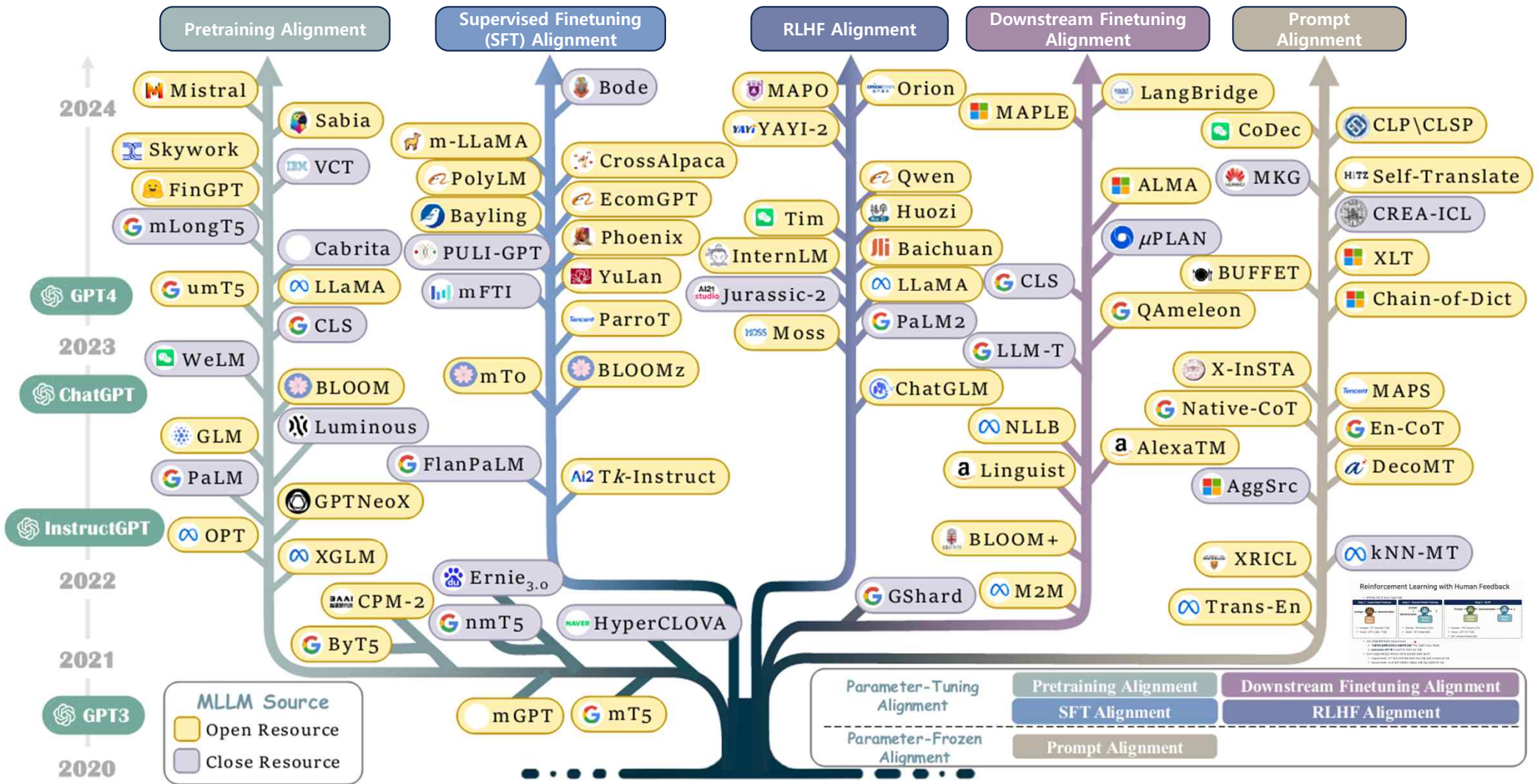
• **Llama 시리즈** : Meta의 이 모델도 변형기 아키텍처를 활용하고 BERT가 개척한 양방향 컨텍스트 이해와 대규모 학습 전략의 혜택을 누립니다.

• 오픈AI의 'GPT-4', 구글의 '제미나이', 앤트로픽의 '클로드'와 같은 대형언어모델(LLM)뿐만 아니라, **비어 모델의 기반에도 영향을 미침** (오픈AI의 '코텍스', 답마인드의 '알파폴드' 등)

결론

BERT는 LLM의 진화에서 기초적인 모델이었습니다. 트랜스포머 아키텍처, 양방향 컨텍스트 처리, 사전 학습/미세 조정 패러다임의 혁신은 보다 진보된 모델 개발을 위한 토대를 마련하여 NLP 분야에서 상당한 개선을 이끌었습니다. 이러한 원칙은 오늘날 최첨단 LLM의 설계와 학습을 뒷받침하고 있습니다

Source: Multilingual Large Language Model - A Survey of Resources, Taxonomy and Frontiers (2404.04925v1)



스케일업 AI 개발의 대안 → 메타인지 !!??



범용 인공지능 도달 전략 논쟁

•수정 2024-08-05 09:21 등록 2024-08-05 09:07

- 이스라엘 스타트업 AI21 랩스:
 - 1) SSM을 기반으로 하는 '맘바(Mamba)'와 트랜스포머 아키텍처의 최고의 특성을 결합한 LLM '잠바(Jamba)'를 출시
- 스타트업 심볼리카 : '심볼릭 AI(Symbolic AI)' 기법을 도입
 - 1) 트랜스포머 아키텍처에 기반한 LLM을 실행하는데 많은 비용이 드는 문제를 해결하기 위해 기호(Symbols)를 조작해 작업을 정의하는 새 기법을 도입
- 구글은 LLM 컨텍스트 창 길이를 무한확장할 수 있는 '인피니-어텐션(Infini-attention)' 기술을 공개.
- 메타도 막대한 양의 메모리 필요없이 컨텍스트 창을 수백만개의 토큰으로 확장 가능한 LLM '메갈로돈(Megalodon)' 모델을 공개.

Transformer의 대체 Architecture : SSM / S4 / RWKV / Hyena / JEPA

트랜스포머 (Transformer) 대안 아키텍처 및 SSM 발전/보완 기술						
논문 제목	발표 연도	발표자	인용 횟수 (대략)	공개된 모델 이름	사용 아키텍처	URL
1 Perceiver: General Perception with Iterative Attention	2021	Andrew Jaegle et al.	1,000+	Perceiver	Perceiver	Link
2 Longformer: The Long-Document Transformer	2020	Iz Beltagy, Matthew E. Peters, Arman Cohan	2,000+	Longformer	Longformer	Link
3 Linformer: Self-Attention with Linear Complexity	2020	Sinong Wang et al.	1,000+	Linformer	Linformer	Link
4 Reformer: The Efficient Transformer	2020	Nikita Kitaev, Lukasz Kaiser, Anselm Levskaya	2,500+	Reformer	Reformer	Link
5 Routing Transformer: Efficient Routing in Transformers with Dynamic Attention	2020	Swadheen Shukla, Sumeet Singh, William Fedus, Jeff Dean, Hugo Larochelle, Balaji Lakshminarayanan	300+	Routing Transformer	Routing Transformer	Link
6 Rethinking Attention with Performers	2021	Krzysztof Choromanski et al.	1,500+	Performer	Performer	Link
7 Generating Long Sequences with Sparse Transformers	2019	Rewon Child, Scott Gray, Alec Radford, Ilya Sutskever	1,200+	Sparse Transformer	Sparse Transformer	Link
8 Neural Turing Machines	2014	Alex Graves, Greg Wayne, Ivo Danihelka	5,000+	Neural Turing Machines (MANNs)	Memory-Augmented Neural Networks (MANNs)	Link
9 Compressive Transformers for Long-Range Sequence Modelling	2019	Jack W. Rae et al.	500+	Compressive Transformer	Compressive Transformer	Link
1 A New Approach to Linear Filtering and Prediction Problems	1960	Rudolf E. Kalman	40,000+	Kalman Filter	SSM (State-Space Model)	Link
2 Filtering and Smoothing of State Vector for Diffuse State-Space Models	2003	S.J. Koopman, J. Durbin	2,000+	N/A	SSM (State-Space Model)	Link
3 A Self-Organizing State-Space Model	1998	Genshiro Kitagawa	1,400+	N/A	SSM (State-Space Model)	Link
4 On Particle Methods for Parameter Estimation in State-Space Models	2014	Christophe Andrieu, Arnaud Doucet, Roman Holenstein	600+	N/A	SSM (State-Space Model)	Link
5 Time Series Analysis by State Space Methods	2012	J. Durbin, S.J. Koopman	2,500+	N/A	SSM (State-Space Model)	Link
6 Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality	2024	Tri Dao, Albert Gu	100+	Falcon Mamba	SSM (State-Space Model)	Link
7 Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling	2024	Various authors	50+	Samba	SSM (State-Space Model)	Link
8 Mamba: Linear-Time Sequence Modeling with Selective State Spaces	2023	Various authors	50+	Falcon Mamba	SSM (State-Space Model)	Link
9 Structured State Space Models (S4): Sequence Modeling with Structured State Spaces	2022	Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, Christopher Ré	1,000+	S4	S4 (Structured State Space Models)	Link
10 RWKV: Reinventing RNNs with Transformer-Level Performance	2023	BlinkDL	200+	RWKV	RWKV RNN과 Transformer의 SSM을 활용하여 긴 문맥을 처리	Link
11 Hyena Hierarchy: Towards General-Purpose Long-Context Sequence Modeling	2023	Mostafa Dehghani, Yi Tay, Vamsi Anubandhi, Dara Bahri, etc.	200+	Hyena	Hyena Hierarchy	Link
12 Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (I-JEPA)	2023	Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr	172	I-JEPA	JEPA	Link

아래는 S4, RWKV, Hyena Hierarchy와 관련된 논문들을 포함하여, 이전에 제공한 표 형식과 동일하게 정리한 표입니다. 이 표에는 각 논문과 관련된 공개된 모델 이름도 함께 포함되어 있습니다.

