Mamba-360: Survey of State Space Models as Transformer Alternative for Long Sequence Modelling: Methods, Applications, and Challenges

نکته : تمامی عناوین و کلمات قرمز رنگ به صورت لینک های قابل کلیک هستند .

چکیده

مدل‌سازی توالی‌ها یکی از حوزه‌های کلیدی در زمینه‌های گوناگون از جمله Natural Language Processing (NLP)، تشخیص گفتار، پیش‌بینی سری‌های زمانی، تولید موسیقی و bioinformatics است. شبکه‌های عصبی بازگشتی (Recurrent Neural Net-works - RNNs) و شبکه‌های حافظه کوتاه‌مدت طولانی (Long Short-Term Memory Networks - LSTMs) در گذشته، بر وظایف مدل‌سازی توالی مانند ترجمه ماشینی و تشخیص موجودیت‌های نام‌دار (Named Entity Recognition - NER) تسلط داشتند. با این حال، پیشرفت مدل‌های transformer باعث تغییر در این پارادایم شد؛ زیرا این مدل‌ها عملکرد برتری را ارائه می‌دهند. با وجود این، ترنسفورمرها از پیچیدگی توجهی با مرتبه O^{N^2} و چالش‌هایی در برخورد با سوگیری استقرایی رنج می‌برند. برای رفع این مشکلات، گونه‌های مختلفی پیشنهاد شده‌اند که از شبکه‌های طیفی یا هم‌نهشتی (convolutional) استفاده می‌کنند و در طیف وسیعی از وظایف عملکرد مناسبی داشته‌اند، اما هنوز در پردازش توالی‌های بسیار بلند دچار مشکل هستند. در این میان، مدل‌های فضای حالت (State Space Models - SSMs) به عنوان جایگزین‌های امیدوارکننده‌ای برای پارادایم مدل‌سازی توالی مطرح شده‌اند، به‌ویژه با ظهور مدل‌هایی مانند S4 و نسخه‌های آن نظیر S4nd، Hippo، Hyena، Diagonal State Spaces (DSS)، Gated State Spaces (GSS)، Linear Recur-rent Unit (LRU)، Liquid-S4 و Mamba. در این مرور، مدل‌های بنیادی SSM بر اساس سه پارادایم اصلی دسته‌بندی می‌شوند: معماری‌های دروازه‌ای (Gating Architectures)، معماری‌های ساختاری (Structural Architectures) و معماری‌های بازگشتی (Recurrent Architectures). همچنین، این مقاله به بررسی کاربردهای متنوع SSM در حوزه‌هایی همچون بینایی رایانه‌ای، ویدئو، صوت، گفتار، زبان (به‌ویژه مدل‌سازی توالی‌های بلند)، پزشکی (از جمله ژنومیکس)، شیمی (مانند طراحی دارو)، سیستم‌های توصیه‌گر و تحلیل سری‌های زمانی از جمله داده‌های جدولی می‌پردازد. افزون بر این، عملکرد SSMها در مجموعه‌داده‌های معیار نظیر Long Range Arena (LRA)، WikiText، GLUE، Pile، ImageNet، Kinetics-400، SSTv2 و همچنین مجموعه‌داده‌های ویدئویی مانند Breakfast، COIN، LVU و چندین مجموعه‌داده سری زمانی، گردآوری شده است. صفحه‌ی پروژه‌ی مربوط به کار Mamba-360 در این وب‌سایت در دسترس است:

https://github.com/badripatro/mamba360

۱ مقدمه

شبکه‌های عصبی بازگشتی (Recurrent Neural Networks - RNNs) مدت‌هاست که سنگ‌بنای مدل‌سازی توالی به شمار می‌روند و در وظایفی مانند ترجمه ماشینی و پیش‌بینی واژه‌ی بعدی عملکرد برجسته‌ای داشته‌اند. RNNها با استفاده از ورودی کنونی و حالت قبلی، حالت بعدی را پیش‌بینی می‌کنند. با این حال، این شبکه‌ها تنها قادرند از آخرین حالت و ورودی فعلی برای پیش‌بینی استفاده کنند، که این امر موجب محدودیت در خروجی آن‌ها می‌شود. شبکه‌های عصبی بازگشتی می‌توانند توالی‌هایی با طول (L) را بدون نیاز به منابع حافظه‌ای بیش از $O(1)$ به‌صورت کارآمد پردازش کنند. با این وجود، به دلیل اینکه محاسبات گرادیان تنها بر حالت پنهان و ورودی فعلی متمرکز است، RNNها ممکن است پنجره‌ی بازبینی یا ظرفیت حافظه‌ی کافی برای یادگیری وابستگی‌های بلندمدت نداشته باشند. افزون بر این، این شبکه‌ها با مشکلاتی همچون انفجار یا ناپدید شدن گرادیان مواجه‌اند و برای پردازش توالی‌های بلند، نیازمند حافظه‌ای با پیچیدگی محاسباتی نمایی هستند. برای رفع این محدودیت‌ها، شبکه‌های حافظه کوتاه‌مدت طولانی (Long Short-Term Memory - LSTM) معرفی شدند. اگرچه LSTMها برخی از مشکلات RNNهای سنتی را برطرف کردند، اما با افزودن سازوکارهای دروازه‌ای (Gating Mechanisms) پیچیدگی بیشتری ایجاد کردند و در یادگیری انتقالی (Transfer Learning) نیز با چالش‌هایی روبه‌رو هستند.

شکل ۱: دسته‌بندی مدل‌های فضای حالت (State Space Models - SSMs) بر اساس ماهیت ساختاری، بازگشتی و دروازه‌ای آن‌ها. در هر دسته، مدل‌های کلیدی معرفی‌شده در متون علمی بررسی شده‌اند.

مدل‌های Transformer به عنوان جایگزینی انقلابی ظهور کرده‌اند که راه‌حل‌هایی برای کاستی‌های هر دو شبکه‌ی RNN و LSTM ارائه می‌دهند و به پارادایم غالب در حوزه‌های NLP و بینایی رایانه‌ای تبدیل شده‌اند. با به‌کارگیری سازوکارهای توجه (Attention Mechanisms)، ترنسفورمرها این امکان را فراهم می‌سازند که هر توکن با تمام توکن‌های دیگر در توالی ورودی تعامل داشته باشد و در نتیجه، وابستگی‌های بلندبرد را به‌صورت کارآمد ثبت کنند. با این حال، پیچیدگی توجه در مرتبه‌ی $O(N^2)$ در ترنسفورمرها چالش‌هایی در مقیاس‌پذیری ایجاد می‌کند، به‌ویژه هنگام پردازش توالی‌های بلند در حوزه‌هایی مانند ژنومیک و تحلیل تصاویر با وضوح بالا. این بدان معناست که اگرچه ترنسفورمرها در بسیاری از وظایف پردازش توالی عملکرد بسیار خوبی دارند، اما ناکارآمدی‌هایی نیز دارند ، به‌ویژه از نظر میزان حافظه و توان محاسباتی موردنیاز که با افزایش طول توالی به‌صورت درجه دوم رشد می‌کند. در این زمینه، مطالعه‌ی حاضر به بررسی چشم‌انداز مدل‌سازی توالی پرداخته و مدل‌های فضای حالت (State Space Models - SSMs) را به عنوان پارادایمی نویدبخش معرفی می‌کند. با بهره‌گیری از SSMها، هدف ما رفع ناکارآمدی‌های RNNها و ترنسفورمرهای سنتی و ارائه‌ی راهکارهایی مقیاس‌پذیر برای وظایف پردازش توالی در حوزه‌های گوناگون است. مدل‌های فضای حالت (State Space Models - SSMs) به‌عنوان جایگزین‌هایی قابل‌توجه برای ترنسفورمرها مطرح شده‌اند، به‌ویژه برای پردازش توالی‌های بلند. می‌توان SSMها را به‌صورت مفهومی، همانند شبکه‌های بازگشتی (RNNs) با طول‌های ثابت در نظر گرفت که اندازه‌ی آن‌ها با طول ورودی افزایش نمی‌یابد. این ویژگی منجر به مزایای چشمگیری از نظر سرعت استنتاج و پیچیدگی محاسباتی و حافظه در مقایسه با ترنسفورمرها می‌شود. با این حال، علیرغم مزایای کارایی، SSMها در برخی حوزه‌های داده، به‌ویژه در وظایف بینایی رایانه‌ای، هنوز به سطح عملکرد مدل‌های پیشرفته‌ی ترنسفورمرها نمی‌رسند. یکی از نقاط ضعف قابل‌توجه SSMها، مصالحه در برخی قابلیت‌های بنیادی است که برای وظایف خاص پردازش توالی ضروری‌اند، از جمله توانایی کپی کردن توالی‌های ورودی بلند [۶۳]، یادگیری درون‌متنی (In-context Learning) و سازوکارهای القایی (Induction Heads) [۱۰۹]. این پژوهش، یک مرور جامع از حوزه‌ی مدل‌های فضای حالت ارائه می‌دهد و نقاط قوت و ضعف آن‌ها را در مقایسه با ترنسفورمرهای پیشرفته بررسی می‌کند. در این مقاله نشان داده می‌شود که چگونه SSMهای وظیفه‌محور برای برخی وظایف خاص مانند وظایف حوزه‌ی Long Range Arena (LRA) [۱۴۹] مناسب‌تر از ترنسفورمرها هستند، هرچند عملکرد ترنسفورمرها در وظایف بینایی رایانه‌ای مانند شناسایی تصویر و تقسیم‌بندی نمونه‌ها همچنان برتر است. در مرور ما، مشاهده می‌کنیم که State Space Models (SSMs)، به‌ویژه مدل‌هایی مانند Mamba، در حوزه‌ها و وظایف گوناگون عملکردی رقابتی با transformers دارند. برای نمونه، در حیطه‌ی Language Domain Tasks، SSMs عملکرد مطلوبی نشان می‌دهند، به‌ویژه در وظایف standard regression in-context learning (ICL). شایان ذکر است که SSMs در وظایفی مانند sparse parity learning از transformers پیشی می‌گیرند. با این حال، در حوزه‌ی Video and Audio Tasks، هرچند transformers در natural language processing امیدبخش‌اند، کارایی آن‌ها در وظایف چندوجهی مانند video and audio understanding نسبتاً کمتر کاوش شده باقی مانده است. برای پر کردن این خلأ، تلاش‌های پژوهشی جاری می‌کوشند از ظرفیت مدل‌سازی بلندبُرد Large Language Models (LLMs) برای بهبود video understanding بهره بگیرند. علاوه بر این، مرور ما بر نیاز به بررسی بیشتر درباره‌ی قوت‌های نسبی SSMs و transformers در طیفی از وظایف time including series prediction، recommendation systems، reinforcement learning و انواع وظایف حوزه‌ی پزشکی تأکید می‌کند. این حوزه‌ها بستری حاصلخیز برای کاوش در این‌باره فراهم می‌آورند که چگونه SSMs و transformers می‌توانند یکدیگر را تکمیل کنند و بینش‌هایی درباره‌ی توانمندی‌ها و محدودیت‌های هر یک ارائه دهند.

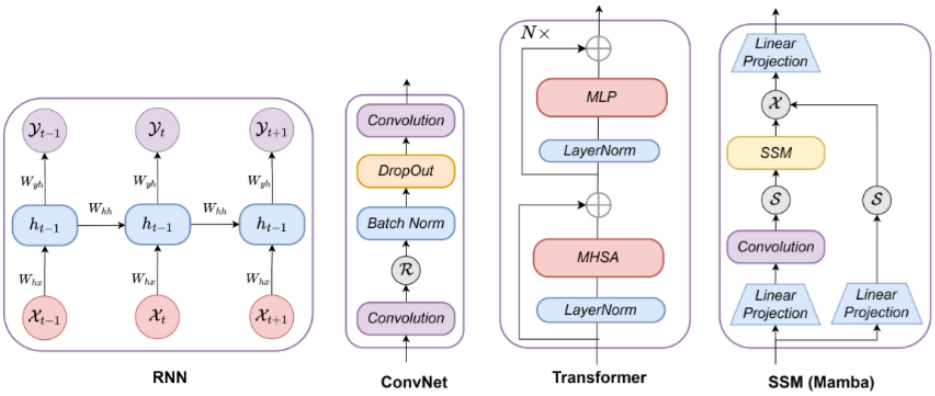
علاوه بر این، مرور ما بر نیاز به بررسی بیشتر درباره‌ی قوت‌های نسبی SSMs و transformers در طیفی از وظایف، از جمله time series prediction، recommendation systems، reinforcement learning و انواع وظایف حوزه‌ی پزشکی تأکید می‌کند. این حوزه‌ها بستری حاصلخیز برای کاوش در این زمینه فراهم می‌کنند که چگونه SSMs و transformers می‌توانند یکدیگر را تکمیل کرده و بینش‌هایی درباره‌ی توانایی‌ها و محدودیت‌های هر یک ارائه دهند. این پژوهش شامل چهار جنبه‌ی مهم است که عبارت‌اند از:

- درک مدل‌های فضای حالت (**State Space Models - SSMs**): این پژوهش به بررسی اصول بنیادی SSMها می‌پردازد و عملکرد درونی و پایه‌های ریاضی آن‌ها را توضیح می‌دهد.
- دسته‌بندی و پیشرفت‌های اخیر در **SSMs**: این بخش، دسته‌بندی نظام‌مندی از SSMها ارائه می‌دهد و به پیشرفت‌های اخیر در این حوزه می‌پردازد. با سازمان‌دهی SSMها، پژوهشگران می‌توانند دیدگاه‌های تازه‌ای درباره‌ی ویژگی‌های منحصربه‌فرد و کاربردهای بالقوه‌ی آن‌ها کسب کنند.
- کاربرد **SSMs** در حوزه‌های مختلف: این مطالعه بررسی می‌کند که چگونه SSMها در حوزه‌های گوناگون، از پردازش زبان طبیعی تا تشخیص‌های پزشکی، به‌کار گرفته می‌شوند. درک این کاربردها به متخصصان کمک می‌کند تا SSMها را برای وظایف خاص به‌طور مؤثر به‌کار برند.
- مقایسه‌ی عملکرد **SSMs** با **Transformers**: با گردآوری نتایج پیشرفته، این پژوهش عملکرد SSMها را در کنار ترنسفورمرها ارزیابی می‌کند. این تحلیل مقایسه‌ای، ما را از نقاط قوت و ضعف هر رویکرد در حوزه‌ها و وظایف مختلف آگاه می‌سازد.

