

▼ Mamba

Mamba, however, is one of an alternative class of models called **State Space Models (SSMs)**. Importantly, for the first time, Mamba promises similar performance (and crucially similar scaling laws) as the Transformer whilst being feasible at long sequence lengths (say 1 million tokens). To achieve this long context, the Mamba authors remove the “quadratic bottleneck” in the Attention Mechanism. Mamba also runs *fast* - like “up to 5x faster than Transformer fast


SSM → RNN + GNN

Structured State Spaces: Combining Continuous-Time, Recurrent, and Convolutional Models · Hazy Research (stanford.edu).


Introduction to State Space Models (SSM)


A Blog post by Loïc BOURDOIS on Hugging Face

 <https://huggingface.co/blog/lbourdois/get-on-the-ssm-train>

 Community Article

Introduction to State Space Models (SSM)

by  lbourdois

 맘바 및 상태 공간 모델에 대한 비주얼 가이드
2312.00752.pdf (arxiv.org).



State Space Model

→ δ , A , B , C 4개의 parameter를 2 stages에 걸쳐 seq2seq Transformation 정의하는 모델



Recursive 관점 → unbounded context에 대해 사용 가능

상태 공간이란 무엇인가?

상태 공간은 시스템을 완전히 설명하는 최소한의 변수들을 포함합니다. 이는 시스템의 가능한 상태들을 정의함으로써 문제를 수학적으로 표현하는 방법입니다.

이를 좀 더 단순화해 보겠습니다. 우리가 미로를 탐색한다고 상상해 봅시다. "상태 공간"은 모든 가능한 위치(상태)의 지도입니다. 각 지점은 미로에서의 고유한 위치를 나타

내며, 출구까지 얼마나 떨어져 있는지와 같은 구체적인 세부 사항을 포함합니다.

"상태 공간 표현"은 이 지도의 간단한 설명입니다. 여기에는 현재 위치(현재 상태), 다음에 갈 수 있는 곳(가능한 미래 상태), 다음 상태로 이동하는 변화(오른쪽이나 왼쪽으로 가는 것)가 표시됩니다.





- State Representation을 업데이트 하기위한 matrix A,B,C
- delta는 skip connection마냥 행동



- A is the transition state matrix. It shows how you transition the current state into the next state. It asks "How should I forget the less relevant parts of the state over time?"
- B is mapping the new input into the state, asking "What part of my new input should I remember?" → Transformer의 Query matrix 마냥 역할을 한다.
- C is mapping the state to the output of the SSM. It asks, "How can I use the state to make a good next prediction?" → Transformer의 Output matrix 마냥 역할
- D is how the new input passes through to the output. It's a kind of modified skip connection that asks "How can I use the new input in my prediction?"



→ Recurrence의 관점

우리의 이산화된 SSM은 연속 신호 대신 특정 시간 단계에서 문제를 공식화할 수 있게 해줍니다. 우리가 이전에 RNNs와 함께 본 것처럼, 여기에서 재귀적 접근 방식이 매우 유용합니다.

우리가 연속 신호 대신 이산 시간 단계를 고려한다면, 우리는 문제를 시간 단계와 함께 재공식화할 수 있습니다:



각 시간 단계에서, 우리는 현재 입력(\mathbf{Bx}_k)이 이전 상태(\mathbf{Ah}_{k-1})에 어떻게 영향을 미치는지를 계산한 다음 예측된 출력(\mathbf{Ch}_k)을 계산합니다.

S4 → S6



Copying v Selective Copying





Mamba → Selective State Space Models Contribution

- linear projection
- selective SSM은 input마다 B,C matrix가 달라지기에 conv 적용이 불가능해짐.
- Kernel Fusion →
- Parallel scan
- Recomputation





Mamba는 언어처리, 유전체학, 오디오분석 같은 복잡한 시퀀스를 더 가볍고 빠르게 처리할 수 있는 모델

최신 하드웨어의 요구사항에 맞춰 메모리 사용과 병렬처리 기능을 모두 최적화함,
end-to-end NN Architecture

Mamba는 attention 없이 트랜스포머보다 5배 빠른 추론과 선형 스케일링

Mamba가 일반 시퀀스 모델 백본이 될 수 있는 강력한 후보

오픈소스

Mamba는 아직 encoder-decoder 개념이 있냐

Selective이므로 추론에 특화될 수 밖에 없다.