No.	논문 제목	발표 연도	발표자	인용 횟수 (대략)	공개된 모델 이름	URL
1	A New Approach to Linear Filtering and Prediction Problems	1960	Rudolf E. Kalman	40,000+	Kalman Filter	Link
2	Filtering and Smoothing of State Vector for Diffuse State-Space Models	2003	S.J. Koopman, J. Durbin	2,000+	N/A	Link
3	A Self-Organizing State-Space Model	1998	Genshiro Kitagawa	1,400+	N/A	Link
4	On Particle Methods for Parameter Estimation in State-Space Models	2014	Christophe Andrieu, Arnaud Doucet, Roman Holenstein	600+	N/A	Link
5	Time Series Analysis by State Space Methods	2012	J. Durbin, S.J. Koopman	2,500+	N/A	Link
6	Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality	2024	Tri Dao, Albert Gu	100+	Falcon Mamba	Link
7	Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling	2024	Various authors	50+	Samba	Link
8	Mamba: Linear-Time Sequence Modeling with Selective State	2023	Various authors	50+	Falcon Mamba	Link

SSM을 제외한 이유:
 SSM은 Transformer의 직접적인 기술적 대안보다는, 다른 응용 분야에서 사용되던 모델로, Transformer 이후의 생성형 AI 모델링 패러다임과는 다른 맥락에서 발전해왔습니다. 따라서 첫 질문에서는 Transformer와 관련한 기술적 대안 아키텍처로 SSM을 포함하지 않음

트랜스포머 (Transformer) 대안 아키텍처 및 SSM 발전보완 기술.xlsx

Transformer의 대체 Architecture : SSM / S4 / RWKV / Hyena / JEPA

1	A New Approach to Linear Filtering and Prediction Problems	1960	Rudolf E. Kalman		45474		Kalman Filter	SSM (State-Space Model)	Link
2	A Self-Organizing State-Space Model	1998	Genshiro Kitagawa	Institute of Statistical Mathematics (ISM), Tokyo	599 + SKP	N/A		SSM (State-Space Model)	Link
3	Filtering and Smoothing of State Vector for Diffuse State-Space Models	2003	S.J. Koopman, J. Durbin		150	N/A		SSM (State-Space Model)	Link
4	Time Series Analysis by State Space Methods	May 2012	J. Durbin, S.J. Koopman James Durbin, Siem Jan Koopman	Oxford Academy	5452 + SKP	N/A		SSM (State-Space Model)	Link
5	On Particle Methods for Parameter Estimation in State-Space Models	Dec. 2014	Nikolas Kantas, Arnaud Doucet, Sumeetpal S. Singh, Jan Maciejowski and Nicolas Chopin	Imperial College London, Univer.of Oxford & Cambridge, CREST-ENSAE and HEC Paris	552			SSM (State-Space Model)	Link
	HiPPO: Recurrent Memory with Optimal Polynomial Projections	Aug. 2020	Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, Christopher Re	Stanford University University at Buffalo	292			모델과는 무관.	Link
6	Efficiently Modeling Long Sequences with Structured State Spaces (S4) <i>(Important of HiPPO)</i>	Oct. 2021	Albert Gu, Karan Goel, and Christopher R'e	Stanford University	865 + SKP	S4		S4 (Structured State Space Models)	Link
	Structured State Space Models (S4): Sequence Modeling with Structured State Spaces	2022	Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, Christopher Ré		1,000+	S4		S4 (Structured State Space Models)	Link
7	The Falcon Series of Open Language Models	Nov. 2023	Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Etienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Nouné, Baptiste Pannier, Guilherme Penedo	The Falcon LLM Team	184	Falcon			Link
8	Mamba: Linear-Time Sequence Modeling with Selective State Spaces	Dec. 2023	Albert Gu*, Tri Dao*	Carnegie Mellon/ Princeton University	690 + SKP	Mamba		Selective State Spaces Model	Link
9	Jamba: A Hybrid Transformer-Mamba Language Model	Mar. 2024	Opher Lieber* Barak Lenz* Hofit Bata Gal Cohen Jhonathan Osin Itay Dalmedigos Erez Safahi Shaked Meir Yonatan Belinkov Shai Shalev-Shwartz Omri Abend Raz Alon Tomer Asida Amir Bergman Roman Glozman Michael Gokhman Avshalom Manevich Nir Ratner Noam Rozen Erez Schwartz Mor Zusman Yoav Shoham	AI21 Labs	50	Jamba		SSM (State-Space Model) + Transformer (Attention Mechanism)	Link
10	Transformers are SSMS: Generalized Models and Efficient Algorithms Through Structured State Space Duality	May, 2024	Tri Dao, Albert Gu	Princeton / Carnegie Mellon University	32	Mamba-2		SSM (State-Space Model)	Link
11	Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling	Jun. 2024	Liliang Ren ^{1,2*} Yang Liu ^{1*} Yadong Lu ^{1*} Yelong Shen ¹ Chen Liang ¹ Weizhu Chen ¹	Microsoft 2, University of Illinois at Urbana-Champaign	36	Samba		SSM (State-Space Model) + Attention Mechanism	Link

트랜스포머 아키텍처의 대안 = SSM ①-2

A Self-Organizing State-Space Model

Genshiro KITAGAWA **Genshiro Kitagawa** = Institute of Statistical Mathematics (ISM), Tokyo, Sept. 1998, 599회 인용

A self-organizing filter and smoother for the general nonlinear non-Gaussian state-space model is proposed. An expanded state-space model is defined by augmenting the state vector with the unknown parameters of the original state-space model. The state of the augmented state-space model, and hence the state and the parameters of the original state-space model, are estimated simultaneously by either a non-Gaussian filter/smoothing or a Monte Carlo filter/smoothing. In contrast to maximum likelihood estimation of model parameters in ordinary state-space modeling, for which the recursive filter computation has to be done many times, model parameter estimation in the proposed self-organizing filter/smoothing is achieved with only two passes of the recursive filter and smoother operations. Examples such as automatic tuning of dispersion and the shape parameters, adaptation to changes of the amplitude of a signal in seismic data, state estimation for a nonlinear state space model with unknown parameters, and seasonal adjustment with a nonlinear model with changing variance parameters are shown to exemplify the usefulness of the proposed method.

KEY WORDS: Bayesian estimation; Filtering; Likelihood; Nonlinear model; Parameter estimation; Self-tuning; Smoothing.

1. INTRODUCTION

Since the development of the Kalman filter in 1960, the state-space model has been widely used in many fields of engineering. However, use of the state-space model became popular in the statistical community only after it was introduced for stochastic system identification by autoregressive moving average (ARMA) modeling (Akaike 1974) and in the dynamic linear model (Harrison and Stevens 1976). By the 1980s, the state-space model was a popular tool for handling nonstationary time series (Harvey 1989; Kitagawa and Gersch 1984).

In the latter half of that decade, many statisticians were interested in the analysis of various types of nonstandard time series, for which the ordinary linear Gaussian state-space model cannot yield reasonable results. To handle such problems, several types of nonlinear non-Gaussian state-space models and related recursive filtering and smoothing algorithms were developed: the dynamic generalized linear model (Fahrmeir 1992; Smith and Miller 1986; West and Harrison 1997; West, Harrison, and Migon 1985) and various extensions of the Kalman filter (Fahrmeir and Kaufmann 1991; Frühwirth-Schnatter 1994; Kitagawa 1994; Meinhold and Singpurwalla 1989; Schnatter 1992), the Gibbs sampler-based method (Carlin, Polson, and Stoffer 1992), and the sequential imputations of Kong, Liu, and Wong (1994).

On the other hand, in earlier work (Kitagawa 1987) I proposed an alternative non-Gaussian filter and smoother that can yield the exact marginal posterior density of the state for fairly general types of state-space models. Because that method is based on numerical integration, its application is limited only to the models with relatively low state di-

mension (say, less than or equal to 4). Despite the development of various refinements of the integration method (e.g., Hodges and Hale 1993; Tanizaki 1993), they did not yield an essential solution to the problem of modeling nonstandard time series.

A Monte Carlo filter and smoother was shown in earlier work (Kitagawa 1993, 1996) which is applicable to very general state-space models forms. A similar "bootstrap filter" algorithm was proposed by Gordon, Salmond, and Smith (1993). The problem that Gordon et al. and I considered is in fact a signal extraction problem that is one-to-one with the smoothing problem (Kohn and Ansley 1988). Gordon et al. did not address smoothing. In Monte Carlo type methods, arbitrary non-Gaussian densities are approximated by many particles that can be considered realizations from the distributions. With the development of these algorithms, it is now possible to use high-dimensional nonlinear non-Gaussian state-space models for the analysis of complex time series.

Nevertheless, a very important question remained: How to operate it without knowledge of system parameters? (Solo 1989). In the statistical community, the maximum likelihood method is commonly used to address that problem. But for nonlinear or non-Gaussian state-space modeling, two factors sometimes render the maximum likelihood

Genshiro Kitagawa is Professor, Department of Prediction and Control, The Institute of Statistical Mathematics, Tokyo 106-8569, Japan. The author thanks Will Gersch of the University of Hawaii and Toshiyuki Higuchi of the Institute of Statistical Mathematics for their useful comments. Careful reading and valuable comments and suggestions by the editor, an associate editor, and the referee—in particular, on the accuracy of the estimates—are gratefully acknowledged. This work was partially supported by Grants-in-Aid for Scientific Research 06680338, 09680318, and 08558021 from the Japanese Ministry of Education, Science, Sports and Culture.