به‌طور خلاصه، ما در بخش ۲ به معرفی اصول پایه‌ای State Space Models (SSMs) پرداخته و اصول بنیادی آن‌ها، شامل فرمول‌بندی ریاضی و چارچوب مفهومی‌شان را مورد بحث قرار می‌دهیم. در بخش ۳، پیشرفت‌های اخیر در ادبیات SSM را مرور کرده و به دستاوردهای جدید و پژوهش‌های نوآورانه در این زمینه می‌پردازیم. در بخش ۴، کاربردهای SSMها را در مدل‌سازی توالی‌های بلند بررسی کرده و کارایی و عملکرد آن‌ها را در مقایسه با سایر مدل‌ها تحلیل می‌کنیم. در بخش ۵، عملکرد SSMها در مدل‌سازی توالی‌های بلند را با ترنسفورمرهای پیشرفته مقایسه می‌کنیم. نشان داده می‌شود که اگرچه SSMها از نظر کارایی نسبت به ترنسفورمرها برتری‌هایی دارند، اما در برخی حوزه‌ها همچنان از نظر عملکرد نسبت به ترنسفورمرهای پیشرفته عقب‌تر هستند. تحول رویکردهای مدل‌سازی داده‌های ترتیبی از شبکه‌های عصبی بازگشتی (Recurrent Neural Networks - RNNs)، شبکه‌های عصبی هم‌نهشتی (Convolutional Neural Networks - CNNs) و ترنسفورمرها تا مدل‌های فضای حالت (State Space Models - SSMs)، مسیری از نوآوری‌ها را نشان می‌دهد که به‌منظور رفع چالش‌های گوناگون در ثبت وابستگی‌های زمانی، سلسله‌مراتب مکانی، تعاملات سراسری و رفتار سیستم‌های پویا شکل گرفته است، همان‌طور که در شکل ۲ نشان داده شده است.

۲ اصول مدل‌های فضای حالت

مدل‌های فضای حالت (State-Space Models) چارچوبی قدرتمند برای مدل‌سازی سیستم‌های پویا ارائه می‌دهند، به‌ویژه از طریق امکان نمایش مشتق‌های مرتبه‌بالا تنها با استفاده از مشتق‌های مرتبه‌اول و کمیت‌های برداری. به عنوان مثال، معادله‌ی دیفرانسیل مرتبه‌دوم که دینامیک یک سیستم جرم–فنر میرا را توصیف می‌کند را در نظر بگیرید:



شکل ۲: این شکل، سیر تکاملی پارادایم‌های مدل‌سازی داده‌های ترتیبی را از شبکه‌های عصبی بازگشتی (Recurrent Neural Networks - RNNs) و شبکه‌های عصبی هم‌نهشتی (Convolutional Neural Networks - CNNs) تا مدل‌های Transformer و مدل‌های فضای حالت (State-Space Models - SSMs) نشان می‌دهد و پیشرفت‌ها در زمینه‌ی ثبت پویایی‌های زمانی، سلسله‌مراتب مکانی و تعاملات پیچیده‌ی سیستم‌ها را برجسته می‌سازد.

$$m\frac{d^2y(t)}{dt^2} + c\frac{dy(t)}{dt} + ky(t) = u(t),$$

در اینجا، $u(t)$ نیروی خارجی وارد بر جرم را نشان می‌دهد و $y(t)$ موقعیت عمودی را مشخص می‌کند. در این معادله، $\frac{dy(t)}{dt}$ و $\frac{d^2y(t)}{dt^2}$ به ترتیب نمایانگر مشتق‌های مرتبه‌ی اول و دوم y هستند.

برای بازنویسی این معادله صرفاً بر حسب مشتق‌های مرتبه‌ی اول و کمیت‌های برداری، بردار حالت زیر را معرفی می‌کنیم:

$$x(t) := \begin{pmatrix} y(t) \\ \dot{y}(t) \end{pmatrix}.$$

هرچند این تبدیل منجر به کار با معادله‌ای برداری به جای معادله‌ای اسکالر می‌شود:

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 \\ -\frac{c}{m} & -\frac{k}{m} \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ \frac{1}{m} \end{pmatrix} u(t).$$

موقعیت $y(t)$ نیز به صورت تابعی خطی از حالت بیان می‌شود:

$$y(t) = Cx(t),$$

که در آن $C = (1, 0)$ است.

۱.۲ سیستم جرم-فنر-میرایی (Spring-Mass-Damper System)

سیستم جرم-فنر-میرایی یک مثال کلاسیک است که برای توضیح اصول دینامیک و نظریه کنترل مورد استفاده قرار می‌گیرد. در اینجا، فرمول‌بندی ریاضی پایه برای مدل فضای حالت سیستم جرم-فنر-میرایی آورده شده است. یک سیستم جرم-فنر-میرایی را در نظر بگیرید که از جرمی به مقدار m تشکیل شده که به دیواری از طریق یک فنر با ثابت فنر k و یک میرایی با ضریب میرایی c متصل است. هدف، توصیف رفتار سیستم با استفاده از متغیرهای حالت است. جابجایی جرم با x، سرعت آن با $\dot{x}$ و نیروی خارجی وارد بر جرم با F نمایش داده می‌شود.

۱. متغیرهای حالت: ما دو متغیر حالت تعریف می‌کنیم: x_1: جابجایی جرم از موقعیت تعادل خود (موقعیت جرم نسبت به نقطه‌ی تعادل فنر). $\dot{x}_1$: سرعت جرم.

۲. دینامیک سیستم: دینامیک سیستم جرم-فنر-میرایی را می‌توان با استفاده از قانون دوم نیوتن به صورت زیر بیان کرد:

$$m\ddot{x}_1 = -kx_1 - c\dot{x}_1$$

که در آن، $\ddot{x}_1$ شتاب جرم را نشان می‌دهد. جمله‌ی اول، $-kx_1$، نیروی فنر (متناسب با جابجایی) را نمایش می‌دهد و جمله‌ی دوم، $-c\dot{x}_1$، نیروی میرایی (متناسب با سرعت) را نشان می‌دهد.

۳. فرمول‌بندی فضای حالت: مدل فضای حالت دینامیک یک سیستم را با استفاده از مجموعه‌ای از معادلات دیفرانسیل مرتبه اول بیان می‌کند. این چارچوب قدرتمندی برای توصیف سیستم‌های خطی با زمان ثابت (Linear Time-Invariant - LTI) است. مؤلفه‌های اصلی آن عبارت‌اند از:

- بردار حالت (**State Vector - x**): بردار حالت شامل متغیرهای حالتی است که وضعیت درونی سیستم را توصیف می‌کنند. آن را با $\mathbf{x}$ نمایش می‌دهیم، به طوری که $\mathbf{x} \in \mathbb{R}^n$ باشد.

- بردار ورودی (**Input Vector - u**): بردار ورودی، ورودی کنترلی یا نیروی خارجی وارد بر سیستم را نشان می‌دهد. آن را با $\mathbf{u}$ نمایش می‌دهیم، که $\mathbf{u} \in \mathbb{R}^m$ است.

- بردار خروجی (**Output Vector - y**): بردار خروجی شامل کمیت‌های قابل اندازه‌گیری مورد نظر است. آن را با $\mathbf{y}$ نمایش می‌دهیم، که $\mathbf{y} \in \mathbb{R}^p$ است.

- دینامیک سیستم (**System Dynamics**): دینامیک حالت‌ها توسط معادله‌ی دیفرانسیل مرتبه‌ی اول توصیف می‌شود. در حالت خطی با زمان ثابت (Linear Time-Invariant - LTI)، ماتریس‌های $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ در طول زمان ثابت باقی می‌مانند. دینامیک حالت در سیستم‌های LTI به صورت زیر بیان می‌شود:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

که در آن $\mathbf{x} = [x_1, \dot{x}_1]^T$ بردار حالت است و $\mathbf{u}$ ورودی (نیروی خارجی یا ورودی کنترلی) را نمایش می‌دهد. ماتریس $\mathbf{A} \in \mathbb{R}^{n\times n}$ ماتریس دینامیک و $\mathbf{B} \in \mathbb{R}^{n\times m}$ ماتریس ورودی است. ماتریس $\mathbf{A}$ و $\mathbf{B}$ به صورت زیر تعریف می‌شوند:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}.$$

در این معادله، $\dot{\mathbf{x}}$ مشتق زمانی بردار حالت را نشان می‌دهد.

معادله‌ی خروجی (Output Equation): معادله‌ی خروجی، خروجی سیستم را به حالت و ورودی مرتبط می‌کند:

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}$$

که در آن، $\mathbf{C} \in \mathbb{R}^{p\times n}$ ماتریس خروجی یا حسگر و $\mathbf{D} \in \mathbb{R}^{p\times m}$ ماتریس بازخور مستقیم است.

$$\mathbf{C} = [1 \quad 0], \quad \mathbf{D} = 0$$

۴. تفسیر (**Interpretation**): بردار حالت $\mathbf{x}$ شامل اطلاعاتی درباره‌ی موقعیت و سرعت جرم است. بردار ورودی $\mathbf{u}$ می‌تواند نیروی خارجی وارد بر جرم را نمایش دهد. بردار خروجی $\mathbf{y}$ معمولاً جابجایی x_1 است. در این سیستم، بردار حالت $\mathbf{x}$ موقعیت و سرعت جرم را نمایش می‌دهد. دینامیک حالت‌ها توسط معادلات حرکت مشتق‌شده از قانون دوم نیوتن کنترل می‌شود. ورودی $\mathbf{u}$ بیانگر هرگونه نیروی خارجی وارد بر جرم است که در این حالت مقدار آن برابر با صفر فرض می‌شود.

۵. پایداری و کنترل (**Stability and Control**): تحلیل پایداری شامل بررسی مقادیر ویژه‌ی ماتریس $\mathbf{A}$ است. طراحی کنترل می‌تواند از طریق تنظیم ورودی کنترلی $\mathbf{u}$ برای دستیابی به رفتار مطلوب سیستم (برای مثال، میرایی نوسانات) انجام گیرد.

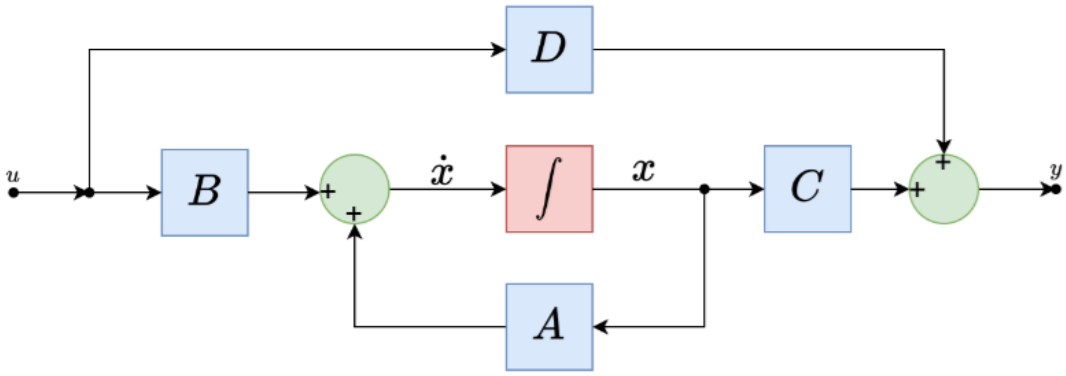
۲.۲ مدل‌های فضای حالت (**State Space Models**)

۱.۲.۲ تعریف (**Definition**)

معادلات خطی فضای حالت چارچوبی چندمنظوره برای مدل‌سازی سیستم‌های دینامیکی با زمان گسسته فراهم می‌کنند:

$$x(t+1) = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t), \quad t = 0, 1, 2, \ldots$$

در اینجا، $x(t) \in \mathbb{R}^n$ وضعیت سیستم را در زمان t نشان می‌دهد و شرایط آن را در بر می‌گیرد. $u(t) \in \mathbb{R}^p$ شامل متغیرهای کنترلی است، در حالی‌که $y(t) \in \mathbb{R}^k$ خروجی‌های خاص مورد نظر را دربر دارد. ماتریس‌های A، B، C و D دارای ابعاد مناسب هستند.



شکل ۳: نمایش شماتیکی از مفهوم مدل فضای حالت (State-Space Model) که دینامیک سیستم را از طریق مجموعه‌ای از معادلات دیفرانسیل مرتبه‌ی اول توصیف می‌کند.

در اصل، یک مدل دینامیکی خطی فرض می‌کند که حالت سیستم در گام زمانی بعدی ترکیبی خطی از حالت در گام‌های زمانی پیشین و در صورت وجود، سایر ورودی‌های برون‌زا است. علاوه بر این، این مدل بیان می‌کند که خروجی تابعی خطی از بردارهای حالت و ورودی است.

در مقابل، مدل با زمان پیوسته به صورت یک معادله‌ی دیفرانسیل بیان می‌شود:

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t), \quad t \geq 0.$$

در نهایت، مدل‌های متغیر با زمان شامل ماتریس‌های A، B، C و D هستند که در طول زمان تغییر می‌کنند و نمایش منعطف‌تری از سیستم‌های دینامیکی ارائه می‌دهند.