JOURNAL ARTICLE


A Self-Organizing State-Space Model

Genshiro Kitagawa

Journal of the American Statistical Association

Vol. 93, No. 443 (Sep., 1998), pp. 1203-1215 (13 pages)

Published By: Taylor & Francis, Ltd.



1: 590회 (8-17-2011) : 무관

2: 559회 (9월 1998)

3: 552회 (2015)

<https://doi.org/10.2307/2669862>

<https://www.jstor.org/stable/2669862>

Abstract:

A self-organizing filter and smoother for the **general nonlinear non-Gaussian state-space model** is proposed. An **expanded state-space model** is defined by augmenting the state vector with the unknown parameters of the original state-space model. The state of the augmented state-space model, and hence the state and the parameters of the original state-space model, are estimated simultaneously by either a non-Gaussian filter/smoothing or a Monte Carlo filter/smoothing. In contrast to maximum likelihood estimation of model parameters in ordinary state-space modeling, for which the recursive filter computation has to be done many times, model parameter estimation in the proposed self-organizing filter/smoothing is achieved with only two passes of the recursive filter and smoother operations.

Examples such as automatic tuning of dispersion and the shape parameters, adaptation to changes of the amplitude of a signal in seismic data, state estimation for a nonlinear state space model with unknown parameters, and seasonal adjustment with a nonlinear model with changing variance parameters are shown to exemplify the usefulness of the proposed method.

초록:

일반적인 비선형 비가우시안 상태 공간 모델을 위한 자기 조직화 필터 및 평활기가 제안됩니다. **확장된 상태 공간 모델**은 원래 상태 공간 모델의 알려지지 않은 매개변수로 상태 벡터를 보강하여 정의됩니다.

증강된 상태 공간 모델의 상태, 따라서 원래 상태 공간 모델의 상태 및 매개변수는 비가우시안 필터/평활기 또는 몬테카를로 필터/평활기에 의해 동시에 추정됩니다.

재귀 필터 계산을 여러 번 수행해야 하는 일반적인 상태 공간 모델링에서 모델 매개변수의 최대 가능성 추정과 달리, 제안된 자기 조직화 필터/평활기에서의 모델 매개변수 추정은 재귀 필터의 두 번의 통과와 더 평활한 작업만으로 달성됩니다.

분산 및 형상 매개변수의 자동 조정, 지진 데이터에서 신호 진폭 변화에 대한 적응, 매개변수가 알려지지 않은 비선형 상태 공간 모델에 대한 상태 추정, 변화하는 분산 매개변수가 있는 비선형 모델을 사용한 계절 조정과 같은 예가 제안된 방법의 유용성을 예시하는 것으로 나타납니다.

트랜스포머 아키텍처의 대안 = SSM ②

OXFORD ACADEMIC

Subject Journals Books Oxford Academic

James Durbin, Siem Jan Koopman = Oxford Academy Oct. 2012, 5,452회 인용

Time Series Analysis by State Space Methods (2nd edn) [Get access >](#)

James Durbin, Siem Jan Koopman

Published: 3 May 2012

“ Cite Permissions Share

Search in this book

<https://doi.org/10.1093/acprof:oso/9780199641178.001.0001>

Online ISBN: 9780191774881
Print ISBN: 9780199641178
Publisher: Oxford University Press

Abstract

This book presents a comprehensive treatment of the state space approach to time series analysis. The distinguishing feature of state space time series models is that observations are regarded as being made up of distinct components such as trend, seasonal, regression elements and disturbance elements, each of which is modelled separately. The techniques that emerge from this approach are very flexible. Part I presents a full treatment of the construction and analysis of linear Gaussian state space models. The methods are based on the Kalman filter and are appropriate for a wide range of problems in practical time series analysis. The analysis can be carried out from both classical and Bayesian perspectives. Part I then presents illustrations to real series and exercises are provided for a selection of chapters. Part II discusses approximate and exact approaches for handling broad classes of non-Gaussian and nonlinear state space models. Approximate methods include the extended Kalman filter and the more recently developed unscented Kalman filter. The book shows that exact treatments become feasible when simulation-based methods such as importance sampling and particle filtering are adopted. Bayesian treatments based on simulation methods are also explored.

Keywords: unobserved components, Kalman filter and smoother, signal extraction, forecasting, maximum likelihood, extended Kalman filter, unscented Kalman filter, simulation-based methods, Monte Carlo, importance sampling

Subject: Probability and Statistics

Collection: Oxford Scholarship Online

Time Series Analysis by State Space Methods

James Durbin, Siem Jan Koopman = May 2012, 5,452회 인용

Abstract

This book presents a **comprehensive treatment of the state space approach to time series analysis**. The distinguishing feature of state space time series models is that observations are regarded as being made up of distinct components such as trend, seasonal, regression elements and disturbance elements, each of which is modelled separately. The techniques that emerge from this approach are very flexible. Part I presents a full treatment of the construction and analysis of linear Gaussian state space models. The methods are based on the Kalman filter and are appropriate for a wide range of problems in practical time series analysis. The analysis can be carried out from both classical and Bayesian perspectives. Part I then presents illustrations to real series and exercises are provided for a selection of chapters. Part II discusses approximate and exact approaches for handling broad classes of non-Gaussian and nonlinear state space models. Approximate methods include the extended Kalman filter and the more recently developed unscented Kalman filter.

The book shows that exact treatments become feasible when simulation-based methods such as **importance sampling and particle filtering** are adopted. **Bayesian treatments** based on simulation methods are also explored.

이 책은 시계열 분석에 대한 상태 공간 접근법의 포괄적인 접근법을 제시합니다. 상태 공간 시계열 모형의 특징은 관측치가 추세, 계절, 회귀 요소 및 교란 요소와 같은 서로 다른 요소로 구성된 것으로 간주되며, 이 요소들은 각각 개별적으로 모델링됩니다. 이 접근법에서 나오는 기법은 매우 유연합니다.

파트 I에서는 선형 가우스 상태 공간 모델의 구성과 분석에 대한 전체적인 처리를 제시합니다. 이 방법들은 칼만 필터를 기반으로 하며 실제 시계열 분석에서 광범위한 문제에 적합합니다. 분석은 고전적 관점과 베이지안 관점 모두에서 수행될 수 있습니다. 그런 다음 파트 I에서는 실제 시리이즈에 대한 삽화를 제시하고 일부 챕터에 대한 연습 문제를 제공합니다. 파트 II에서는 광범위한 클래스의 비가우시안 및 비선형 상태 공간 모델을 처리하기 위한 근사적이고 정확한 접근 방식에 대해 설명합니다. 대략적인 방법에는 확장 칼만 필터와 최근에 개발된 무향 칼만 필터가 포함됩니다.

이 책은 **중요도 샘플링과 입자 필터링**과 같은 시뮬레이션 기반 방법을 채택할 때 정확한 치료가 가능하다는 것을 보여줍니다. 시뮬레이션 방법을 기반으로 한 **베이지안 치료법**도 살펴봅니다.

시계열 분석의 획기적 발전을 주도한 James Durbin은 **시계열 분석에 대한 상태 공간 접근법의 포괄적인 접근법을 제시** 하였다. 또한 이 저자는 **중요도 샘플링과 입자 필터링** 접근방법과 **베이지안 치료법**을 통한 효과적인 SSM 기술에 대해 정의.

트랜스포머 아키텍처의 대안 = SSM ③

arXiv > stat > arXiv:1412.8695
Nikolas Kantas, Arnaud Doucet, Sumeetpal S. Singh, Jan Maciejowski and Nicolas Chopin = Stanford University
Dec. 2014, 552회 인용

Statistics > Computation

On Particle Methods for Parameter Estimation in State-Space Models

Nikolas Kantas, Arnaud Doucet, Sumeetpal S. Singh, Jan Maciejowski, Nicolas Chopin

Nonlinear non-Gaussian state-space models are ubiquitous in statistics, econometrics, information engineering and signal processing. Particle methods, also known as Sequential Monte Carlo (SMC) methods, provide reliable numerical approximations to the associated state inference problems. However, in most applications, the state-space model of interest also depends on unknown static parameters that need to be estimated from the data. In this context, standard particle methods fail and it is necessary to rely on more sophisticated algorithms. The aim of this paper is to present a comprehensive review of particle methods that have been proposed to perform static parameter estimation in state-space models. We discuss the advantages and limitations of these methods and illustrate their performance on simple models.

Comments: Published at this [http URL](#) in the Statistical Science (this [http URL](#)) by the Institute of Mathematical Statistics (this [http URL](#))
 Subjects: **Computation (stat.CO)**; Methodology (stat.ME)
 Report number: IMS-STS-STS511
 Cite as: arXiv:1412.8695 [stat.CO]
 (or arXiv:1412.8695v2 [stat.CO] for this version)
<https://doi.org/10.48550/arXiv.1412.8695>
 Journal reference: Statistical Science 2015, Vol. 30, No. 3, 328-351
 Related DOI: <https://doi.org/10.1214/14-STS511>

Submission history
 From: Nikolas Kantas [view email] [via VTEX proxy]
 [v1] Tue, 30 Dec 2014 17:21:00 UTC (417 KB)
 [v2] Thu, 10 Sep 2015 12:02:31 UTC (2,501 KB)

Key words and phrases: Bayesian inference, maximum likelihood inference, particle filtering, Sequential Monte Carlo, state-space models.