۲.۲.۲ فرمول‌بندی مدل (Model Formulation)

برای مدل‌سازی یک توالی بزرگ، به جای استفاده از Multi-headed Self-Attention به دلیل پیچیدگی آن، از مدل‌های فضای حالت استفاده می‌کنیم. مدل فضای حالت [۴۴، ۴۱] به‌طور معمول به‌عنوان یک سیستم خطی با زمان ثابت شناخته می‌شود که تحریک ورودی $x(t) \in \mathbb{R}^L$ را از طریق یک فضای پنهان $h(t) \in \mathbb{R}^N$ به پاسخ $y(t)$ نگاشت می‌کند. مدل‌های توالی فضای حالت ساختاریافته (Structured State Space Sequence Models - S4) نسل جدیدی از مدل‌های توالی برای یادگیری عمیق هستند که به‌طور گسترده با RNNs، CNNs و مدل‌های کلاسیک فضای حالت مرتبط‌اند. از نظر ریاضی، فضاهای حالت نهفته با زمان پیوسته را می‌توان به‌صورت معادلات دیفرانسیل خطی معمولی مدل کرد که از پارامتر تحول $A \in \mathbb{R}^{N\times N}$ و پارامتر تصویر $B \in \mathbb{R}^{N\times 1}$ و $C \in \mathbb{R}^{N\times 1}$ استفاده می‌کنند، به‌صورت زیر:

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{۱}$$
$$y(t) = Cx(t) + Du(t)$$

۳.۲.۲ مدل فضای حالت گسسته‌زمان (**Discrete-time SSM**)

شکل گسسته‌ی مدل SSM از یک پارامتر مقیاس زمانی Δ برای تبدیل پارامترهای پیوسته‌ی A، B و C به پارامترهای گسسته‌ی $\bar{A}$، $\bar{B}$ و $\bar{C}$ با استفاده از فرمول ثابت زیر بهره می‌برد: $\bar{A} = f_A(\Delta, A)$، $\bar{B} = f_B(\Delta, A, B)$. زوج f_A, f_B قاعده‌ی گسسته‌سازی را نشان می‌دهد که از نگه‌داری مرتبه صفر (Zero-Order Hold - ZOH) برای این تبدیل استفاده می‌کند. معادلات به صورت زیر بیان می‌شوند:

$$\begin{aligned} x_k &= \bar{A}x_{k-1} + \bar{B}u_k & \bar{A} &= (I - \Delta/2 \cdot A)^{-1}(I + \Delta/2 \cdot A) \\ y_k &= \bar{C}x_k & \bar{B} &= (I - \Delta/2 \cdot A)^{-1}\Delta B & \bar{C} &= C. \end{aligned} \tag{۲}$$

۴.۲.۲ نمایش هسته‌ی کانولوشنی (**Convolutional Kernel Representation**)

شکل گسسته‌ی مدل بازگشتی SSM در معادله‌ی ۲ به دلیل ماهیت ترتیبی آن، از نظر عملی قابل آموزش نیست. برای دستیابی به نمایش کارآمدتر، هم‌نهشتی پیوسته را به‌عنوان هم‌نهشتی گسسته مدل می‌کنیم، زیرا این سیستم یک سیستم خطی با زمان ثابت (Linear Time-Invariant System - LTI) است. برای سادگی، حالت اولیه را $x_{-1} = 0$ در نظر می‌گیریم. سپس، بازکردن صریح معادله‌ی ۲ نتایج زیر را به دست می‌دهد:

$$\begin{aligned} x_0 &= \bar{B}u_0 & x_1 &= \overline{AB}u_0 + \bar{B}u_1 & x_2 &= \bar{A}^2\bar{B}u_0 + \overline{AB}u_1 + \bar{B}u_2 & &\cdots \\ y_0 &= \overline{CB}u_0 & y_1 &= \overline{CAB}u_0 + \bar{C}Bu_1 & y_2 &= \overline{CA}^2\bar{B}u_0 + \overline{CAB}u_1 + \overline{CB}u_2 & &\cdots \end{aligned}$$

این نتایج را می‌توان به یک convolution ۳ برداری‌سازی کرد، همراه با فرمول صریح برای convolution kernel ۴.

$$\begin{aligned} y_k &= \overline{CA}^k\overline{B}u_0 + \overline{CA}^{k-1}\overline{B}u_1 + \cdots + \overline{CAB}u_{k-1} + \overline{CB}u_k \\ y &= \overline{K} \times u. \end{aligned} \tag{۳}$$

$$\overline{K} \in \mathbb{R}^L := k_L(\overline{A}, \overline{B}, \overline{C}) := (\overline{C}\overline{A}^i\overline{B})_{i\in[L]} = (\overline{C}\overline{B}, \overline{C}\overline{A}\overline{B}, \cdots, \overline{C}\overline{A}^{L-1}\overline{B}.) \quad (۴)$$

$\bar{K}$ در معادله‌ی ۳ می‌تواند به‌صورت یک هم‌نهشتی منفرد (غیر دایره‌ای) نمایش داده شود که با استفاده از FFTs می‌توان آن را با کارایی بالا محاسبه کرد. با این حال، محاسبه‌ی $\bar{K}$ در معادله‌ی ۴ ساده نیست و به‌عنوان $\bar{K}$، یعنی SSM Convolution Kernel یا فیلتر مدل‌سازی می‌شود.

به طور خاص، ما توالی ورودی را با استفاده از مدل پیشرفته‌ی فضای حالت Mamba [۴۱] مدل‌سازی می‌کنیم. مدل Mamba یک ضعف اساسی در مدل‌های موجود را شناسایی می‌کند: ناتوانی در انجام استدلال مبتنی بر محتوا. برای رفع این مشکل، Mamba فضاهای حالت انتخابی (Selective State Spaces - SSMs) را معرفی می‌کند که به مدل اجازه می‌دهند تا به‌صورت انتخابی اطلاعات را در امتداد بُعد طول توالی، بر اساس توکن فعلی، انتشار داده یا فراموش کند. در حالی که ما بلوک Mamba را برای وظیفه‌ی بینایی اعمال می‌کنیم، با مشکل ناپایداری (مانند از دست رفتن همگرایی) نسبت به مدل‌های دیگر مانند S4 یا Hippo مواجه می‌شویم. ما نوعی راه‌حل برای مشکل ناپایداری ارائه می‌دهیم که تنها مقادیر ویژه‌ی منفی را حفظ می‌کند. برای انجام این کار، به یک ماژول اضافی نیاز داریم که آن را «ترکیب کانال» (Channel Mixing) می‌نامیم، که در بلوک Mamba وجود نداشت. ما ماژول ترکیب کانال را با ماژول ترکیب توالی ادغام کرده و معماری ساده‌شده‌ای بر پایه‌ی Mamba ایجاد می‌کنیم که Simplified Mamba Based Architecture (SiMBA) نام دارد، همان‌طور که در SiMBA [۱۱۷] نشان داده شده است. توالی ورودی توکن‌ها X_{1-1} ابتدا توسط لایه‌ی نرمال‌سازی نرمال می‌شود. سپس، توالی نرمال‌شده به‌صورت خطی بر روی x و z با اندازه‌ی بُعد E نگاشت می‌شود. در ادامه، x از جهت‌های رو به جلو و عقب پردازش می‌شود. برای هر جهت، ابتدا هم‌نهشتی یک‌بعدی (1-D Convolution) را بر روی x اعمال کرده و x'_0 را به دست می‌آوریم. سپس x'_0 به‌صورت خطی بر روی B_0، C_0، Δ_0 نگاشت می‌شود. در نهایت، Δ_0 برای تبدیل $\bar{A}_0$ و $\bar{B}_0$ به کار گرفته می‌شود.

۳ پیشرفت‌های اخیر در مدل‌های فضای حالت (Recent Advances in State Space Models)

مدل‌های مبتنی بر توجه (Attention-based Transformers) انقلابی در پردازش زبان طبیعی و سایر وظایف توالی‌به‌توالی ایجاد کرده‌اند. با این حال، این مدل‌ها با محدودیت‌هایی روبه‌رو هستند، به‌ویژه هنگام پردازش توالی‌های ورودی بلند، زمانی که وابستگی‌ها فراتر از اندازه‌ی پنجره‌ی توجه گسترش می‌یابند. این محدودیت در کاربردهایی مانند تحلیل تصاویر با وضوح بالا و ژنومیک اهمیت ویژه‌ای دارد. تلاش‌هایی برای رفع این محدودیت‌ها توسط Efficient 360 مورد بررسی قرار گرفته است، که تمرکز آن بر بهینه‌سازی و بهبود کارایی از نظر پیچیدگی محاسباتی است. جنبه‌های مختلفی از ترنسفورمرها، از جمله تحلیل طیفی (Spectral Analysis)، ملاحظات عدالت (Fairness Considerations)، روش‌های تقریب، بهبود مقاومت (Robustness Enhancements) و بهینه‌سازی پیچیدگی محاسباتی مورد بحث قرار گرفته‌اند. در این گزارش، به بررسی این محدودیت‌ها پرداخته و مدل‌های فضای حالت (State Space Models - SSMs) را به‌عنوان رویکردی جایگزین مورد بررسی قرار می‌دهیم.

- پیچیدگی محاسباتی (**Computational Complexity**) ترنسفورمرها دارای نیازهای محاسباتی بالایی هستند، به‌ویژه در مدل‌های بزرگ. این پیچیدگی چالش‌هایی را برای آموزش و پیاده‌سازی آن‌ها بر روی دستگاه‌هایی با منابع محدود ایجاد می‌کند.

- نیاز زیاد به حافظه (**Large Memory Requirements**) ترنسفورمرها به منابع حافظه‌ی قابل‌توجهی برای ذخیره‌ی تعبیه ها (Embeddings) و فعال‌سازی‌های میانی نیاز دارند. این امر می‌تواند مقیاس‌پذیری را، به‌ویژه در توالی‌های بسیار بلند، محدود کند، زیرا ممکن است از ظرفیت حافظه‌ی موجود فراتر رود.

- طول ثابت توالی (**Fixed Sequence Length**) ترنسفورمرها به‌دلیل استفاده از تعبیه های موقعیتی (Positional Embeddings) به ورودی‌هایی با طول ثابت متکی هستند. مدیریت کارآمد ورودی‌هایی با طول متغیر، چالشی قابل‌توجه در معماری‌های مبتنی بر ترنسفورمر محسوب می‌شود.

- مقیاس‌پذیری سازوکار توجه (**Attention Mechanism Scalability**): در حالی که سازوکار توجه ابزاری قدرتمند است، دارای رشد درجه‌دوم نسبت به طول توالی ورودی است. این ویژگی آن را برای توالی‌های بسیار بلند کم‌بازده می‌کند.

- فقدان علیت در توجه استاندارد (**Lack of Causality in Standard Attention**) سازوکار توجه خودکار استاندارد که در ترنسفورمرها استفاده می‌شود، ذاتاً علیت را در نظر نمی‌گیرد. این سازوکار تمام موقعیت‌ها را به‌صورت برابر در نظر می‌گیرد که می‌تواند در وظایفی که علیت در آن‌ها اهمیت دارد، مشکل‌ساز شود.

با وجود پیشرفت‌ها، ترنسفورمرهای مبتنی بر توجه همچنان در پردازش توالی‌های بلند با چالش مواجه‌اند و در آزمون‌های مرجع بلندبرد، مانند وظیفه‌ی path-X، مشکلات حل‌نشده‌ای دارند. برای رفع این محدودیت‌ها، مدل‌های فضای حالت (State-Space Models - SSMs) رویکردی نویدبخش ارائه می‌دهند؛ مدل‌های پیشگامی مانند S4 از نخستین مدل‌هایی بودند که توانستند به‌طور مؤثر مسئله‌ی path-X را حل کنند. مدل‌های SSM با کارایی بالا توالی‌های بلند را مدل‌سازی کرده و در عین حال وابستگی‌های بلندمدت را به‌خوبی ثبت می‌کنند. در بخش‌های بعدی، مدل‌های کلیدی فضای حالت که در متون علمی معرفی شده‌اند را دسته‌بندی و بررسی می‌کنیم. دسته‌بندی مدل‌های اصلی فضای حالت در شکل ۴ نشان داده شده است:

- مدل‌های ساختاریافته‌ی فضای حالت (**Structured SSMs**): این مدل‌ها بر پایه‌ی S4 و نسخه‌های گوناگون آن ساخته شده‌اند و شامل Hippo، H3، HyenaHierarchy، Liquid-S4، S4nd، DSS و هم‌نهشتی‌های سراسری (Global Convolutions) و نسخه‌های آن مانند LongConv، FFTFlashConv و SG-Conv هستند. همچنین مدل‌های پایه‌ای مانند LD-Stack و مدل مشتق‌شده‌ی آن S5 نیز در این دسته قرار می‌گیرند. این مدل‌ها رویکردی اصولی برای مدیریت وابستگی‌های بلندبرد فراهم می‌کنند.

- مدل‌های بازگشتی فضای حالت (**Recurrent SSMs**): این مدل‌ها بر اساس RNNs و نسخه‌های آن‌ها مانند RWKV، LRU و HGRN ساخته شده‌اند و جایگزینی برای رویکردهای مبتنی بر توجه در مدل‌سازی توالی ارائه می‌دهند.

- مدل‌های دروازه‌دار فضای حالت (**Gated SSMs**): مدل‌هایی مانند GSS، Mega و TNN در این دسته قرار می‌گیرند که با استفاده از تکنیک‌های دروازه‌ای (Gating Techniques) عملکرد مدل را در توالی‌های بلند بهبود می‌بخشند.

- مدل‌های متفرقه‌ی فضای حالت (**Miscellaneous SSMs**): مدل‌هایی مانند MambaFormer، Mamba-Byte و Mamba-MoE از تکنیک‌های گوناگونی فراتر از سازوکار توجه استاندارد استفاده می‌کنند و با ترکیب ایده‌هایی از دسته‌های مختلف، مدل‌سازی کارآمد توالی‌ها را دنبال می‌کنند.

با این حال باید توجه داشت که به عنوان مثال، مدل Mamba از هر دو مدل Hippo (به‌عنوان یک Structured SSM) مشتق شده و همچنین از فناوری دروازه‌ای (Gating Technology) بهره می‌برد. این ارتباط در نمودار با پیکان‌ها نشان داده شده است. به‌طور مشابه، مدل GSS که یکی از مدل‌های پایه در دسته‌ی ساختاریافته محسوب می‌شود، از DSS مشتق شده اما از سازوکار دروازه‌ای نیز استفاده می‌کند. همچنین، مدل S5 از یکی از مدل‌های پایه، یعنی LDStack و S4، مشتق شده است. مدل‌های فضای حالت (State Space Models) راه‌حل‌هایی نویدبخش برای مدیریت توالی‌های بلند ارائه می‌دهند و کارایی و اثربخشی آن‌ها، این مدل‌ها را در برخی سناریوها به جایگزین‌هایی ارزشمند برای ترنسفورمرهای مبتنی بر توجه تبدیل کرده است.

۱.۳ مدل‌های ساختاریافته‌ی فضای حالت (**Structured State Space Models**)

مدل‌های ساختاریافته‌ی فضای حالت (Structured State Space Models - SSMs) شامل رویکردهای نوآورانه‌ی گوناگونی برای مدل‌سازی توالی هستند، از جمله S4، HiPPO، H3 و Liquid-S4. این مدل‌ها از سازوکارهای پیشرفته‌ای مانند عملگرهای تصویر چندجمله‌ای ، سیستم‌های چندورودی–چندخروجی (Multi-Input Multi-Output Systems) و هسته‌های هم‌نهشتی (Convolutional Kernels) برای ثبت وابستگی‌های بلندبرد به‌صورت کارآمد بهره می‌برند. این مدل‌ها در مجموعه‌داده‌های مرجع متنوع عملکردی رقابتی از خود نشان داده‌اند و توانایی آن‌ها در مدیریت داده‌های ترتیبی با کارایی محاسباتی بهبودیافته را به نمایش می‌گذارند.

۱.۱.۳ مدل **Structured State Space Sequence (S4)**

مدل S4 نوعی مدل نوآورانه برای مدل‌سازی توالی‌ها است که با هدف ثبت وابستگی‌های بلندمدت درون توالی‌ها و بر پایه‌ی مدل فضای حالت (State Space Model - SSM) طراحی شده است. این مدل ماهیتی پیوسته در زمان دارد و سه سازوکار کلیدی را معرفی می‌کند:

- عملگر تصویر چندجمله‌ای مرتبه‌بالا (**Higher-Order Polynomial Projection Operator - HiPPO**): HiPPO بر روی ماتریس‌های حالت و انتقال ورودی عمل می‌کند تا تاریخچه‌ی سیگنال را به‌صورت مؤثر در حافظه نگه دارد، و به مدل امکان می‌دهد تا وابستگی‌های بلندمدت را ثبت کند.

- پارامتردهی قطری به‌علاوه‌ی تصحیح رتبه‌پایین (**Diagonal Plus Low-Rank Parametrization**): در این سازوکار، مدل S4 ماتریس فضای حالت A را با یک اصلاح رتبه‌پایین تنظیم می‌کند تا پایداری و قابلیت قطری‌سازی آن تضمین شود.

- محاسبه‌ی کارآمد هسته‌ی هم‌نهشتی (**Efficient (Convolutional) Kernel Computation**): مدل S4 برای محاسبه‌ی مؤثر ماتریس‌های انتقال از FFT و iFFT استفاده می‌کند و بدین ترتیب پیچیدگی کلی محاسبات را به $\mathcal{O}(N \log N)$ کاهش می‌دهد.

مدل S4 پارامتردهی جدیدی برای SSM ارائه می‌دهد که در آن ماتریس A با تصحیح رتبه‌پایین تنظیم می‌شود تا پایداری و قابلیت قطری‌سازی آن تضمین گردد. این تبدیل، محاسبات SSM را به یک هسته‌ی Cauchy تبدیل می‌کند که به‌خوبی مطالعه شده است. در نتیجه، مدل‌های S4 نتایج تجربی بسیار قوی و در عین حال کارایی محاسباتی بالا را ارائه می‌دهند. مدل S4 نخستین مدل فضای حالت بود که توانست مسئله‌ی path-X را در معیار LRA حل کند و پیچیدگی محاسباتی را به $O(N \log N)$ کاهش دهد. S4 در مجموعه‌داده‌های مختلف عملکردی استثنایی از خود نشان می‌دهد؛ به‌عنوان مثال، در مجموعه‌داده‌ی ترتیبی CIFAR-10 بدون استفاده از افزایش داده یا ضرایب کمکی، به دقت ٪۹۱ دست یافت. این مدل فاصله‌ی عملکردی با ترنسفورمرها را در وظایف مدل‌سازی تصویر و زبان کاهش می‌دهد، در حالی‌که سرعت بالاتری نیز دارد. تلاش‌های بعدی، از جمله HiPPO و Long Convolutions، با هدف بهبود کارایی مدل‌های فضای حالت صورت گرفتند، اما هنوز فاصله‌ای میان عملکرد آن‌ها و ترنسفورمرهای پیشرفته وجود دارد.

۲.۱.۳ عملگرهای تصویر چندجمله‌ای مرتبه‌بالا (**High-Order Polynomial Projection Operators - HiPPO**)

مدل HiPPO بر روی ماتریس‌های حالت و انتقال ورودی اعمال می‌شود تا تاریخچه‌ی سیگنال‌ها را به‌صورت مؤثر در حافظه نگه دارد. مشاهده شده بود که تفسیر ریاضی دقیقی از ماتریس خاصی که در مدل S4 استفاده می‌شد وجود نداشت؛ این ماتریس در ابتدا برای سیستم‌های دینامیکی وابسته به زمان تعریف شده بود اما در مدل‌های فضای حالت مستقل از زمان (Time-Invariant SSMs) به کار گرفته شد. مدل HiPPO تفسیری ریاضی از S4 به‌عنوان چندجمله‌ای‌های لژاندر (Legendre Polynomials) با تغییر نمایی ارائه داد و بدین ترتیب توانایی S4 در ثبت وابستگی‌های بلندبرد را توضیح داد. چارچوب HiPPO مفهومی جالب است که مدل‌های فضای حالت (SSMs) را با تصویرسازی بر پایه‌های متعامد تعمیم‌یافته (Generalized Orthogonal Basis Projections) ترکیب می‌کند.

مدل HiPPO شامل چهار نسخه است:

- نسخه‌ی نخست از چندجمله‌ای‌های پایه‌ی فوریه‌ی کوتاه‌شده با نام HiPPO-FouT استفاده می‌کند.
- نسخه‌ی دوم با نام LagT بر پایه‌ی چندجمله‌ای‌های لاگور (Laguerre Polynomials) بنا شده است.
- نسخه‌ی سوم LegT است که از چندجمله‌ای‌های لژاندر (Legendre Polynomials) استفاده می‌کند.
- نسخه‌ی چهارم LegS نام دارد و از چندجمله‌ای‌های لژاندر با پنجره‌ی لغزان بهره می‌برد.

مدل HiPPO عملکرد S4 را به‌طور چشمگیری بهبود می‌دهد، به‌گونه‌ای که در معیار Long Range Arena (LRA) به دقت ٪۸۶ و در دشوارترین وظیفه مانند path-X به دقت ٪۹۶ دست یافته است.

۳.۱.۳ مدل **Hungry Hungry HiPPO (H3)**

مدل H3 دو چالش اساسی را که مدل‌های فضای حالت (SSMs) پیشین با آن روبه‌رو بودند، شناسایی کرد: نخست، دشواری در یادآوری توکن‌های اولیه به این معنا که SSMهای سنتی در حفظ مؤثر اطلاعات مربوط به توکن‌های ابتدایی در یک توالی ناتوان بودند؛ و دوم، دشواری در مقایسه‌ی توکن‌ها در میان توالی‌های مختلف. برای رفع این محدودیت‌ها، مدل H3 رویکردی نوآورانه ارائه می‌دهد که بر سه مؤلفه‌ی کلیدی استوار است:

- انباشته‌سازی مدل‌های **SSM** با تعامل‌های ضربی (**Stacked SSMs with Multiplicative Interactions**): در این روش، دو مدل فضای حالت به‌صورت انباشته با تعامل‌های ضربی بین تصویرهای ورودی و خروجی ترکیب می‌شوند. این طراحی موجب بهبود حفظ حافظه و توانایی مقایسه‌ی بین‌توالی می‌شود.

- استفاده از **FlashConv** برای افزایش کارایی آموزش (**FlashConv for Training Efficiency**): H3 برای بهبود کارایی آموزش بر روی سخت‌افزارهای مدرن از روش FlashConv استفاده می‌کند. FlashConv یک الگوریتم FFT بلوکی ترکیبی (Fused Block Fast Fourier Transform) است که به‌طور ویژه برای توالی‌هایی با طول حداکثر ۸۰۰۰ طراحی شده است.

- الگوریتم انتقال حالت برای مقیاس‌پذیری (**State-Passing Algorithm for Scaling**): برای گسترش مقیاس مدل‌های فضای حالت فراتر از محدودیت طول ،□۸ مدل H3 الگوریتمی به نام انتقال حالت معرفی می‌کند. این الگوریتم ورودی را به بزرگ‌ترین بخش‌های ممکن تقسیم می‌کند که بتوانند در حافظه‌ی SRAM پردازنده‌ی گرافیکی (GPU) جای گیرند.

با وجود این پیشرفت‌ها، هنوز فاصله‌ای از نظر پیچیدگی (Perplexity) میان H3 و ترنسفورمرها (با ۳.۱ میلیارد پارامتر) وجود دارد. با این حال، مدل H3 در سناریوهای یادگیری بدون‌نمونه (Zero-Shot) و کم‌نمونه (Few-Shot) عملکردی برتر نسبت به ترنسفورمرها نشان می‌دهد، به‌ویژه در آزمون‌های مرجع SuperGLUE. به‌طور چشمگیری، H3 در معیار Long Range Arena (LRA) سرعتی دو برابر ($2\times$) نسبت به ترنسفورمرها دارد و در مدل‌های زبانی ترکیبی، تولید متن را با سرعتی $2.4\times$ بیشتر از ترنسفورمرها ممکن می‌سازد.

۴.۱.۳ هم‌نهشتی سراسری (**Global Convolution**)

نویسندگان Long Convolution بر این باورند که مدل‌های فضای حالت (SSMs) برای آموزش مؤثر در شبکه‌های عمیق به ساختارهای ریاضی پیچیده‌ای متکی هستند. در این مدل، یک هسته‌ی هم‌نهشتی به طول توالی ورودی تولید می‌شود که با ضرب مکرر یک ماتریس حالت پنهان ایجاد می‌گردد. این فرآیند منجر به ناپایداری می‌شود و نیاز به مقداردهی اولیه‌ی دستی و تنظیم دقیق اَبَرپارامترها دارد. برای رفع این چالش‌ها، نویسندگان اصلی Long Convolution روشی نوین را معرفی کردند که در آن، هسته‌های هم‌نهشتی بلند مستقیماً پارامتردهی می‌شوند. در پیاده‌سازی‌های معمول، هم‌نهشتی‌های بلند از تبدیل فوریه سریع (Fast Fourier Transform - FFT) استفاده می‌کنند، که در برخی شرایط سیستمی می‌تواند حتی از توجه درجه‌دوم نیز کندتر عمل کند. مدل Long Convolution این مشکلات را با به‌کارگیری روش‌های ساده‌ای مانند منظم‌سازی (Regularization) شامل هموارسازی و فشرده‌سازی و الگوریتمی آگاه از ورودی/خروجی با نام FlashButterfly برطرف می‌کند. به‌طور چشمگیری، Long Convolution در معیار WikiText103 عملکردی بهتر از ترنسفورمرها ارائه می‌دهد، به‌طوری‌که میزان پیچیدگی (Perplexity) را به اندازه‌ی 0.2 کاهش می‌دهد در حالی‌که ۳۰٪ پارامتر کمتر دارد. علاوه بر این، Long Convolution مدعی است که در معیار سرعت LRA، تا 7.2 برابر سریع‌تر از ترنسفورمرها عمل می‌کند. تلاش مرتبط دیگری، FlashFFTConv است که از دو الگوریتم هم‌نهشتی خلوت (Sparse Convolution) شامل Frequency-Sparse Convolution و Partial Convolution بهره می‌گیرد. FlashFFTConv از تکنیک تجزیه‌ی ماتریسی استفاده می‌کند که FFT را از طریق ضرب ماتریسی محاسبه می‌کند و این امر ادغام هسته‌ها را برای توالی‌های بلند با کاهش عملیات ورودی/خروجی ممکن می‌سازد. رویکردهای مشابهی مانند Structural Global Convolution (SGConv) و Hyena Hierarchy نیز در این زمینه ارائه

شده‌اند.

۵.۱.۳ سلسله‌مراتب هاینا (Hyena Hierarchy - HH)

مدل Hyena Hierarchy (HH) نوعی مدل فضای حالت (SSM) است که با هدف رفع شکاف پیچیدگی (Perplexity Gap) میان مدل‌های مبتنی بر توجه و سایر روش‌ها معرفی شده است. سازوکارهای توجه دارای هزینه‌ی محاسباتی درجه‌دوم نسبت به طول توالی هستند، که این امر دسترسی به بافت (Context) را محدود می‌کند. در حالی که روش‌های موجود از لایه‌های زیر درجه‌دوم (Sub-Quadratic Layers) بر پایه‌ی تقریب‌های رتبه‌پایین و خلوت (Low-Rank and Sparse Approximations) استفاده می‌کنند، برای دستیابی به عملکردی معادل با ترنسفورمرها، اغلب به لایه‌های توجه‌ی متراکم نیاز دارند. مدل HH نشان می‌دهد که جایگزین‌های توجه زیر درجه‌دوم پیشین هنوز دچار شکاف پیچیدگی نسبت به سازوکار توجه هستند، به‌ویژه در توالی‌های متنی بلند این موضوع آشکارتر است. مدل Hyena رویکردی زیر درجه‌دوم اتخاذ می‌کند که در آن، توجه با هم‌نهشتی‌های بلند با پارامتردهی ضمنی (Implicitly Parametrized Long Convolutions) و دروازه‌گذاری داده‌ای (Data-Gating) جایگزین می‌شود. نتایج نشان داده‌اند که نقطه‌ی برتری عملکردی بین HH و سازوکار توجه در طول توالی حدود $6K$ رخ می‌دهد، و در طول توالی $100K$، مدل HH تا ۱۰۰ برابر سریع‌تر از Flash Attention یکی از بهینه‌ترین اشکال سازوکار توجه عمل می‌کند. علاوه بر این، HH تقریباً شکاف عملکردی با ترنسفورمرها را در معیار Pile Benchmark از بین برده است، که نشان‌دهنده‌ی کارایی بالای آن در پردازش توالی‌های بلند و حفظ عملکرد رقابتی با مدل‌های مبتنی بر توجه است.

۶.۱.۳ مدل RWKV

مدل RWKV یکی از شبکه‌های عصبی بازگشتی (RNN) جدید برای مدل‌سازی زبان است که از سازوکار تقریب توجه خطی با نام WKV استفاده می‌کند. این مدل از بازگشت‌های خطی با زمان ثابت (Linear Time-Invariant - LTI) بهره می‌برد و از دید مفهومی به‌عنوان نسبت دو مدل فضای حالت (State Space Models - SSMs) در نظر گرفته می‌شود. مدل RWKV بر پایه‌ی تقریب توجه خطی طراحی شده است که با سازوکار توجه خودکار سنتی (Self-Attention) مورد استفاده در ترنسفورمرها تفاوت دارد. نویسندگان RWKV بیان می‌کنند که اگرچه ترنسفورمرها انقلابی در پردازش زبان طبیعی (Natural Language Processing - NLP) و حوزه‌های دیگر ایجاد کرده‌اند، اما از پیچیدگی محاسباتی و حافظه‌ای درجه‌دوم رنج می‌برند. در مقابل، RNNها از مقیاس‌پذیری خطی در هر دو بُعد حافظه و محاسبه برخوردارند، اما از نظر عملکرد با ترنسفورمرها فاصله دارند. با این حال، RNNها معمولاً در زمینه‌ی موازی‌سازی و مقیاس‌پذیری از ترنسفورمرها عقب‌تر هستند. مدل RWKV خود را مدلی ترکیبی معرفی می‌کند که نقاط قوت هر دو دسته یعنی RNNها و ترنسفورمرها را با هم ترکیب می‌کند. با این حال، لازم است توجه شود که بر اساس بهترین شواهد موجود، RWKV در واقع نوعی ترنسفورمر با سازوکار توجه خطی است و ویژگی‌های بازگشتی واقعی که برای RNNهای سنتی تعریف می‌شود را ندارد.