On particle methods for parameter estimation in state space models
 N Kantas, A Doucet, SS Singh, J Maciejowski... August 2015
 ... of inference in **statespace models**, consider On Particle Methods for Parameter Estimation in State-Space Models
 is only feasible analytically for simple **models** su
 ☆ 저장 552회 인용 552회 인용 관련 학술자료
 Statist. Sci. 30(3): 328-351 (August 2015). DOI: 10.1214/14-STS511

ABOUT FIRST PAGE CITED BY REFERENCES

On Particle Methods for Parameter Estimation in State-Space Models

Abstract
 Nonlinear non-Gaussian state-space models are ubiquitous in statistics, econometrics, information engineering and signal processing.
 Particle methods, also known as Sequential Monte Carlo (SMC) methods, provide reliable numerical approximations to the associated state inference problems.
 However, in most applications, the state-space model of interest also depends on unknown static parameters that need to be estimated from the data. In this context, **standard particle methods fail** and it is necessary to rely on more sophisticated algorithms.

The aim of this paper is to present a comprehensive review of particle methods that have been proposed to perform static parameter estimation in state-space models.
 We discuss the advantages and limitations of these methods and illustrate their performance on simple models..

비선형 비 가우시안 상태 공간 모델은 통계, 계량 경제학, 정보 공학 및 신호 처리 분야에서 널리 사용됩니다. 순차 몬테카를로(SMC) 방법으로도 알려진 입자 방법은 관련 상태 추론 문제에 대한 신뢰할 수 있는 수치 근사치를 제공합니다.

그러나 대부분의 응용 분야에서 관심 있는 상태 공간 모델은 데이터에서 추정해야 하는 알려지지 않은 정적 매개 변수에도 의존합니다. 이와 같은 맥락에서 **표준 입자 방법은 실패**하고 더 정교한 알고리즘에 의존해야 합니다.

이 논문의 목적은 상태 공간 모델에서 **정적 매개변수 추정을 수행하기** 위해 제안된 **입자 방법에 대한 포괄적인 검토를 제시**하는 것입니다.
 우리는 이러한 방법의 장점과 한계에 대해 논의하고 간단한 모델에 대한 성능을 설명합니다.

Key words: 베이지안 추론, 최대 우도 추론, 입자 필터링, 순차 몬테카를로, 상태 공간 모델.

트랜스포머 아키텍처의 대안 = SSM ④

arXiv > cs > arXiv:2111.00396 새로운 매개변수화(HiPPO)를 통한 성능개선 Search... Help

Computer Science > Machine Learning

[Submitted on 31 Oct 2021 (v1), last revised 5 Aug 2022 (this version, v3)]

Efficiently Modeling Long Sequences with Structured State Spaces

Albert Gu, Karan Goel, Christopher Ré *Albert Gu, Karan Goel, and Christopher Ré = Stanford University Oct. 2021, 865회 인용*

A central goal of sequence modeling is designing a single principled model that can address sequence data across a range of modalities and tasks, particularly on long-range dependencies. Although conventional models including RNNs, CNNs, and Transformers have specialized variants for capturing long dependencies, they still struggle to scale to very long sequences of 10000 or more steps. A promising recent approach proposed modeling sequences by simulating the fundamental state space model (SSM) $x'(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$, and showed that for appropriate choices of the state matrix A , this system could handle long-range dependencies mathematically and empirically. However, this method has prohibitive computation and memory requirements, rendering it infeasible as a general sequence modeling solution. We propose the Structured State Space sequence model (S4) based on a new parameterization for the SSM, and show that it can be computed much more efficiently than prior approaches while preserving their theoretical strengths. Our technique involves conditioning A with a low-rank correction, allowing it to be diagonalized stably and reducing the SSM to the well-studied computation of a Cauchy kernel. S4 achieves strong empirical results across a diverse range of established benchmarks, including (i) 91% accuracy on sequential CIFAR-10 with no data augmentation or auxiliary losses, on par with a larger 2-D ResNet, (ii) substantially closing the gap to Transformers on image and language modeling tasks, while performing generation $60\times$ faster (iii) SoTA on every task from the Long Range Arena benchmark, including solving the challenging Path-X task of length 16k that all prior work fails on, while being as efficient as all competitors.

Comments: ICLR 2022 (Outstanding Paper HM)
 Subjects: Machine Learning (cs.LG)
 Cite as: arXiv:2111.00396 [cs.LG]
 (or arXiv:2111.00396v3 [cs.LG] for this version)
<https://doi.org/10.48550/arXiv.2111.00396>

Submission history
 From: Albert Gu [view email]
 [v1] Sun, 31 Oct 2021 03:32:18 UTC (1,116 KB)
 [v2] Fri, 4 Mar 2022 19:43:50 UTC (2,956 KB)
 [v3] Fri, 5 Aug 2022 17:54:38 UTC (3,167 KB)

they still struggle to scale to very long sequences of 10000 or more steps. A promising recent approach proposed modeling sequences by simulating the fundamental state space model (SSM) $x'(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$, and showed that for appropriate choices of the state matrix A , this

Efficiently Modeling Long Sequences with Structured State Spaces (S4)

Abstract
 A central goal of sequence modeling is designing a single principled model that can address sequence data across a range of modalities and tasks, particularly on long-range dependencies. Although conventional models including RNNs, CNNs, and Transformers have specialized variants for capturing long dependencies, they still struggle to scale to very long sequences of 10000 or more steps. A promising recent approach proposed modeling sequences by simulating the fundamental state space model (SSM) $x'(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$, and showed that for appropriate choices of the state matrix A , this system could handle long-range dependencies mathematically and empirically. However, this method has prohibitive computation and memory requirements, rendering it infeasible as a general sequence modeling solution. We propose the Structured State Space sequence model (S4) based on a new parameterization for the SSM, and show that it can be computed much more efficiently than prior approaches while preserving their theoretical strengths. Our technique involves conditioning A with a low-rank correction, allowing it to be diagonalized stably and reducing the SSM to the well-studied computation of a Cauchy kernel.

S4 achieves strong empirical results across a diverse range of established benchmarks, including (i) 91% accuracy on sequential CIFAR-10 with no data augmentation or auxiliary losses, on par with a larger 2-D ResNet, (ii) substantially closing the gap to Transformers on image and language modeling tasks, while performing generation $60\times$ faster (iii) SoTA on every task from the Long Range Arena benchmark, including solving the challenging Path-X task of length 16k that all prior work fails on, while being as efficient as all competitors.

시퀀스 모델링의 핵심 목표는 특히 장거리 종속성에 대해 다양한 양식과 작업에 걸쳐 시퀀스 데이터를 처리할 수 있는 단일 원칙 모델을 설계하는 것입니다.
 RNN, CNN 및 트랜스포머를 포함한 기존 모델에는 긴 종속성을 캡처하는 데 특화된 변형이 있지만 여전히 10000 단계 이상의 매우 긴 시퀀스로 확장하는 데 어려움을 겪고 있습니다. <https://arxiv.org/pdf/2111.00396>
 최근 유망한 접근 방식은 기본 상태 공간 모델(SSM) $x'(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$ 를 시뮬레이션하여 모델링 시퀀스를 제안했으며, 상태 행렬 A 의 적절한 선택을 위해 이 시스템이 장거리 종속성을 수학적 및 경험적으로 처리할 수 있음을 보여주었습니다.
 그러나 이 방법은 계산 및 메모리 요구 사항이 너무 많아 일반적인 시퀀스 모델링 솔루션으로는 불가능합니다.
 우리는 SSM에 대한 새로운 매개변수화를 기반으로 한 구조화된 상태 공간 시퀀스 모델(S4)을 제안하고, 이론적 강점을 유지하면서 이전 접근 방식보다 훨씬 효율적으로 계산할 수 있음을 보여줍니다.
 우리의 기술은 낮은 순위 보정으로 A 를 조건화하여 안정적으로 대각선화할 수 있도록 하고 SSM을 잘 연구된 코시 커널 계산으로 줄이는 것을 포함합니다.

S4는 (i) 데이터 증강 또는 보조 손실 없이 순차적 CIFAR-10에서 91% 정확도를 포함하여 다양한 범위의 확립된 벤치마크에서 강력한 경험적 결과를 달성하며, (ii) 더 큰 2-D ResNet과 동등하게 이미지 및 언어 모델링 작업에서 트랜스포머와의 격차를 실질적으로 좁히는 동시에, 모든 이전 작업에서 실패하는 길이 16k의 까다로운 Path-X 작업을 해결하는 것을 포함하여 장거리 아레나 벤치마크의 모든 작업에서 60배 더 빠른 (iii) SoTA 생성을 수행하는 동시에 모든 경쟁자만큼 효율적입니다.

3차 시부후기를 주도하는 RNN, CNN 및 트랜스포머를 포함하는 기존 모델의 한계점인 10000단계 이상의 매우 긴 시퀀스를 확장하고자 함. 기존의 SSM 방식 역시, 계산 및 메모리 요구사항이 높은 것을 개선하여 새로운 매개변수화를 기반으로 한 구조화된 공간 시퀀스 모델(S4)을 제안

베이지안 치료법

트랜스포머 아키텍처의 대안 = SSM ④ - A

새로운 매개변수화(HiPPO)를 통한 성능개선

HiPPO: Recurrent Memory with Optimal Polynomial Projections
 Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, Christopher Re = Stanford University Aug. 2020, 292회 인용

arXiv > cs > arXiv:2008.07669 Search... Help

Computer Science > Machine Learning

[Submitted on 17 Aug 2020 (v1), last revised 23 Oct 2020 (this version, v2)]

HiPPO: Recurrent Memory with Optimal Polynomial Projections

Albert Gu, [Tri Dao](#), [Stefano Ermon](#), [Atri Rudra](#), [Christopher Re](#)

A central problem in learning from sequential data is representing cumulative history in an incremental fashion as more data is processed. We introduce a general framework (HiPPO) for the online compression of continuous signals and discrete time series by projection onto polynomial bases. Given a measure that specifies the importance of each time step in the past, HiPPO produces an optimal solution to a natural online function approximation problem. As special cases, our framework yields a short derivation of the recent Legendre Memory Unit (LMU) from first principles, and generalizes the ubiquitous gating mechanism of recurrent neural networks such as GRUs. This formal framework yields a new memory update mechanism (HiPPO-LegS) that scales through time to remember all history, avoiding priors on the timescale. HiPPO-LegS enjoys the theoretical benefits of timescale robustness, fast updates, and bounded gradients. By incorporating the memory dynamics into recurrent neural networks, HiPPO RNNs can empirically capture complex temporal dependencies. On the benchmark permuted MNIST dataset, HiPPO-LegS sets a new state-of-the-art accuracy of 98.3%. Finally, on a novel trajectory classification task testing robustness to out-of-distribution timescales and missing data, HiPPO-LegS outperforms RNN and neural ODE baselines by 25-40% accuracy.