۷.۱.۳ مدل LDStack

مدل LDStack نشان می‌دهد که چگونه یک شبکه‌ی عصبی بازگشتی (Recurrent Neural Network - RNN) می‌تواند به‌صورت یک سیستم دینامیکی خطی چندورودی-چندخروجی (Multiple-Input Multiple-Output Linear Dynamical System - MIMO LDS) نمایش داده شود. این مدل اثبات می‌کند که RNN را می‌توان به‌طور مؤثر به‌صورت یک LDS مدل‌سازی کرد، و این

دیدگاه امکان درک عمیق‌تری از رفتار و ویژگی‌های RNNها را فراهم می‌سازد. در این چارچوب، RNN به‌عنوان سیستمی با چند ورودی و چند خروجی مشابه با یک LDS در نظر گرفته می‌شود. برای حل چالش‌های محاسباتی، مدل LDStack روشی به نام «پویش موازی» (Parallel Scan) را معرفی می‌کند. این روش شامل تقریب MIMO LDS از طریق تجمیع آن به یک سیستم تک‌ورودی-چندخروجی (Single Input Multiple Output - SIMO LDS) است. این تقریب ضمن حفظ ویژگی‌های اساسی، محاسبات را ساده‌تر می‌کند. نویسندگان تأکید می‌کنند که بسیاری از سیستم‌های دینامیکی خطی گسسته‌زمان (Discrete-Time LDS) را می‌توان با استفاده از معادلات فضای حالت مدل‌سازی کرد. این امر نشان می‌دهد که LDS را می‌توان معادل با مدل‌های فضای حالت متغیر با زمان در نظر گرفت. رویکرد LDStack عملکردی در سطح پیشرفته (State-of-the-Art) به‌ویژه در وظایف حافظه‌ی کپی (Copy Memory Tasks) نشان داده است. این معادل‌سازی پلی میان چارچوب RNN و مدل‌سازی رسمی‌تر فضای حالت ایجاد می‌کند. مدل LDStack با دستیابی به عملکرد سطح بالا در مسائل حافظه‌ی کپی، اثربخشی روش خود را به نمایش می‌گذارد. به‌طور قابل‌توجهی، این کار به درک گسترده‌تر از RNNها و ارتباط آن‌ها با سیستم‌های دینامیکی خطی کمک کرده و دیدگاه‌های ارزشمندی درباره‌ی رابطه‌ی میان RNN و LDS ارائه می‌دهد، که افق‌های جدیدی برای پژوهش‌ها و کاربردهای عملی در این زمینه می‌گشاید.

۸.۱.۳ مدل S5

مدل S5 اصول مدل LDStack که در آن RNNها به‌صورت سیستم‌های دینامیکی خطی چند ورودی چندخروجی (Multiple-Input Multiple-Output Linear Dynamical Systems - MIMO LDS) مدل‌سازی شده بودند را گسترش داده و آن را به مدل‌های فضای حالت (State Space Models - SSMs) تعمیم می‌دهد و رویکردی کلی‌تر ارائه می‌کند. برخلاف LDStack که مدل‌های فضای حالت مستقل را به‌صورت ترتیبی پردازش می‌کند، لایه‌ی S5 چندین ورودی و خروجی را به‌طور هم‌زمان (Concurrently) پردازش می‌کند. نتایج ارزیابی عملکرد نشان می‌دهند که مدل S5 به دقت چشمگیری دست یافته است؛ به‌طوری‌که در معیار Long-Range Arena (LRA) دقت ۸۷٪.۲ و در وظیفه‌ی دشوار path-X درون LRA دقت ۹۸٪.۵ را کسب کرده است. در حالی که مدل S4 از مجموعه‌ای از مدل‌های فضای حالت تک‌ورودی-تک‌خروجی مستقل (Single-Input Single-Output SSMs) که مبتنی بر چارچوب HiPPO هستند استفاده می‌کند، لایه‌ی S5 از یک مدل فضای حالت چندورودی-چندخروجی (Multi-Input Multi-Output SSM) واحد بهره می‌برد که امکان پردازش موازی را فراهم می‌سازد. این پردازش موازی، کارایی محاسباتی را به‌طور چشمگیری افزایش می‌دهد. مدل S5 ترکیبی جالب از ایده‌های مدل‌سازی مبتنی بر LDS و مدل‌های فضای حالت ارائه می‌دهد و بر اهمیت پردازش موازی و کارایی محاسباتی تأکید دارد.

۹.۱.۳ مدل S4nd

هدف اصلی مدل S4nd گسترش کاربرد مدل‌های فضای حالت (State Space Models - SSMs) فراتر از داده‌های ترتیبی (مانند داده‌های متنی و سری‌های زمانی) به حوزه‌ی داده‌های پیوسته است. این مدل چالش استفاده از مدل‌های فضای حالت برای داده‌های پیوسته مانند تصاویر و ویدئوها را برطرف می‌کند. مدل S4nd یک لایه‌ی جدید در یادگیری عمیق معرفی می‌کند که یک مدل فضای حالت استاندارد را که معمولاً به‌صورت یک معادله‌ی دیفرانسیل معمولی یک‌بعدی (1D Ordinary Differential Equation - ODE) بیان می‌شود به یک معادله‌ی دیفرانسیل جزئی چندبعدی (Multi-Dimensional Partial Differential Equation - PDE) تبدیل می‌کند.

این تبدیل به مدل‌های فضای حالت اجازه می‌دهد تا وابستگی‌های مکانی را در ابعاد مختلف ثبت کنند. علاوه بر این، S4nd نشان می‌دهد که یک مدل فضای حالت چندبعدی می‌تواند به‌طور معادل به‌صورت هم‌نهشتی پیوسته‌ی چندبعدی (ND Continuous Convolution) نمایش داده شود، که در آن هر بُعد به‌صورت مستقل تحت هم‌نهشتی فضای حالت یک‌بعدی (1D SSM Convolution) قرار می‌گیرد. برای ارزیابی کارایی، مدل S4nd بر روی مجموعه‌داده‌ی ImageNet با استفاده از ConvNeXt به‌عنوان مدل پایه آزمایش شد و عملکرد بهتری نسبت به روش‌های سنتی نشان داد. به‌ویژه، عملکرد مدل‌های بینایی ترنسفورمر (Vision Transformers - ViTs) را بهبود بخشید. همچنین، S4nd در وظایف شناسایی ویدئو عملکردی برتر نسبت به ConvNeXt نشان داد، به‌ویژه در مجموعه‌داده‌هایی مانند HMDB . مدل S4nd قابلیت‌های مدل‌های فضای حالت را برای پردازش داده‌های پیوسته گسترش داده و پلی میان مدل‌سازی ترتیبی و مکانی ایجاد می‌کند.

۱۰.۱.۳ مدل Diagonal State Spaces (DSS)

مدل DSS به بررسی پارامتردهی ماتریس‌های فضای حالت از طریق ماتریس‌های قطری به‌همراه تصحیح رتبه‌پایین می‌پردازد. نکته‌ی قابل‌توجه این است که DSS نشان می‌دهد می‌توان بدون نیاز به تصحیح رتبه‌پایین، عملکردی در سطح مدل S4 به‌دست آورد. با استفاده‌ی صرف از ماتریس‌های فضای حالت قطری، مدل DSS در معیارهای گوناگون از جمله طبقه‌بندی گفتار خام (Raw Speech Classification) و معیار Long-Range Arena (LRA) به عملکردی مشابه با مدل S4 دست یافته است. علاوه بر این، نسخه‌ای از DSS که در منبع پیشنهاد شده، دیدگاه‌های تازه‌ای درباره‌ی مقداردهی اولیه‌ی خاص ماتریس فضای حالت ارائه می‌دهد. این مقداردهی‌های اولیه، که بر پایه‌ی تقریب ماتریس مدل S4 طراحی شده‌اند، منجر به عملکرد مؤثر در وظایف مدل‌سازی توالی‌های بلند می‌شوند. این نسخه از مدل، انعطاف‌پذیری و پتانسیل بالای DSS را در دستیابی به عملکرد قوی در کاربردها و سناریوهای مختلف نشان می‌دهد. نوآوری کلیدی DSS در حذف مؤلفه‌ی تصحیح رتبه‌پایین از ماتریس HiPPO (که در مدل S4 استفاده می‌شود) نهفته است.

۱۱.۱.۳ مدل Liquid Structural State Spaces (Liquid-S4)

شبکه‌های با ثابت زمانی مایع (Liquid Time-Constant Networks - LTCs) نوعی شبکه‌ی عصبی پیوسته در زمان و علی (Causal Continuous-Time Neural Network) هستند که با انتقال‌های حالت وابسته به ورودی مشخص می‌شوند. با ترکیب شبکه‌های LTC با مدل S4، پژوهشگران مدل جدیدی با نام Liquid-S4 معرفی کردند که دینامیک آن به‌صورت زیر تعریف می‌شود:

$$\dot{x} = (A + Bu)x + Bu, y = Cx$$

که با دینامیک s4متفاوت است .

$$\dot{x} = Ax + Bu, y = Cx$$

انگیزه‌ی اصلی پشت مدل Liquid-S4 بهبود مدل‌های فضای حالت (SSMs) از طریق بهره‌گیری از ویژگی‌های شبکه‌های LTC است. مدل Liquid-S4 یک هسته‌ی هم‌نهشتی (Convolutional Kernel) را بر پایه‌ی نسخه‌ی خطی‌شده‌ی شبکه‌های LTC ایجاد می‌کند. این ساختار هسته شباهت میان نمونه‌های توالی ورودی را هم در طول آموزش و هم در زمان استنتاج در نظر می‌گیرد. مدل Liquid-S4

یک ماژول انتقال حالت وابسته به ورودی معرفی می‌کند که امکان سازگاری با ورودی‌های متغیر را فراهم می‌سازد. همچنین، Liquid-S4 هسته‌ی هم‌نهشتی‌ای ایجاد می‌کند که متناظر با نسخه‌ی خطی‌شده‌ی LTC است. مدل Liquid-S4 در وظایف مدل‌سازی توالی با وابستگی‌های بلندمدت (شامل تصویر، متن، صوت و سری‌های زمانی پزشکی) به تعمیمی در سطح پیشرفته (State-of-the-Art) دست یافته است. به‌طور خاص، Liquid-S4 در معیار Long-Range Arena (LRA) میانگین عملکرد ۸۷٪.۳۲ را به‌دست آورده است، در حالی‌که تنها مدل‌های دروازه‌دار فضای حالت (Gated State Space Models - GSS) عملکردی قابل مقایسه دارند. مدل Liquid-S4 در مجموعه‌داده‌ی شناسایی گفتار خام (Raw Speech Command Recognition Dataset) عملکردی بهتر از ConvNeXt ارائه داده است، به‌طوری‌که با ٪۳۰ پارامتر کمتر نسبت به S4، دقتی برابر با ۹۶٪.۸۷ به‌دست آورده است. همچنین، نمایش هم‌نهشتی مدل Liquid-SSMs مشابه با فرم ارائه‌شده در معادله‌ی ۳ معرفی شده است. در این چارچوب، مدل Liquid-SSM ابتدا در طول زمان باز می‌شود تا هسته‌ی هم‌نهشتی آن ساخته شود. با فرض $x_{-1} = 0$:

$$
\begin{aligned}
&x_0 = \overline{B}u_0, \quad y_0 = \overline{CB}u_0, \\
&x_1 = \overline{AB}u_0 + \overline{B}u_1 + \overline{B}^2 u_0 u_1, \quad y_1 = \overline{CAB}u_0 + \overline{CB}u_1 + \overline{CB}^2 u_0 u_1 \\
&x_2 = \overline{A}^2\overline{B}u_0 + \overline{AB}u_1 + \overline{B}u_2 + +\overline{AB}^2 u_0 u_1 + \overline{AB}^2 u_0 u_2 + \overline{B}^2 u_1 u_2 + \overline{B}^3 u_0 u_1 u_2, \\
&y_2 = \overline{CA}^2\overline{B}u_0 + \overline{CAB}u_1 + \overline{CB}u_2 + \overline{CAB}^2 u_0 u_1 + \overline{CAB}^2 u_0 u_2 + \overline{CB}^2 u_1 u_2 + \overline{CB}^3 u_0 u_1 u_2, \ldots
\end{aligned}
\tag{۵}
$$

مدل Liquid-S4 نقاط قوت شبکه‌های LTC و مدل‌های فضای حالت (SSMs) را با یکدیگر ترکیب می‌کند که نتیجه‌ی آن، بهبود عملکرد و افزایش سازگاری در طیف گسترده‌ای از وظایف مدل‌سازی توالی است.

۱۲.۱.۳ مدل State Space Augmented Transformer (SPADE)

مدل SPADE برای حل چالش ثبت کارآمد هم‌زمان اطلاعات سراسری و محلی از توالی‌های بلند طراحی شده است. در حالی‌که بهینه‌سازی‌های سازوکار توجه در ترنسفورمرها موجب افزایش کارایی محاسباتی می‌شوند، این مدل‌ها اغلب در ثبت مؤثر بافت سراسری (Global Context) از توالی‌های بلند دچار ضعف هستند. در مقابل، مدل‌های فضای حالت (State Space Models - SSMs) در ثبت اطلاعات سراسری از توالی‌های بلند عملکرد بسیار خوبی دارند، اما در ثبت وابستگی‌های محلی (Local Dependencies) چندان کارآمد نیستند. مدل SPADE رویکردی نوآورانه پیشنهاد می‌دهد که در آن یک مدل فضای حالت به‌ویژه مدل S4 به‌عنوان لایه‌ی پایینی ترنسفورمر ادغام می‌شود. این یکپارچگی به SPADE اجازه می‌دهد تا از نقاط قوت هر دو مدل بهره‌مند شود و اطلاعات سراسری را به‌طور کارآمد ثبت کند. علاوه بر این، SPADE از لایه‌های توجه محلی مبتنی بر پنجره و بخش‌بندی (Window and Chunk-Based Local Attention Layers) برای ثبت دقیق وابستگی‌های محلی استفاده می‌کند. با ترکیب اطلاعات سراسری ثبت‌شده توسط لایه‌ی فضای حالت با اطلاعات محلی به‌دست‌آمده از سازوکار توجه، مدل SPADE به بهبود چشمگیری در عملکرد در مجموعه‌داده‌های مختلف از جمله Long-Range Arena (LRA)، WikiText و GLUE دست یافته است. این ترکیب مدل‌های SSM و ترنسفورمر در SPADE جهتی نویدبخش برای ارتقای کارایی و عملکرد در وظایف مدل‌سازی توالی محسوب می‌شود.

۲.۳ مدل‌های دروازه‌دار فضای حالت (Gated SSMs)

مدل‌های Gated State Spaces (GSS)، Toeplitz Neural Network (TNN) و Mamba به‌عنوان رویکردهای نوآورانه در حوزه‌ی مدل‌های دروازه‌دار فضای حالت شناخته می‌شوند. مدل GSS با استفاده از واحدهای دروازه‌ای (Gating Units) عملیات تبدیل فوریه سریع (FFT) را بهینه‌سازی می‌کند، که نتیجه‌ی آن پردازش کارآمد توالی‌ها و دستیابی به عملکردی رقابتی است. مدل TNN با معرفی ماتریس توپلیتز دارای رمزگذاری مکانی (Position-Encoded Toeplitz Matrix) برای ترکیب توکن‌ها (Token Mixing)، به‌طور قابل‌توجهی پیچیدگی زمان و فضا را کاهش می‌دهد، در حالی‌که نتایجی در سطح پیشرفته (State-of-the-Art) حفظ می‌کند. مدل Mamba ناکارآمدی‌های محاسباتی در مدل‌های فضای حالت سنتی را از طریق ترکیب شبکه‌ی پرسپترون چندلایه‌ی دروازه‌دار (Gated MLP) و الگوریتم‌های آگاه از سخت‌افزار (Hardware-Aware Algorithms) برطرف می‌کند. این مدل دارای پیچیدگی زمانی خطی است و کارایی بالاتری نسبت به ترنسفورمرهای متداول ارائه می‌دهد.

۱.۲.۳ مدل Mega

مدل Mega دو ضعف اساسی در ترنسفورمرها را شناسایی می‌کند: نخست، سوگیری استقرایی ضعیف (Weak Inductive Bias)، و دوم، پیچیدگی درجه‌دوم سازوکار توجه (Quadratic Attention Complexity) نسبت به طول توالی. سوگیری استقرایی ضعیف به این معناست که ترنسفورمرها هیچ فرضی درباره‌ی الگوهای تعامل یا وابستگی میان توکن‌ها ندارند و همه‌ی موقعیت‌ها را به‌طور یکسان در نظر می‌گیرند، که این امر برای مدل‌سازی توالی‌های بلند بهینه نیست. برای رفع این مشکل، Mega سازوکار جدیدی با نام Moving Average Equipped-Gated Attention معرفی می‌کند. در این روش، سازوکار توجه دروازه‌دار (Gated Attention Mechanism) از نوع تک‌سری (Single-Head) است که میانگین متحرک نمایی کلاسیک (Exponential Moving Average - EMA) را با سازوکار توجه دروازه‌دار تک‌سری ترکیب می‌کند. این ترکیب وابستگی‌های محلی در سطح موقعیت را درون توجه غیرمکانی (Positional-Agnostic Attention) ادغام کرده و به این ترتیب، سوگیری استقرایی معنادارتری به سازوکار توجه می‌افزاید. علاوه بر این، Mega نسخه‌ای با نام Mega-Chunk ارائه می‌دهد که به مشکل پیچیدگی درجه‌دوم توجه پاسخ می‌دهد. این نسخه با معرفی پیچیدگی زمانی و فضایی خطی (Linear Time and Space Complexity) و تقسیم کارآمد کل توالی به بخش‌هایی با طول ثابت (Fixed-Length Chunks)، کیفیت مدل را بدون افت محسوس حفظ می‌کند. با ترکیب قدرت EMA و توجه دروازه‌دار، Mega سوگیری استقرایی ترنسفورمرها را بهبود داده و پیچیدگی محاسباتی ناشی از سازوکارهای توجه را کاهش می‌دهد. این نوآوری‌ها گامی مهم در جهت توسعه‌ی مدل‌سازی توالی‌های کارآمدتر و مؤثرتر به‌شمار می‌آیند. نتایج آزمایش‌های انجام‌شده توسط Ma et al. نشان می‌دهد که Mega عملکردی بهتر از سایر مدل‌های توالی، از جمله انواع مختلف ترنسفورمر، در مجموعه‌داده‌های متنوعی مانند ترجمه‌ی ماشینی عصبی (Neural Machine Translation)، مدل‌سازی زبان (Language Modeling)، و طبقه‌بندی تصویر و گفتار دارد.

۲.۲.۳ مدل Gated State Spaces (GSS)

مدل GSS بر پایه‌ی پژوهش‌های پیشین Lia et al. توسعه یافته است، که در آن‌ها مشاهده شد جایگزینی لایه‌ی MLP در ترنسفورمر با واحدهای دروازه‌ای (Gating Units) می‌تواند بُعد داده را در طی فرایند ترکیب توکن‌ها (Token Mixing) کاهش دهد. با گسترش این ایده، مدل GSS لایه‌ای دروازه‌دار معرفی می‌کند که با هدف کاهش ابعاد در طول عملیات تبدیل فوریه سریع (Fast Fourier Transform)

FFT -) در مدل‌های فضای حالت (State Space Models - SSMs) طراحی شده است. این طراحی باعث می‌شود که GSS بتواند به‌صورت کارآمد بر روی توالی‌های ورودی بلند عمل کند. مدل GSS توانایی تعمیم بدون نمونه (Zero-Shot Generalization) به توالی‌های بلندتر را دارد در حالی‌که پیاده‌سازی آن همچنان ساده باقی می‌ماند. همچنین، این مدل بینش‌های تازه‌ای ارائه می‌دهد؛ از جمله مقداردهی اولیه‌ی متغیرهای فضای حالت با مقادیر تصادفی، که برخلاف مدل‌هایی مانند HiPPO، S4 و DSS است که برای مقداردهی اولیه از تکنیک‌های جبر خطی (مانند ماتریس HiPPO) استفاده می‌کنند. به‌طور شگفت‌انگیزی، مدل GSS عملکردی بهتر از این مدل‌ها ارائه می‌دهد و شکاف پیچیدگی (Perplexity Gap) با مدل‌های بازگشتی بلوکی ترنسفورمر (Block Recurrent Transformers) را در معیارهای مختلف مانند PG-19، Arxiv و GitHub کاهش می‌دهد، در حالی‌که از کارایی بالاتری نیز برخوردار است. این یافته‌ها نشان‌دهنده‌ی کارایی و انعطاف‌پذیری بالای مدل GSS در وظایف مدل‌سازی توالی هستند. GSS در واحدهای پردازش تنسور (Tensor Processing Units - TPUs) به‌طور قابل‌توجهی سریع‌تر از نسخه‌ی قطری مدل S4 (DSS) آموزش می‌بیند و عملکردی رقابتی در مقایسه با چندین مدل ترنسفورمر با تنظیم دقیق ارائه می‌دهد.

۳.۲.۳ شبکه‌ی عصبی توپلیتز (**Toeplitz Neural Network - TNN**)

مدل TNN به‌تازگی با هدف رفع دو جنبه‌ی اساسی ترنسفورمرها معرفی شده است: نخست، سازوکار توجه (Attention Mechanism) که همبستگی‌های جفتی میان توکن‌های ورودی را می‌آموزد، و دوم، جاسازی مکانی (Positional Embedding) که سوگیری استقرایی مکانی (Positional Inductive Bias) را فرا می‌گیرد. مدل TNN از یک ماتریس توپلیتز دارای رمزگذاری مکانی (Position-Encoded Toeplitz Matrix) برای ثبت روابط میان جفت‌توکن‌ها به‌عنوان ترکیب‌کننده‌ی توکن (Token Mixer) استفاده می‌کند و با بهره‌گیری از تکنیک ضرب ماتریس-بردار توپلیتز (Toeplitz Matrix-Vector Product)، پیچیدگی زمان و فضا را به $O(N \log N)$ کاهش می‌دهد. نویسندگان ادعا می‌کنند که TNN از ترنسفورمرها کاراتر است، زیرا پیچیدگی زمانی آن به‌صورت لگاریتمی-خطی (Log-Linear Complexity) است، در حالی‌که ترنسفورمرها دارای پیچیدگی درجه‌دوم هستند. ویژگی کلیدی TNN استفاده از رمزگذار موقعیت نسبی (Relative Position Encoder - RPE) برای تولید ضرایب موقعیتی نسبی با بودجه‌ی پارامتر ثابت است. رمزگذار RPE با استفاده از پارامترهای موقعیتی نسبی، ماتریس توپلیتز را بازسازی کرده و باعث می‌شود که تعداد پارامترهای مدل TNN مستقل از طول توالی باشد. مدل TNN از دو مؤلفه‌ی اصلی تشکیل شده است:

- واحد توپلیتز دروازه‌دار (**Gated Toeplitz Unit - GTU**): از عملگر عصبی توپلیتز (Toeplitz Neural Operator - TNO) برای ترکیب توکن‌ها با استفاده از ماتریس توپلیتز بهره می‌برد.
- واحد خطی دروازه‌دار (**Gated Linear Unit - GLU**): عملکردی مشابه لایه‌های دروازه‌ای در ترنسفورمرها دارد و جریان اطلاعات را تنظیم می‌کند.

رمزگذار RPE نیز از یک شبکه‌ی کاملاً متصل (Fully Connected Network) استفاده می‌کند که با اطلاعات موقعیت رمزگذاری شده است تا ضرایب موقعیت نسبی را تولید نماید. نتایج تجربی نشان داده‌اند که TNN در معیارهای مرجع مانند GLUE و Long-Range Arena (LRA) به عملکردی در سطح پیشرفته (State-of-the-Art) دست یافته است. این یافته‌ها نشان‌دهنده‌ی کارایی و اثربخشی بالای TNN در مقایسه با معماری‌های سنتی ترنسفورمر است.

۴.۲.۳ مدل **Mamba**

مدل Mamba به پیچیدگی محاسباتی و حافظه‌ای درجه‌دوم (Quadratic Computational and Memory Complexity) در ترنسفورمرها و همچنین شکاف پیچیدگی (Perplexity Gap) میان مدل‌های فضای حالت سنتی (State Space Models - SSMs) و ترنسفورمرها اشاره دارد. پیش از معرفی Mamba، مدل‌های فضای حالت در انجام وظایفی مانند کپی انتخابی (Selective Copying) و سر القایی (Induction Head) با ناکارآمدی‌هایی مواجه بودند. با این حال، مشاهدات به‌دست‌آمده از ماژول انتقال حالت وابسته به ورودی در مدل Liquid-S4 نشان داد که چنین سازوکاری می‌تواند در حل مؤثر این وظایف مفید واقع شود. مدل Mamba با معرفی یک رویکرد پارامتردهی جدید برای مدل‌های فضای حالت بر اساس ویژگی‌های ورودی و افزودن یک سازوکار انتخابی ساده، این چالش‌ها را برطرف می‌کند. علاوه بر این، Mamba یک الگوریتم کارآمد آگاه از سخت‌افزار (Hardware-Aware Algorithm) بر پایه‌ی پویش انتخابی (Selective Scan) ارائه می‌دهد. مشابه با مدل Gated State Spaces (GSS)، مدل Mamba از یک تکنیک دروازه‌دار (Gated Technique) برای کاهش ابعاد در عملیات هسته‌ی سراسری (Global Kernel Operations) استفاده می‌کند. افزون بر این، Mamba با ترکیب شبکه‌ی پرسپترون چندلایه‌ی دروازه‌دار (Gated MLP) با ماژول فضای حالت (SSM Module) توانایی مدل را به‌طور قابل‌توجهی افزایش می‌دهد. پیچیدگی زمانی خطی (Linear-Time Complexity) در Mamba باعث می‌شود که این مدل نسبت به ترنسفورمرهای سنتی کارایی محاسباتی بسیار بالاتری داشته باشد.

۳.۳ مدل‌های بازگشتی فضای حالت (**Recurrent SSMs**)

مدل‌های بازگشتی فضای حالت (Recurrent SSMs) شامل واحد بازگشتی خطی (Linear Recurrent Unit - LRU) و شبکه‌ی عصبی بازگشتی دروازه‌دار سلسله‌مراتبی (Hierarchically Gated Recurrent Neural Network - HGRN) هستند. مدل LRU به همراه نسخه‌های گسترش‌یافته‌ی آن یعنی Griffin و Hawk، اثربخشی بازگشت خطی (Linear Recurrence)، بلوک‌های پرسپترون چندلایه (MLP Blocks) و سازوکارهای توجه (Attention Mechanisms) را در بهبود مدل‌سازی توالی‌های بلند برجسته می‌سازند. در مقابل، مدل HGRN با افزودن دروازه‌های فراموشی پویا (Dynamic Forget Gates) به شبکه‌های بازگشتی خطی (Linear RNNs)، منجر به بهبود چشمگیر در کارایی محاسباتی و ارائه‌ی عملکردی رقابتی در مجموعه‌ای از معیارهای مرجع می‌شود.

۱.۳.۳ واحد بازگشتی خطی (**Linear Recurrent Unit - LRU**)

مدل LRU مشاهده می‌کند که اگرچه مدل‌های فضای حالت (SSMs) در وظایف مدل‌سازی توالی‌های بلند عملکرد خوبی دارند، اما دلیل دقیق این موفقیت به‌طور کامل روشن نیست. به‌عنوان مثال، عملکرد مطلوب مدل S4 در ابتدا به مقداردهی اولیه‌ی خاص (HiPPO) ماتریس‌های حالت و گسسته‌سازی (Discretization) سیستم پیوسته‌ی معادلات دیفرانسیل نسبت داده می‌شد. با این حال، همان‌طور که پیش‌تر اشاره شد، مدل‌های GSS و DSS نشان داده‌اند که این مقداردهی‌های اولیه‌ی خاص ضروری نیستند و حتی مقداردهی‌های تصادفی نیز می‌توانند به بهبود قابل‌توجه عملکرد منجر شوند. بنابراین، مدل LRU رویکردی جدید اتخاذ می‌کند و با شروع از شبکه‌های عصبی بازگشتی (Recurrent Neural Networks - RNNs) بررسی می‌کند که آیا می‌توان عملکردی مشابه مدل‌های فضای حالت در وظایف مدل‌سازی توالی‌های بلند به‌دست آورد یا خیر. در این راستا، چندین بهبود کلیدی در RNNها اعمال شده است:

- بازگشت خطی (**Linear Recurrence**): حذف توابع غیرخطی از فرایند بازگشت و انباشتن لایه‌های RNN خطی همراه با بلوک‌های غیرخطی MLP.
- پارامتردهی (**Parametrization**): تبدیل پارامتری RNNها به فرم قطری مختلط (Complex Diagonal Form) برای امکان‌پذیر شدن آموزش موازی، الهام‌گرفته از مدل‌های DSS و S5.
- پارامتردهی نمایی پایدار (**Stable Exponential Parametrization**): اعمال پارامتردهی نمایی پایدار بر روی ماتریس بازگشتی قطری برای ساده‌سازی فرایند آموزش.
- نرمال‌سازی (**Normalization**): نرمال‌سازی فعال‌سازی‌های پنهان در گذر رو به جلو (Forward Pass) برای تضمین پایداری در زمان آموزش.