Subjects: **Machine Learning (cs.LG)**, Machine Learning (stat.ML)
 Cite as: arXiv:2008.07669 [cs.LG]
 (or arXiv:2008.07669v2 [cs.LG] for this version)
<https://doi.org/10.48550/arXiv.2008.07669>

Submission history
 From: Albert Gu [[view email](#)]
[v1] Mon, 17 Aug 2020 23:39:33 UTC (988 KB)
[v2] Fri, 23 Oct 2020 02:48:03 UTC (1,028 KB)

Abstract

A central problem in learning from sequential data is representing cumulative history in an incremental fashion as more data is processed. We introduce a general framework (HiPPO) for the online compression of continuous signals and discrete time series by projection onto polynomial bases. Given a measure that specifies the importance of each time step in the past, HiPPO produces an optimal solution to a natural online function approximation problem.

As special cases, our framework yields a short derivation of the recent Legendre Memory Unit (LMU) from first principles, and generalizes the ubiquitous gating mechanism of recurrent neural networks such as GRUs. This formal framework yields a new memory update mechanism (HiPPO-LegS) that scales through time to remember all history, avoiding priors on the timescale.

HiPPO-LegS enjoys the theoretical benefits of timescale robustness, fast updates, and bounded gradients. By incorporating the memory dynamics into recurrent neural networks, HiPPO RNNs can empirically capture complex temporal dependencies. On the benchmark permuted MNIST dataset, HiPPO-LegS sets a new state-of-the-art accuracy of 98.3%.

Finally, on a novel trajectory classification task testing robustness to out-of-distribution timescales and missing data, HiPPO-LegS outperforms RNN and neural ODE baselines by 25-40% accuracy.

순차적 데이터에서 학습하는 데 있어 핵심적인 문제는 더 많은 데이터가 처리됨에 따라 누적 기록을 점진적으로 나타내는 것입니다.

[다항식 베이스에 투영하여 연속 신호와 이산 시계열의 온라인 압축을 위한 일반 프레임워크\(HiPPO\)를 소개합니다.](#)

과거 각 시간 단계의 중요도를 지정하는 척도가 주어지면 HiPPO는 자연스러운 온라인 함수 근사 문제에 대한 최적의 솔루션을 생성합니다.
 특별한 경우로, 우리의 프레임워크는, **첫 번째 원칙에서 최근의 LMU(Legendre Memory Unit)를 간략하게 도출하고 GRU와 같은 순환 신경망의 유비쿼터스 게이팅 메커니즘을 일반화합니다.**

이 공식 프레임워크는 시간을 통해 확장되어 모든 이력을 기억하는 새로운 메모리 업데이트 메커니즘(HiPPO-LegS)을 생성하여 시간 척도의 선행을 피합니다.

HiPPO-LegS는 시간 척도 견고성, 빠른 업데이트 및 제한된 그라디언트의 이론적 이점을 누리고 있습니다. 메모리 역학을 순환 신경망에 통합함으로써 HiPPO RNN은 복잡한 시간적 종속성을 경험적으로 포착할 수 있습니다. 벤치마크 순열 MNIST 데이터 세트에서 HiPPO-LegS는 98.3%의 새로운 최첨단 정확도를 설정합니다.

마지막으로, 분포 이탈 시간 척도 및 누락된 데이터에 대한 견고성을 테스트하는 새로운 궤적 분류 작업에서 HiPPO-LegS는 RNN 및 신경 ODE 기준선을 25~40% 정확도로 능가합니다.

트랜스포머 아키텍처의 대안 = SSM : Falcon

arXiv > cs > arXiv:2311.16867

Any authoris The Falcon LLM Team, Nov.. 2023, 184회 인용

Computer Science > Computation and Language

Source: <https://arxiv.org/abs/2311.16867>

[Submitted on 28 Nov 2023 (v1), last revised 29 Nov 2023 (this version, v2)]

The Falcon Series of Open Language Models

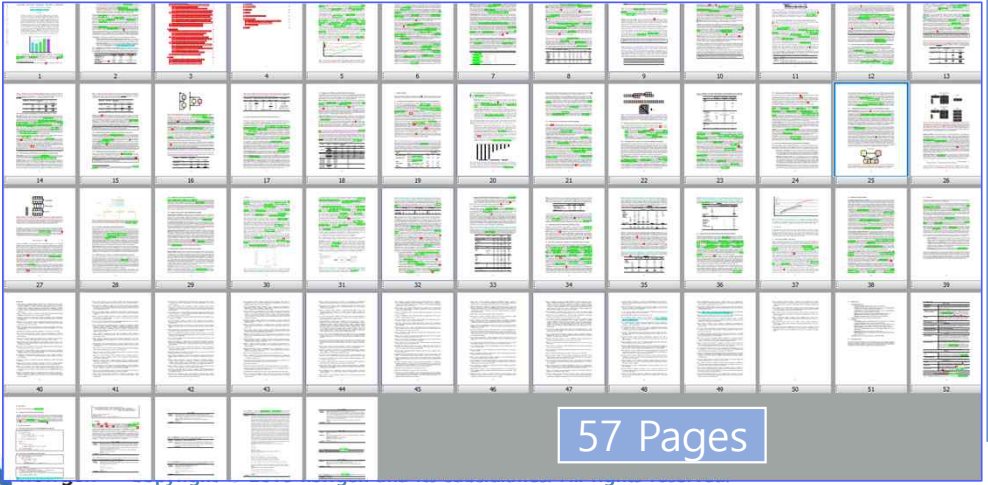
Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, M rouane Debbah,  tienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noun, Baptiste Pannier, Guilherme Penedo

We introduce the Falcon series: 7B, 40B, and 180B parameters causal decoder-only models trained on a diverse high-quality corpora predominantly assembled from web data. The largest model, Falcon-180B, has been trained on over 3.5 trillion tokens of text—the largest openly documented pretraining run. Falcon-180B significantly outperforms models such as PaLM or Chinchilla, and improves upon concurrently developed models such as LLaMA 2 or Inflection-1. It nears the performance of PaLM-2-Large at a reduced pretraining and inference cost, making it, to our knowledge, one of the three best language models in the world along with GPT-4 and PaLM-2-Large. We report detailed evaluations, as well as a deep dive into the methods and custom tooling employed to pretrain Falcon. Notably, we report on our custom distributed training codebase, allowing us to efficiently pretrain these models on up to 4,096 A100s on cloud AWS infrastructure with limited interconnect. We release a 600B tokens extract of our web dataset, as well as the Falcon-7/40/180B models under a permissive license to foster open-science and accelerate the development of an open ecosystem of large language models.

Subjects: **Computation and Language (cs.CL)**, Artificial Intelligence (cs.AI)

Cite as: arXiv:2311.16867 [cs.CL]
(or arXiv:2311.16867v2 [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.2311.16867>

Submission history
From: Julien Launay [view email]
[v1] Tue, 28 Nov 2023 15:12:47 UTC (1,452 KB)
[v2] Wed, 29 Nov 2023 19:45:10 UTC (1,453 KB)



The Falcon Series of Open Language Models

Abstract

We introduce the Falcon series: 7B, 40B, and 180B parameters causal decoder-only models trained on a diverse high-quality corpora predominantly assembled from web data.

The largest model, Falcon-180B, has been trained on over 3.5 trillion tokens of text—the largest openly documented pretraining run. Falcon-180B significantly outperforms models such as PaLM or Chinchilla, and improves upon concurrently developed models such as LLaMA 2 or Inflection-1.

It nears the performance of PaLM-2-Large at a reduced pretraining and inference cost, making it, to our knowledge, one of the three best language models in the world along with GPT-4 and PaLM-2-Large.

We report detailed evaluations, as well as a deep dive into the methods and custom tooling employed to pretrain Falcon. Notably, we report on our custom distributed training codebase, allowing us to efficiently pretrain these models on up to 4,096 A100s on cloud AWS infrastructure with limited interconnect.

We release a 600B tokens extract of our web dataset, as well as the Falcon-7/40/180B models under a permissive license to foster open-science and accelerate the development of an open ecosystem of large language models.

Falcon 시리즈를 소개합니다. 7B, 40B, 180B 매개변수 인과 디코더 전용 모델은 주로 웹 데이터에서 조립된 다양한 고품질 코퍼스에서 학습되었습니다.

가장 큰 모델인 Falcon-180B는 3.5조 개가 넘는 텍스트 토큰에서 학습되었습니다. 이는 가장 큰 공개 문서화된 사전 학습 실행입니다. Falcon-180B는 PaLM 또는 Chinchilla와 같은 모델보다 상당히 성능이 뛰어나며 LLaMA 2 또는 Inflection-1과 같은 동시에 개발된 모델보다 개선되었습니다.