این بهبودها باعث می‌شوند که مدل LRU بتواند عملکردی قابل‌مقایسه با سایر مدل‌های فضای حالت و ترنسفورمرها در معیارهایی مانند Long-Range Arena (LRA) به‌دست آورد. ترکیب دیدگاه‌های به‌دست‌آمده از مدل‌های کلاسیک SSM و معماری‌های شبکه‌های عصبی راه‌های جدیدی برای بهبود مدل‌سازی توالی‌های بلند فراهم می‌کند. مدل‌های Griffin و Hawk، که توسط همان نویسندگان توسعه یافته‌اند، قابلیت‌های مدل LRU را از طریق افزودن لایه‌ها و سازوکارهای اضافی برای مدل‌سازی بهتر توالی گسترش می‌دهند. در مدل Griffin، لایه‌های LRU با بلوک‌های MLP و سازوکارهای توجه محلی (Local Attention Mechanisms) به‌صورت متناوب ترکیب می‌شوند. این معماری از سه بخش اصلی تشکیل شده است: بلوک باقیمانده (Residual Block)، بلوک MLP و بلوک ترکیب زمانی (Temporal-Mixing Block). بلوک ترکیب زمانی با سه جایگزین مورد ارزیابی قرار گرفته است: (۱) توجه سراسری چندپرسشی (Global Multi-Query Attention)، (۲) توجه محلی مبتنی بر پنجره لغزان (Local Sliding-Window Attention)، و (۳) بلوک‌های بازگشتی مبتنی بر LRU. در مقابل، مدل Hawk که توسط همان گروه توسعه یافته است رویکردی کمی متفاوت اتخاذ می‌کند و لایه‌های LRU را با بلوک‌های MLP به‌صورت متناوب ادغام می‌کند. مشابه Griffin، مدل Hawk نیز شامل یک بلوک ترکیب زمانی است که سه گزینه‌ی جایگزین برای آن تعریف شده است: توجه سراسری چندپرسشی، توجه محلی، و بلوک‌های بازگشتی مبتنی بر LRU. هر دو مدل Griffin و Hawk در مقایسه با مدل‌های قدرتمندی مانند LLaMA 2 و Mamba مورد ارزیابی قرار گرفته‌اند و اثربخشی خود را در وظایف مدل‌سازی توالی به‌خوبی نشان داده‌اند. این مدل‌ها پتانسیل بالای ترکیب سازوکارهای بازگشتی و توجه را برای بهبود مدل‌سازی توالی‌های بلند برجسته می‌کنند.

۲.۳.۳ شبکه‌ی عصبی بازگشتی دروازه‌دار سلسله‌مراتبی (**Hierarchically Gated Recurrent Neural Network - HGRN**)

مدل HGRN تلاشی نوین در اصلاح شبکه‌های عصبی بازگشتی (RNNs) است که از شبکه‌های بازگشتی خطی دروازه‌دار (Gated Linear RNNs) با دروازه‌های فراموشی (Forget Gates) استفاده می‌کند؛ دروازه‌هایی که وزن‌های قابل یادگیری دارند و از لایه‌های پایین‌تر به لایه‌های بالاتر منتقل می‌شوند. این طراحی به HGRN اجازه می‌دهد تا وابستگی‌های کوتاه‌مدت (Short-Term Dependencies) را در لایه‌های پایین‌تر، که بیشتر به اطلاعات محلی مرتبط هستند، مدیریت کند. در حالی‌که در لایه‌های بالاتر، مدل وابستگی‌های بلندمدت (Long-Term Dependencies) را که اطلاعات سراسری را دربر می‌گیرند، پردازش می‌کند. مدل HGRN برخی از نواقص اصلی

RNNها را برطرف می‌کند؛ از جمله پیچیدگی به‌روزرسانی حالت پنهان که معمولاً شامل ضرب کامل ماتریس‌ها و وجود توابع غیرخطی در واحد بازگشتی است عواملی که مانع از موازی‌سازی محاسبات می‌شوند. HGRN با استفاده از یک لایه‌ی بازگشتی خطی در سطح عنصر (Element-Wise Linear Recurrent Layer) غیرخطی بودن در بازگشت را حذف کرده و امکان آموزش موازی را فراهم می‌کند. در این مدل، به‌روزرسانی حالت پنهان از طریق ضرب در سطح عنصر (Element-Wise Multiplication) انجام می‌شود. شبکه‌های بازگشتی خطی معمولاً از دو روش استفاده می‌کنند: میانگین متحرک نمایی (Exponential Moving Average - EMA) و طرح‌های دروازه‌دار (Gating Schemes). باید توجه داشت که مدل‌های فضای حالت متداول مانند S4، RWKV، S4nd، Mega و LRU همگی از روش EMA استفاده می‌کنند که در آن نرخ‌های زوال (Decay Rates) مستقل از داده بوده و در تمامی گام‌های زمانی ثابت هستند. در مقابل، تنها دو مدل HGRN و Liquid-S4 از نرخ‌های زوال وابسته به داده (Data-Dependent) یا پویا (Dynamic Decay Rates) استفاده می‌کنند. در HGRN این نرخ‌های زوال پویا از طریق دروازه‌های فراموشی حاصل می‌شوند، در حالی‌که در Liquid-S4 از ماتریس انتقال دینامیکی (Dynamic Transition Matrix) استفاده می‌شود که نوعی محدود از FFT است. مدل HGRN نتایج چشمگیری در مجموعه‌داده‌های WikiText، GLUE، Long-Range Arena (LRA) و Pile Benchmarks به‌دست آورده است. با این حال، در حالی‌که HGRN نسبت به مدل‌های فضای حالت و ترنسفورمرهای سنتی کارایی محاسباتی بالاتری دارد، تنها مدل TNN توانسته است امتیاز پیچیدگی (Perplexity) پایین‌تری نسبت به HGRN کسب کند. به این ترتیب ، HGRN ضمن حفظ کارایی بالا، به‌طور مؤثری شکاف پیچیدگی میان مدل‌های بازگشتی و ترنسفورمرهای مبتنی بر توجه را کاهش می‌دهد.

۴.۳ مدل‌های متفرقه‌ی فضای حالت (**Miscellaneous SSMs**)

۱.۴.۳ مدل ترکیب کارشناسان (**Mixture of Experts - MoE**)

رویکرد Mixture of Experts (MoE) به‌عنوان یکی از روش‌های برجسته برای ارتقای عملکرد مدل‌های زبانی بزرگ (Large Language Models - LLMs) مطرح شده است. تلاش‌های متعددی برای ادغام MoE با مدل‌های فضای حالت (State Space Models - SSMs) صورت گرفته که منجر به نوآوری‌هایی همچون BlackMamba ، MoE-Mamba و Jamba شده است. مدل BlackMamba با رویکردی نوآورانه، سازوکار توجه خودکار (Self-Attention Mechanism) در معماری ترنسفورمر را با مدل‌های فضای حالت Mamba جایگزین کرده است تا فرآیند ترکیب توالی‌ها (Sequence Mixing) را انجام دهد. علاوه بر این، BlackMamba از معماری MoE-Transformer بهره می‌گیرد، که در آن چندین شبکه‌ی پرسپترون چندلایه (Multilayer Per-ceptrons - MLPs) برای ترکیب کانال‌ها (Channel Mixing) مورد استفاده قرار می‌گیرند. مولفه‌ی MoE در BlackMamba تنها یک زیرمجموعه‌ی خلوت از پارامترها را در هر گام پیش‌رو (Forward Pass) فعال می‌کند، که منجر به کاهش قابل‌توجه هزینه‌ی محاسباتی می‌شود. همچنین، این مدل دارای یک سازوکار مسیریاب پویا (Router Mechanism) است که به‌صورت تطبیقی یاد می‌گیرد توکن‌ها را به کارشناس مناسب هدایت کند. نتایج تجربی BlackMamba نشان داده‌اند که این مدل توانسته است شکاف پیچیدگی (Per-plexity Gap) میان ترنسفورمرهای مبتنی بر توجه و مدل‌های فضای حالت را کاهش دهد. مدل MoE-Mamba نیز معماری مشابهی با BlackMamba دارد و ترکیبی از مدل‌های فضای حالت Mamba و ساختار MoE-Transformer را به کار می‌گیرد. این مدل نیز عملکردی رقابتی از خود نشان داده و در کاهش شکاف پیچیدگی با ترنسفورمرها موفق بوده است. در مقابل، مدل Jamba از یک

معماری ترکیبی (Hybrid Architecture) بهره می‌برد که در آن لایه‌های Mamba و ترنسفورمر به‌صورت متناوب قرار گرفته‌اند و سازوکار MoE تنها در برخی از لایه‌ها ادغام شده است. Jamba به‌طور خاص برای وظایف استدلال عقل سلیم (Common Sense Reasoning) طراحی شده و عملکردی رقابتی، حتی برتر از مدل Mixtral-8x7B در برخی معیارها از خود نشان داده است. همچنین، Jamba از طول توکن ورودی تا ۲۵۶هزار ($256K$) پشتیبانی می‌کند، که مقیاس‌پذیری و انعطاف بالای آن را در مدیریت توالی‌های بسیار بلند نشان می‌دهد. این پیشرفت‌ها اثربخشی ادغام رویکرد MoE با مدل‌های فضای حالت را برجسته می‌سازند و مسیر تازه‌ای برای بهبود کارایی، مقیاس‌پذیری و عملکرد مدل‌های زبانی بزرگ فراهم می‌کنند. مدل MambaByte تلاشی نوین در جهت معرفی معماری‌های کارآمد از نظر سخت‌افزاری (Hardware-Efficient Architectures) برای پردازش مؤثر توالی‌ها است. در این مدل، از مدل فضای حالت Mamba State Space Model (SSM) همراه با یک حالت حافظه با اندازه‌ی ثابت (Fixed-Size Memory State) و تکنیک‌های رمزگشایی بهینه (Optimized Decoding Techniques) استفاده می‌شود. منطق طراحی MambaByte بر پایه‌ی این مشاهده استوار است که مدل Mamba دارای یک حالت حافظه‌ی ثابت و بزرگ است که مشابه با حالت پنهان در شبکه‌های عصبی بازگشتی (Recurrent Neural Networks - RNNs) عمل می‌کند. این حالت حافظه مستقل از طول بافت (Context Length) باقی می‌ماند و همین ویژگی امکان پردازش کارآمد توالی‌های بسیار بلند را فراهم می‌سازد. علاوه بر این، MambaByte از یک الگوریتم رمزگشایی پیش‌بینی‌گرانه (Speculative Decoding Algorithm) استفاده می‌کند که به‌طور ویژه برای مدل‌های سطح بایت (Byte-Level Models) طراحی شده است. نتایج تجربی نشان می‌دهند که MambaByte عملکردی برتر نسبت به Mamba دارد، به‌ویژه در پردازش توالی‌های بلند با طول تا $524K$. همچنین، مدل MambaByte در مقایسه با ترنسفورمرهایی مانند MegaByte در مجموعه‌داده‌ی PG19 که یکی از معیارهای باز مدل‌سازی زبان با واژگان آزاد مبتنی بر کتاب‌ها است رقابت‌پذیری بالایی از خود نشان داده است. این پیشرفت‌ها در مدل MambaByte نشان‌دهنده‌ی کارایی بالای آن در دستیابی به پیچیدگی کمتر (Lower Perplexity) و بهبود عملکرد در پردازش توالی‌های بلند است، و به‌طور کلی سهم مهمی در تکامل معماری‌های کارآمد و سخت‌افزارمحور خانواده‌ی Mamba دارد.

۲.۴.۳ مدل‌های فضای حالت و یادگیری درون‌متنی (SSMs and In-Context Learning)

در بررسی مفهوم یادگیری درون‌متنی (In-Context Learning - ICL)، تمرکز اصلی بر رابطه‌ی میان عملکرد موفق در انجام وظایف و اطلاعات موجود در داده‌های آموزشی است. یکی از پرسش‌های اساسی در این زمینه آن است که آیا می‌توان مدل‌ها را به‌گونه‌ای آموزش داد که بتوانند در زمان استنتاج، بدون به‌روزرسانی پارامترها، یادگیری درون‌متنی مؤثری درون کلاس‌های خاصی از توابع برای مثال توابع خطی از خود نشان دهند یا خیر. پژوهش‌های اخیر به بررسی این پرسش پرداخته‌اند، به‌ویژه در ارزیابی توانایی‌های ترنسفورمرهای استاندارد و معماری‌های تخصصی مانند Mamba. به‌عنوان نمونه، Garg et al. نشان داده‌اند که ترنسفورمرهای استاندارد قادرند برای یادگیری درون‌متنی توابع خطی آموزش ببینند، به‌طوری‌که این یادگیری تنها در زمان استنتاج رخ می‌دهد و هیچ به‌روزرسانی پارامتری در طول آموزش انجام نمی‌شود. به‌طور مشابه، Grazzi et al. امکان یادگیری درون‌متنی با استفاده از مدل Mamba را بررسی کرده‌اند. یافته‌های آن‌ها نشان می‌دهد که مدل Mamba عملکردی مشابه ترنسفورمرها در وظایف یادگیری درون‌متنی دارد، به‌ویژه در سناریوهای استانداردی مانند رگرسیون خطی نامتقارن (Skewed Linear Regression)، شبکه‌های عصبی با تابع فعال‌سازی ReLU، و درخت‌های تصمیم (Decision Trees). علاوه بر این، مقاله‌ی مربوط به مدل MambaFormer این بررسی را گسترش داده و نشان می‌دهد که این مدل توانایی بالایی در یادگیری درون‌متنی در وظایف اضافی، از جمله وظایف برداری Vector-Valued MQAR، دارد. جالب آن‌که

MambaFormer در مواردی موفق عمل می‌کند که مدل Mamba به‌تنهایی دچار ضعف است. با این حال، باید توجه داشت که پیچیدگی محاسباتی MambaFormer همچنان درجه‌دوم (Quadratic Complexity) باقی می‌ماند، زیرا در هر لایه سازوکار توجه (Attention Mechanism) با Mamba ادغام شده است. این پژوهش‌ها به‌صورت جمعی دیدگاه‌های مهمی درباره‌ی پتانسیل مدل‌های ترنسفورمر و معماری‌های تخصصی مانند Mamba و MambaFormer در زمینه‌ی یادگیری درون‌متنی ارائه می‌دهند و به درک عمیق‌تری از توانایی‌ها و محدودیت‌های آن‌ها در این حوزه کمک می‌کنند.