사전 학습 및 추론 비용이 감소한 PaLM-2-Large의 성능에 근접하여, 저희가 아는 한 GPT-4 및 PaLM-2-Large와 함께 세계 최고의 언어 모델 3개 중 하나입니다.

자세한 평가와 Falcon을 사전 학습하는 데 사용된 방법 및 사용자 지정 도구에 대한 심층 분석을 보고합니다. 특히, 우리는 맞춤형 분산 학습 코드베이스에 대해 보고하며, 이를 통해 제한된 상호 연결을 통해 클라우드 AWS 인프라에서 최대 4,096개의 A100에서 이러한 모델을 효율적으로 사전 학습할 수 있습니다.

우리는 개방형 과학을 육성하고 대규모 언어 모델의 개방형 생태계 개발을 가속화하기 위해 허용 라이선스에 따라 웹 데이터 세트의 600B 토큰 추출물과 Falcon-7/40/180B 모델을 릴리스합니다.

트랜스포머 아키텍처의 대안 = SSM : Falcon-1

- 아부다비 정부의 첨단기술연구위원회(ATRC) 산하 기술혁신연구소(TII)가 'Falcon Mamba 7B'를 오픈 소스로 출시
- Falcon Mamba 7B
 - 1) 처리 가능한 최대 컨텍스트 길이를 확인하는 테스트에서 SOTA 트랜스포머 기반 모델보다 더 큰 시퀀스를 처리 할 수 있다고 함
 - 2) 처리량 테스트에서도 모든 토큰을 일정한 속도로 생성
 - 3) 미스트랄 7B'의 슬라이딩 윈도우 어텐션 아키텍처를 능가한 것
 - 4) 일반 벤치마크에서는 트랜스포머 모델은 물론, 기존 SSM보다 뛰어나거나 비슷한 성능을 기록
 - 5) 추론(Arc), 환각방지(TruthfulQA), 수학(GSM8K) 벤치마크에서 각각 62.03%, 53.42%, 52.54%를 기록
 - 6) '라마 3 8B'나 '라마 3.1 8B' '젬마 7B' '미스트랄 7B' 등 동급 모델을 확실하게 능가
 - 7) 반면, 언어 이해(MMLU)와 상식(Hellaswag) 벤치마크에서는 나머지 모델보다 뒤처졌다.
- 이전 버전 Falcon 40B (2023년 6월 출시)
 - 1) 허깅페이스 리더보드 1위에 오르는 등 큰 인기를 누리며, 4500만번 이상 다운로드 된 모델
 - 2) 단번에 UAE를 중동 LLM의 핵심으로 만들어준 모델

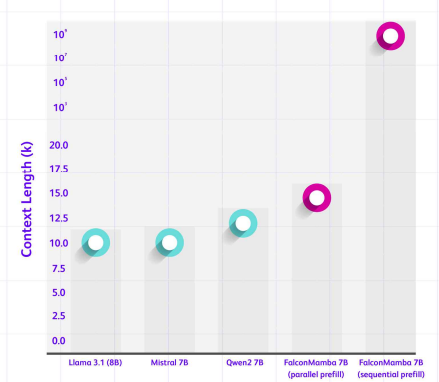
Performance Benchmarks

Architecture	Model	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU -PRO	Avg - 6 tasks
Mamba (SSLM)	Falcon Mamba 7B	33.36	19.88	3.63	8.05	10.86	14.47	15.04
Transformer	Mistral (7B)	22.66	24.04	2.64	5.59	8.36	21.7	14.17
	Llama 3.1 (8B)	12.7	25.29	4.61	6.15	8.98	24.95	13.78
	Falcon 2 (11B)	32.61	21.94	2.34	2.8	7.53	15.44	13.78

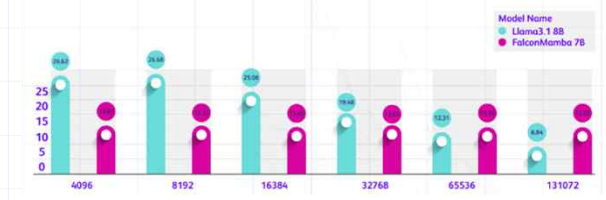
Falcon Mamba 7B: Highlights

- Falcon Mamba 7B is the first open source released State Space Language Model (SSLM), a new revolutionary architecture for Falcon models
- Falcon Mamba 7B is the no. 1 globally performing open source SSLM in the world, as independently verified by Hugging Face
- SSLMs have a low memory cost and don't require additional memory to generate arbitrary long blocks of text
- Falcon Mamba 7B also outperforms traditional transformer architecture models such as Meta's Llama 3.1 8B and Mistral's 7B
- New model reflects the innovation and pioneering approach of Abu Dhabi in AI research and development

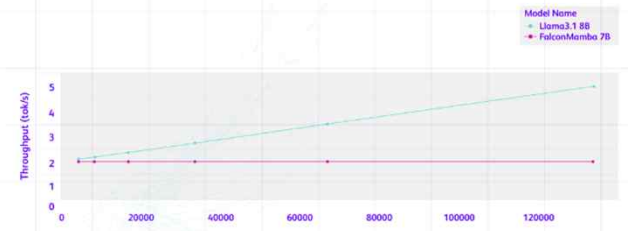
Maximum context length fitting on a single 24GB A10 GPU



Throughput vs Generated tokens (high is better) - using Hugging Face transformers library



Maximum GPU memory occupied by tensors vs Generated tokens (lower is better) - using Hugging Face transformers library



트랜스포머 아키텍처의 대안 = SSM : Mamba

arXiv > cs > arXiv:2312.00752v2

Computer Science > Machine Learning

[Submitted on 1 Dec 2023 (v1), last revised 31 May 2024 (this version, v2)]

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu, Tri Dao

Albert Gu*, Tri Dao* = Carnegie Mellon/ Princeton University, Dec., 2023, 694회 인용

Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language. We identify that a key weakness of such models is their inability to perform content-based reasoning, and make several improvements. First, simply letting the SSM parameters be functions of the input addresses their weakness with discrete modalities, allowing the model to selectively propagate or forget information along the sequence length depending on the current token. Second, even though this change prevents the use of efficient convolutions, we design a hardware-aware parallel algorithm in recurrent mode. We integrate these selective SSMs into a simplified end-to-end neural network architecture without attention or even MLP blocks (Mamba). Mamba enjoys fast inference ($5\times$ higher throughput than Transformers) and linear scaling in sequence length, and its performance improves on real data up to million-length sequences. As a general sequence model backbone, Mamba achieves state-of-the-art performance across several modalities such as language, audio, and genomics. On language modeling, our Mamba-3B model outperforms Transformers of the same size and matches Transformers twice its size, both in pretraining and downstream evaluation.

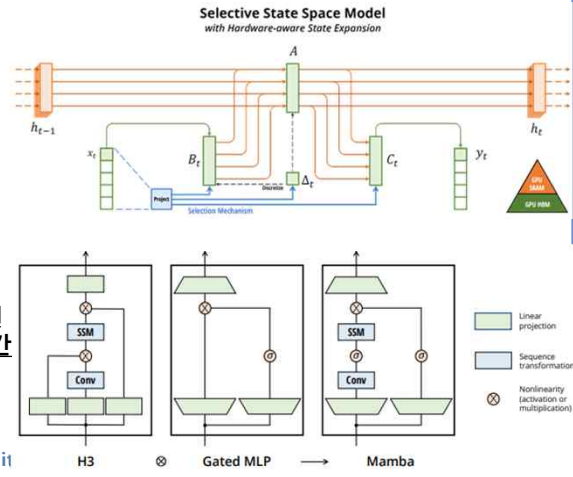
Subjects: Machine Learning (cs.LG), Artificial Intelligence (cs.AI)
 Cite as: arXiv:2312.00752 [cs.LG]
 (or arXiv:2312.00752v2 [cs.LG] for this version)
<https://doi.org/10.48550/arXiv.2312.00752>

Submission history

From: Albert Gu [view email]
 [v1] Fri, 1 Dec 2023 18:01:34 UTC (1,264 KB)
 [v2] Fri, 31 May 2024 17:55:27 UTC (1,017 KB)

- Mamba: S4의 개념을 바탕으로 개발
- 긴 시퀀스를 효율적으로 처리하면서도
 - 고성능을 유지할 수 있도록 설계

Mamba는 특히 메모리 사용과 계산 자원 효율성을 극대화하기 위해 **S4의 상태 공간 모델링 기술을 적극 활용**합니다.



Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu*, Tri Dao* = Carnegie Mellon/ Princeton University

Abstract

Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs=S4) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language. We identify that a key weakness of such models is their inability to perform content-based reasoning, and make several improvements. First, simply letting the SSM parameters be functions of the input addresses their weakness with discrete modalities, allowing the model to selectively propagate or forget information along the sequence length dimension depending on the current token. Second, even though this change prevents the use of efficient convolutions, we design a hardware-aware parallel algorithm in recurrent mode. We integrate these selective SSMs into a simplified end-to-end neural network architecture without attention or even MLP blocks (Mamba). Mamba enjoys fast inference ($5\times$ higher throughput than Transformers) and linear scaling in sequence length, and its performance improves on real data up to million-length sequences. As a general sequence model backbone, Mamba achieves state-of-the-art performance across several modalities such as language, audio, and genomics. On language modeling, our Mamba-3B model outperforms Transformers of the same size and matches Transformers twice its size, both in pretraining and downstream evaluation.

현재 딥 러닝 분야의 **대부분의 흥미로운 애플리케이션을 지원하는 기반 모델은 거의 보편적으로 트랜스포머 아키텍처와 핵심 주의 모듈을 기반으로 합니다.**

선형 주의, 게이트 컨볼루션 및 순환 모델, 구조화된 상태 공간 모델(SSMs=S4)과 같은 많은 2차 시간 아키텍처가 긴 시퀀스에서 트랜스포머의 계산 비효율을 해결하기 위해 개발되었지만 **언어와 같은 중요한 양식에 대한 주의만큼 잘 수행되지 않았습니다. (보다 완전한 해결책이 되지 못함)**

우리는 이러한 모델의 핵심 약점은 콘텐츠 기반 추론을 수행할 수 없다는 것을 확인하고 몇 가지 개선 사항을 제공합니다.

첫째, SSM 매개변수를 입력의 함수로 두는 것만으로도 개별 양식으로 약점을 해결하여 모델이 현재 토큰에 따라 시퀀스 길이 차원을 따라 선택적으로 정보를 전파하거나 잊어버릴 수 있습니다.

둘째, 이러한 변경으로 인해 효율적인 컨볼루션을 사용할 수 없음에도 불구하고 반복 모드에서 하드웨어 인식 병렬 알고리즘을 설계합니다.

우리는 이러한 선택적 SSM을 주의를 기울이지 않고 심지어 MLP 블록(Mamba) 없이 단순화된 엔드투엔드 신경망 아키텍처에 통합합니다.

Mamba는 빠른 추론(트랜스포머보다 5배 높은 처리량)과 시퀀스 길이의 선형 확장을 즐기며, 최대 백만 길이의 시퀀스까지 실제 데이터에서 성능이 향상됩니다. 일반적인 시퀀스 모델 백본인 Mamba는 언어, 오디오 및 유전체학과 같은 여러 양식에 걸쳐 최첨단 성능을 달성합니다. 언어 모델링에서 당사의 Mamba-3B 모델은 동일한 크기의 트랜스포머를 능가하며 사전 훈련 및 다운스트림 평가 모두에서 트랜스포머보다 두 배 큰 크기와 일치합니다.

트랜스포머 아키텍처의 대안 = SSM : Jamba

arXiv > cs > arXiv:2403.19887

22명의 공동저자로 구성 AI21 Labs, March 2024, 50회 인용

Computer Science > Computation and Language

[Submitted on 28 Mar 2024 (v1), last revised 3 Jul 2024 (this version, v2)]

Jamba: A Hybrid Transformer-Mamba Language Model

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirum, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avshalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, Yoav Shoham

We present Jamba, a new base large language model based on a novel hybrid Transformer-Mamba mixture-of-experts (MoE) architecture. Specifically, Jamba interleaves blocks of Transformer and Mamba layers, enjoying the benefits of both model families. MoE is added in some of these layers to increase model capacity while keeping active parameter usage manageable. This flexible architecture allows resource- and objective-specific configurations. In the particular configuration we have implemented, we end up with a powerful model that fits in a single 80GB GPU. Built at large scale, Jamba provides high throughput and small memory footprint compared to vanilla Transformers, and at the same time state-of-the-art performance on standard language model benchmarks and long-context evaluations. Remarkably, the model presents strong results for up to 256K tokens context length. We study various architectural decisions, such as how to combine Transformer and Mamba layers, and how to mix experts, and show that some of them are crucial in large scale modeling. We also describe several interesting properties of these architectures which the training and evaluation of Jamba have revealed, and plan to release checkpoints from various ablation runs, to encourage further exploration of this novel architecture. We make the weights of our implementation of Jamba publicly available under a permissive license.

Comments: Webpage: [this https URL](https://arxiv.org/abs/2403.19887)

Subjects: **Computation and Language (cs.CL)**, Machine Learning (cs.LG)

Cite as: arXiv:2403.19887 [cs.CL]
(or arXiv:2403.19887v2 [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.2403.19887>

Submission history

From: Yonatan Belinkov [[view email](mailto:yonatan@ai21.com)]

[v1] Thu, 28 Mar 2024 23:55:06 UTC (941 KB)

[v2] Wed, 3 Jul 2024 14:30:33 UTC (1,121 KB)

➤ 이스라엘 스타트업 AI21 랩스:
SSM을 기반으로 하는 '맘바(Mamba)'와 트랜스포머 아키텍처의 최고의 특성을 결합한 LLM '잠바(Jamba)'를 출시

Jamba: A Hybrid Transformer-Mamba Language Model

Abstract
 We present Jamba, a new base large language model based on a novel hybrid Transformer-Mamba mixture-of-experts (MoE) architecture. Specifically, Jamba interleaves blocks of Transformer and Mamba layers, enjoying the benefits of both model families. MoE is added in some of these layers to increase model capacity while keeping active parameter usage manageable. This flexible architecture allows resource- and objective-specific configurations. In the particular configuration we have implemented, we end up with a powerful model that fits in a single 80GB GPU. Built at large scale, Jamba provides high throughput and small memory footprint compared to vanilla Transformers, and at the same time state-of-the-art performance on standard language model benchmarks and long-context evaluations. Remarkably, the model presents strong results for up to 256K tokens context length. We study various architectural decisions, such as how to combine Transformer and Mamba layers, and how to mix experts, and show that some of them are crucial in large scale modeling.

We also describe several interesting properties of these architectures which the training and evaluation of Jamba have revealed, and plan to release checkpoints from various ablation runs, to encourage further exploration of this novel architecture. We make the weights of our implementation of Jamba publicly available under a permissive license.

우리는 새로운 하이브리드 트랜스포머-맘바 전문가 혼합(MoE) 아키텍처를 기반으로 하는 새로운 기본 대형 언어 모델인 잠바를 소개합니다. 특히 잠바는 트랜스포머와 맘바 레이어 블록을 인터리빙하여 두 모델 패밀리의 이점을 모두 누릴 수 있습니다.

MoE는 활성 파라미터 사용을 관리 가능하게 유지하면서 모델 용량을 늘리기 위해 이러한 계층 중 일부에 추가됩니다. 이 유연한 아키텍처를 통해 리소스 및 목표별 구성이 가능합니다. 우리가 구현한 특정 구성에서 우리는 단일 80GB GPU에 맞는 강력한 모델로 마무리됩니다. 대규모로 구축된 잠바는 바닐라 트랜스포머에 비해 높은 처리량과 작은 메모리 공간을 제공하며, 동시에 표준 언어 모델 벤치마크와 긴 컨텍스트 평가에서 최첨단 성능을 제공합니다.

놀랍게도, 이 모델은 최대 256,000개의 토큰 컨텍스트 길이에 대해 강력한 결과를 제시합니다. 저희는 트랜스포머와 맘바 레이어를 결합하는 방법, 전문가를 혼합하는 방법과 같은 다양한 아키텍처 결정을 연구하고 그 중 일부가 대규모 모델링에서 중요하다는 것을 보여줍니다.

우리는 또한 잠바의 훈련과 평가가 밝힌 이러한 아키텍처의 몇 가지 흥미로운 특성을 설명하고, 이 새로운 아키텍처에 대한 추가 탐색을 장려하기 위해 다양한 절제 실행에서 체크포인트를 공개할 계획입니다. 저희는 Jamba 구현의 가중치를 허용 라이선스에 따라 공개적으로 제공합니다.

트랜스포머 아키텍처의 대안 = SSM : Transformer are SSMs (Mamba-2)

arXiv > cs > arXiv:2405.21060 *Tri Dao, Albert Gu, Princeton / Carnegie Mellon University, Princeton / Carnegie Mellon University, May 2024, 32회 인용*

Computer Science > Machine Learning

[Submitted on 31 May 2024]

Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality

Tri Dao, Albert Gu

While Transformers have been the main architecture behind deep learning's success in language modeling, state-space models (SSMs) such as Mamba have recently been shown to match or outperform Transformers at small to medium scale. We show that these families of models are actually quite closely related, and develop a rich framework of theoretical connections between SSMs and variants of attention, connected through various decompositions of a well-studied class of structured semiseparable matrices. Our state space duality (SSD) framework allows us to design a new architecture (Mamba-2) whose core layer is a refinement of Mamba's selective SSM that is 2-8X faster, while continuing to be competitive with Transformers on language modeling.

Comments: ICML 2024
 Subjects: Machine Learning (cs.LG)
 Cite as: arXiv:2405.21060 [cs.LG]
 (or arXiv:2405.21060v1 [cs.LG] for this version)
<https://doi.org/10.48550/arXiv.2405.21060>

Submission history
 From: Albert Gu [view email]
 [v1] Fri, 31 May 2024 17:50:01 UTC (1,815 KB)

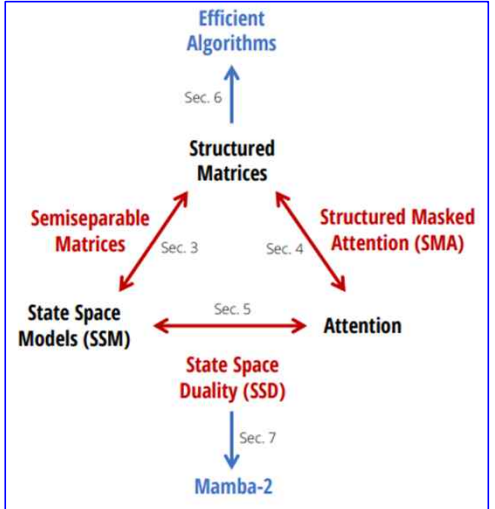


Figure 1: (Structured State-Space Duality.) This paper fleshes out the relationship between state space models and attention through the bridge of structured matrices.

Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality (Mamba-2)

Abstract

While Transformers have been the main architecture behind deep learning's success in language modeling, state-space models (SSMs) such as **Mamba** have recently been shown to match or outperform Transformers at small to medium scale.

We show that these families of models are actually quite closely related, and develop a rich framework of theoretical connections between SSMs and variants of attention, connected through various decompositions of a well-studied class of structured semiseparable matrices.

Our state space duality (SSD) framework allows us to design a new architecture (Mamba-2) whose core layer is a refinement of Mamba's selective SSM that is 2-8X faster, while continuing to be competitive with Transformers on language modeling.

트랜스포머는 딥러닝의 언어 모델링 성공 배경에 있는 주요 아키텍처였지만, 맘바(Mamba)와 같은 상태 공간 모델 (SSM)은 최근 **중소 규모에서 트랜스포머와 일치하거나 능가하는 것**으로 나타났습니다.

우리는 이러한 모델 제품군이 실제로 매우 밀접하게 관련되어 있음을 보여주고, 잘 연구된 구조화된 반분리 행렬 클래스의 다양한 분해를 통해 연결된 SSM과 관심 변형 간의 풍부한 이론적 연결 프레임워크를 개발합니다.

SSD(State Space Duality) 프레임워크를 통해 언어 모델링에서 트랜스포머와 경쟁력을 유지하면서도 2-8배 빠른 맘바의 선택적 SSM을 개선한 코어 계층의 새로운 아키텍처(Mamba-2)를 설계할 수 있습니다.

2.1 구조화된 상태 공간 모델

구조화된 상태 공간 시퀀스 모델(S4)은 RNN, CNN 및 고전적인 상태 공간 모델과 광범위하게 관련된 딥 러닝을 위한 최근 클래스의 시퀀스 모델입니다.

이들은 암시적 잠재 상태 $h \in \mathbb{R}(T, N)$ 을 통해 1차원 시퀀스 $x \in \mathbb{R}^T \rightarrow y \in \mathbb{R}^T$ 를 매핑하는 특정 연속 시스템(1)에서 영감을 받았습니다.

구조화된 SSM의 일반적인 이산 형태는 식 (1)의 형태를 취합니다

구조화된 SSM은 심층 신경망에서 사용될 수 있을 만큼 충분히 효율적으로 이 시퀀스-투-시퀀스 변환을 계산하기 위해 시간 역학을 제어하는 A 행렬을 구조화해야 하기 때문에 그렇게 명명되었습니다. 도입된 원래 구조는 대각선 플러스 로우 랭크(DPLR)(Gu, Goel 및 Ré 2022)와 대각선(Gu, Gupta, et al. 2022; Gupta, Gupta 및 Berant 2022; J.T. Smith, Warrington 및 Linderman 2023)으로 가장 인기 있는 구조로 남아 있습니다.

이 작업에서는 구조화된 SSM을 나타내는 상태 공간 모델(SSM)이라는 용어를 사용합니다. 이러한 SSM에는 연속 시간, 반복 및 컨볼루션 모델과 같은 신경 시퀀스 모델의 여러 주요 패러다임과 깊은 관련이 있는 많은 맛이 있습니다(Gu, Johnson, Goel 등 2021).

아래에 간략한 개요를 제공하며, 보다 자세한 맥락과 내용은 선행 작업을 참고합니다(Gu 2023; Gu and Dao 2023).

트랜스포머 아키텍처의 대안 = SSM : Transformer are SSMs (Mamba-2)

State Space Duality와 Transformer의 융합 (ChatGPT 4o)

- 기술적 발전:** 최근 연구에서는 Transformer와 SSM이 본질적으로 연결되어 있으며, 둘 사이의 이론적 연계를 통해 새로운 아키텍처를 설계할 수 있음을 제안합니다. 이 접근법은 Transformer의 강력한 기능과 SSM의 효율성을 결합하여, 고성능 모델을 개발하는 데 기여합니다.
- 변화된 연구 방법:** 이론적 분석을 통해 Transformer의 self-attention 메커니즘을 SSM의 구조적 이점으로 대체하거나 보완하는 연구가 진행되었으며, 이는 높은 성능의 LLM을 만드는 데 중요한 역할을 했습니다.
- 관련 논문:** "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality" (2024) [Link](#)

트랜스포머 아키텍처의 대안 = SSM : Samba

<https://github.com/microsoft/Samba>

arXiv > cs > arXiv:2406.07522 **Liliang Ren^{1,2*} Yang Liu^{1†} Yadong Lu^{1†} Yelong Shen¹ Chen Liang¹ Weizhu Chen¹ / Microsoft², University of Illinois, May 2024, 36회 인용**

Computer Science > Computation and Language [Submitted on 11 Jun 2024]
Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, Weizhu Chen

Efficiently modeling sequences with infinite context length has been a long-standing problem. Past works suffer from either the quadratic computation complexity or the limited extrapolation ability on length generalization. In this work, we present Samba, a simple hybrid architecture that layer-wise combines Mamba, a selective State Space Model (SSM), with Sliding Window Attention (SWA). Samba selectively compresses a given sequence into recurrent hidden states while still maintaining the ability to precisely recall memories with the attention mechanism. We scale Samba up to 3.8B parameters with 3.2T training tokens and show that Samba substantially outperforms the state-of-the-art models based on pure attention or SSMs on a wide range of benchmarks. When trained on 4K length sequences, Samba can be efficiently extrapolated to 256K context length with perfect memory recall and show improved token predictions up to 1M context length. As a linear-time sequence model, Samba enjoys a 3.73x higher throughput compared to Transformers with grouped-query attention when processing user prompts of 128K length, and 3.64x speedup when generating 64K tokens with unlimited streaming. A sample implementation of Samba is publicly available in [this https URL](https://github.com/microsoft/Samba).

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)
 Cite as: arXiv:2406.07522 [cs.CL]
 (or arXiv:2406.07522v1 [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.2406.07522>

Submission history
 From: Liliang Ren [view email]
 [v1] Tue, 11 Jun 2024 17:50:51 UTC (2,175 KB)

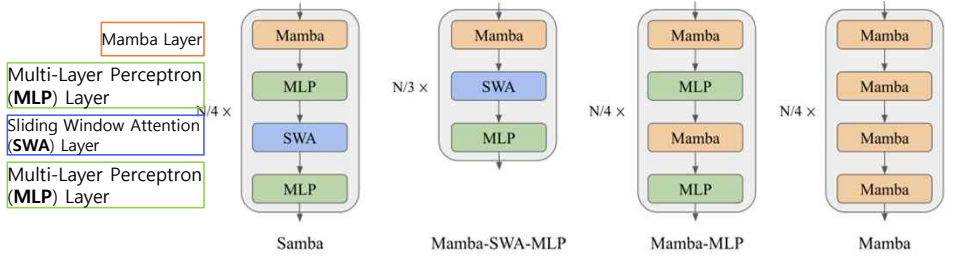


Figure 2: From left to right: Samba, Mamba-SWA-MLP, Mamba-MLP, and Mamba. The illustrations depict the layer-wise integration of Mamba with various configurations of Multi-Layer Perceptrons (MLPs) and Sliding Window Attention (SWA). We assume the total number of intermediate layers to be N , and omit the embedding layers and output projections for simplicity. Pre-Norm [XYH⁺20, ZS19] and skip connections [HZRS16] are applied for each of the intermediate layers.

Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling

Abstract

Efficiently modeling sequences with infinite context length has been a long-standing problem. Past works suffer from either the quadratic computation complexity or the limited extrapolation ability on length generalization. In this work, we present **Samba**, a simple hybrid architecture that layer-wise combines Mamba, a selective State Space Model (SSM), with Sliding Window Attention (SWA). Samba selectively compresses a given sequence into recurrent hidden states while still maintaining the ability to precisely recall memories with the attention mechanism. We scale Samba up to 3.8B parameters with 3.2T training tokens and show that Samba substantially outperforms the state-of-the-art models based on pure attention or SSMs on a wide range of benchmarks. When trained on 4K length sequences, Samba can be efficiently extrapolated to 256K context length with perfect memory recall and show improved token predictions up to 1M context length. As a linear-time sequence model, Samba enjoys a 3.73x higher throughput compared to Transformers with grouped-query attention when processing user prompts of 128K length, and 3.64x speedup when generating 64K tokens with unlimited streaming. A sample implementation of Samba is publicly available in [this https URL](https://github.com/microsoft/Samba).

컨텍스트 길이가 무한한 시퀀스를 효율적으로 모델링하는 것은 오랜 문제였습니다. 과거 작업은 2차 계산 복잡성이나 길이 일반화에 대한 제한된 외삽 능력으로 어려움을 겪었습니다.

이 작업에서는 선택적 상태 공간 모델(SSM)인 맘바와 슬라이딩 윈도우 어텐션(SWA)을 레이어별로 결합한 간단한 하이브리드 아키텍처인 삼바를 제시합니다.

삼바는 주의 메커니즘으로 기억을 정확하게 기억하는 기능을 유지하면서 주어진 시퀀스를 반복적인 숨겨진 상태로 선택적으로 압축합니다. 우리는 3.2T 훈련 토큰으로 삼바를 최대 3.8B 매개변수로 확장하고, 광범위한 벤치마크에서 순수 주의 또는 SSM을 기반으로 삼바가 최첨단 모델을 크게 능가한다는 것을 보여줍니다. 4K 길이 시퀀스로 훈련하면 완벽한 메모리 리콜로 Samba를 256K 컨텍스트 길이로 효율적으로 외삽할 수 있으며 최대 1M 컨텍스트 길이까지 향상된 토큰 예측을 보여줍니다. 선형 시간 시퀀스 모델인 Samba는 128K 길이의 사용자 프롬프트를 처리할 때 그룹화된 쿼리 주의로 트랜스포머에 비해 3.73배 높은 처리량을 제공하며, 무제한 스트리밍으로 64K 토큰을 생성할 때는 3.64배의 속도를 제공합니다. 이 [https URL](https://github.com/microsoft/Samba)에서는 Samba의 구현 예를 공개적으로 확인할 수 있습니다.

<https://github.com/microsoft/Samba>