Params (M)	↓PPL (test)	↓PPL (val)	Model
			Attn-based
۴۴.۶۵	۲۴.۷۸	۲۴.۴۰	Transformer
۴۲.۱۷	۲۶.۷۰	۲۵.۹۲	FLASH
۴۴.۶۵	۲۸.۰۵	۲۷.۴۴	1+elu
۴۴.۶۵	۶۳.۱۶	۶۲.۵۰	Performer
۴۴.۶۵	۲۷.۰۶	۲۶.۵۳	cosFormer
			MLP-based
۴۶.۷۵	۳۲.۴۳	۳۱.۳۱	Syn(D)
۴۴.۶۵	۳۴.۷۸	۳۳.۶۸	Syn(R)
۴۷.۸۳	۲۹.۱۳	۲۸.۰۸	gMLP
			SSM-based
۴۵.۶۹	۳۹.۶۶	۳۸.۳۴	S4
۴۵.۷۳	۴۱.۰۷	۳۹.۳۹	DSS
۴۳.۸۴	۳۰.۷۴	۲۹.۶۱	GSS
۴۶.۲۳	۲۵.۰۷	۲۴.۳۱	RWKV
۴۶.۲۴	۳۱.۱۲	۲۹.۸۶	LRU
۴۶.۲۵	۲۴.۸۲	۲۴.۱۴	HGRN
۴۸.۶۸	۲۴.۶۷	۲۳.۹۸	TNN

جدول ۱: نتایج به‌دست‌آمده بر روی مجموعه‌داده Wikitext-103، که در آن نماد ↓ به معنی بهتر بودن مقادیر کمتر است.

۴ کاربردهای مدل‌های فضای حالت (Applications of State Space Models)

مدل‌های فضای حالت (SSMs) در ابتدا با هدف پردازش توالی‌های ورودی بلند، در مقایسه با مدل‌های Transformer پیشنهاد شدند. ازاین‌رو، این مدل‌ها در حوزه‌های گوناگونی که نیاز به پردازش توالی‌های بلند دارند، کاربردهای گسترده‌ای پیدا کرده‌اند.

۱.۴ حوزه‌ی زبان (Language Domain - Long Sequence)

در حوزه‌ی پردازش زبان طبیعی (Natural Language Processing - NLP)، مدل‌های Transformer به‌طور سنتی انتخاب اصلی برای مدل‌سازی داده‌های متنی بوده‌اند، زیرا به‌خوبی قادرند وابستگی‌های پیچیده را از طریق مکانیزم توجه (Attention Mechanism)

به‌دست آورند. بااین‌حال، کارایی این مدل‌ها به دلیل پیچیدگی درجه‌دوم $O(N^2)$ کاهش می‌یابد، به‌ویژه هنگام پردازش توالی‌های طولانی. با افزایش طول توالی، نیاز به حافظه و توان محاسباتی نیز افزایش می‌یابد و این موضوع باعث می‌شود آموزش مدل بر روی ورودی‌های بلند، از نظر منابع محاسباتی غیرعملی شود. برای رفع این ناکارآمدی‌ها، مجموعه‌ای از مدل‌های فضای حالت (State Space Models - SSMs) معرفی شده‌اند، از جمله: S4, S4nd, S5, HiPPO, Hyena, H3, LDStack, Liquid-S4, DSS, GSS, Mega, LRU, HGRN, TNN, Mamba و سایر مدل‌های مشابه. برخلاف Transformerها که بر پایه‌ی مکانیزم توجه عمل می‌کنند، مدل‌های SSM داده‌های ورودی را در یک فضای پنهان با اندازه‌ی ثابت (Fixed-size Latent State) فشرده می‌کنند. این حافظه‌ی ایستا در طول تولید توالی ثابت باقی می‌ماند و همین امر باعث می‌شود SSMها در پردازش ورودی‌های بلند، کاراتر باشند. بااین‌حال، یک مصالحه (Trade-off) وجود دارد: در حالی‌که مدل‌های SSM از نظر کارایی برتری دارند، قابلیت بازیابی یا کپی بخش‌هایی از زمینه‌ی ورودی را از دست می‌دهند؛ ویژگی‌ای که برای وظایفی مانند یادگیری با داده‌ی کم (Few-shot Learning) و بازیابی اطلاعات حیاتی است. در مقابل، مدل‌های Transformer در چنین وظایفی عملکرد بسیار بهتری دارند. در تحلیل‌های عملکردی این پژوهش، امتیازهای GLUE در جدول ۲، نتایج معیار Wikitext در جدول ۱ و معیار Pile در جدول ۱.۴ برای داده‌های متنی با توالی بلند، بین مدل‌های Transformer و SSM مقایسه شده‌اند. بحث میان این دو رویکرد همچنان ادامه دارد، زیرا هرکدام دارای نقاط قوت و محدودیت‌های خاص خود در حوزه‌ی NLP هستند.

Params (M)	AVG	CoLA	MRPC	SST-2	QQP	QNLI	MNLI	Method
								Attn-based
۱۲۴.۷۰	۷۸.۷۹	۳۸.۶۳	۸۸.۳۵	۹۰.۲۵	۸۸.۰۴	۸۷.۷۹	۷۹.۳۷ / ۷۹.۰۷	Transr
۱۲۸.۲۸	۷۷.۰۱	۴۰.۶۵	۸۲.۶۵	۹۰.۲۵	۸۶.۸۵	۸۴.۸۶	۷۷.۰۱ / ۷۶.۷۸	LS
۱۲۷.۱۲	۷۶.۸۷	۲۹.۴۰	۸۲.۵۰	۹۰.۷۱	۸۸.۸۳	۸۷.۱۰	۷۹.۴۵ / ۸۰.۰۸	FLASH
۱۲۴.۷۰	۷۰.۰۰	–	۸۳.۰۳	۸۷.۲۷	۸۶.۹۰	۸۲.۵۹	۷۴.۸۷ / ۷۵.۳۷	1+elu
۱۲۴.۷۰	۶۳.۴۱	۱۹.۴۱	۸۲.۱۱	۸۱.۴۲	۷۹.۱۰	۶۳.۴۴	۵۸.۸۵ / ۵۹.۵۲	Performer
۱۲۴.۷۰	۷۴.۸۸	۳۳.۰۳	۸۱.۹۳	۸۹.۴۵	۸۶.۱۲	۸۲.۶۱	۷۵.۱۰ / ۷۵.۹۵	cosFormer
								MLP-based
۱۳۱.۰۰	۵۸.۶۰	–	۸۱.۷۹	۸۲.۳۴	۸۱.۳۳	۶۲.۸۰	۵۰.۹۳ / ۵۱.۰۲	Syn(D)
۱۲۹.۴۲	۵۹.۰۸	۴.۶۳	۸۱.۳۸	۸۲.۲۲	۷۸.۱۱	۶۲.۲۹	۵۲.۸۲ / ۵۲.۱۳	Syn(R)
۱۳۱.۰۸	۷۴.۶۵	۳۶.۰۶	۸۲.۳۰	۹۰.۲۵	۸۶.۴۸	۸۰.۵۶	۷۳.۳۰ / ۷۳.۶۰	gMLP
								FFT-based
۱۲۴.۷۰	۶۳.۵۳	–	۸۲.۹۱	۸۱.۸۸	۷۹.۴۳	۷۳.۳۱	۶۲.۴۵ / ۶۴.۷۱	FNet
۱۳۰.۰۶	۶۵.۱۹	۹.۶۲	۸۲.۴۴	۸۴.۴۰	۸۰.۲۵	۶۵.۴۲	۶۶.۷۵ / ۶۷.۴۵	GFNet
۱۲۱.۵۷	۷۱.۹۷	۳۶.۱۹	۸۲.۳۵	۸۸.۸۸	۸۵.۱۲	۷۳.۲۰	۶۸.۷۹ / ۶۹.۲۸	AFNOt
								SSM-based
۱۳۱.۷۹	۶۹.۵۸	۲۳.۰۱	۸۳.۳۶	۸۷.۰۴	۸۴.۶۱	۷۲.۱۴	۶۸.۴۵ / ۶۸.۴۲	S4
۱۲۳.۷۶	۴۸.۴۵	۶.۱۴	۸۰.۹۵	۶۵.۳۷	۶۵.۱۸	۵۰.۸۰	۳۵.۴۶ / ۳۵.۲۲	DSS
۱۲۲.۸۰	۶۰.۰۰	۶.۵۶	۸۲.۱۱	۸۵.۶۷	۸۰.۹۸	۶۲.۵۸	۵۰.۵۳ / ۵۱.۵۸	GSS
۱۲۶.۴۰	۷۸.۵۱	۴۹.۸۵	۸۲.۹۶	۹۰.۶۰	۸۸.۳۰	۸۵.۰۶	۷۶.۷۲ / ۷۶.۰۶	TNN

جدول ۲: مقایسه‌ی عملکرد مدل‌های مختلف در مدل‌سازی توالی بر روی مجموعه‌داده‌ی GLUE. نتایج MNLI به‌صورت جداگانه برای دو بخش match و mismatch گزارش شده‌اند. عملکرد MRPC با استفاده از نمره‌ی F1 اندازه‌گیری شده است، CoLA با ضریب همبستگی Matthews سنجیده شده، و سایر وظایف با معیار دقت (Accuracy) ارزیابی شده‌اند. بهترین نتیجه با بولد و دومین نتیجه‌ی برتر با زیرخط مشخص شده‌اند. نماد «–» نشان‌دهنده‌ی مقادیر ناموجود است. عبارت Attn به مدل‌های توجه، SSM به مدل‌های فضای حالت، Trans به مدل‌های ترنسفورمر و LS به مدل‌های Transformer-LS اشاره دارد.

با توجه به این ناکارآمدی‌ها، مجموعه‌ای از مدل‌های فضای حالت (State Space Models - SSMs) پدیدار شده‌اند که شامل مدل‌هایی همچون S4، S4nd، S5، HiPPO، Hyena، H3، LDStack، Liquid-S4، DSS، GSS، Mega، LRU، HGRN، TNN و Mamba می‌باشند. برخلاف مدل‌های Transformer که متکی بر مکانیزم توجه (Attention Mechanism) هستند، مدل‌های SSM داده‌های ورودی را در یک فضای پنهان با اندازه‌ی ثابت (Fixed-size Latent State) فشرده می‌کنند. این تخصیص حافظه‌ی ایستا در طول فرآیند تولید توالی ثابت باقی می‌ماند و همین ویژگی موجب افزایش کارایی SSMها در پردازش ورودی‌های بلند می‌شود. با این حال، نوعی مصالحه (Trade-off) وجود دارد: در حالی که مدل‌های SSM از نظر کارایی برتری دارند، قابلیت بازیابی و کپی بخش‌هایی از زمینه‌ی ورودی را از دست می‌دهند؛ قابلیتی که برای وظایفی نظیر یادگیری با داده‌ی اندک (Few-shot Learning) و بازیابی اطلاعات (Retrieval) ضروری است. در مقابل، مدل‌های Transformer در این حوزه‌ها عملکرد بسیار بهتری دارند. در تحلیل عملکرد، امتیازهای GLUE در جدول ۲، نتایج معیار Wikitext در جدول ۱ و معیار Pile در جدول ۱.۴ برای داده‌های متنی بلند بین مدل‌های Transformer و SSM مورد مقایسه قرار گرفته‌اند. بحث میان این دو رویکرد همچنان ادامه دارد، زیرا هر یک از آن‌ها در حوزه‌ی NLP دارای نقاط قوت و محدودیت‌های منحصربه‌فردی هستند.

۲.۴ حوزه‌ی بینایی ماشین (Vision Domain)

مدل‌های فضای حالت (State Space Models - SSMs) کارایی قابل‌توجهی را در حوزه‌ی بینایی ماشین (Computer Vision) در وظایفی مانند طبقه‌بندی تصویر (Image Classification)، بخش‌بندی (Segmentation) و تشخیص اشیاء (Object Detection) از خود نشان داده‌اند. تعدادی از پژوهش‌ها از جمله SiMBA، V-Mamba، Vim، localMamba، plainMamba و Vim2 مدل‌های SSM را برای وظایف طبقه‌بندی تصویر تطبیق داده‌اند. به‌طور مشابه، مدل‌هایی مانند U-Mamba، SegMamba، Swin U-Mamba، P-Mamba (برای بخش‌بندی بطنی) و VM-UNet برای وظایف بخش‌بندی در بینایی ماشین معرفی شده‌اند. تطبیق‌های خاص بینایی از معماری Mamba مانند Vision Mamba، V-Mamba و SiMBA از مدل‌های فضای حالت دیداری و دوجهته (Bidirectional and Visual State Space Models) برای انجام وظایف بینایی ماشین استفاده می‌کنند. با این حال، شکاف عملکردی میان این مدل‌ها و مدل‌های ترنسفورمر پیشرفته (State-of-the-Art Transformers) مانند SpectFormer، SVT، WaveViT، Volo و SCT وجود دارد. مدل SiMBA تلاش می‌کند این شکاف را با جایگزینی شبکه‌های توجه با Mamba برای ترکیب توکن‌ها (Token Mixing) و استفاده از Einstein FFT (EinFFT) برای ترکیب کانال‌ها (Channel Mixing) کاهش دهد. SiMBA روش جدیدی به نام Einstein Blending را برای ترکیب کانال‌ها معرفی می‌کند، که برخلاف بسیاری از مدل‌ها، نیازی به داشتن ابعاد مربعی کامل برای طول توالی و ابعاد کانال ندارد. همچنین، SiMBA از نسخه‌ی هرمی معماری ترنسفورمر (Pyramid Transformer Architecture) استفاده می‌کند، که منجر به بهبود قابل‌توجه عملکرد، به‌ویژه در مجموعه‌داده‌هایی مانند ImageNet و وظایف سری‌های زمانی (Time Series Tasks) می‌شود. با وجود این پیشرفت‌ها، هنوز شکاف عملکردی میان SiMBA و مدل‌های ترنسفورمر پیشرفته‌ای مانند SCT باقی مانده است. جدول ۵ نتایج مربوط به وظیفه‌ی طبقه‌بندی تصویر در مجموعه‌داده‌ی ImageNet را خلاصه می‌کند و نشان می‌دهد که SiMBA به‌عنوان بهترین مدل در میان معماری‌های فضای حالت برای وظایف تشخیص تصویر عمل می‌کند. از سوی دیگر، مدل RSMamba بر روی طبقه‌بندی تصاویر سنجش از دور (Remote Sensing Image Classification) تمرکز دارد، در حالی‌که Res-VMamba برای طبقه‌بندی دیداری دقیق دسته‌بندی مواد غذایی (Fine-grained Food Category Visual Classification) با استفاده از مدل‌های فضای حالت انتخابی و یادگیری عمیق باقیمانده (Deep Residual Learning) طراحی شده است. افزون بر این، تلاش‌هایی مانند SiMBA، V-Mamba و Vim عملکرد امیدوارکننده‌ای را در وظایف تشخیص اشیاء (Object Detection) نشان داده‌اند. مدل U-Mamba بر بهبود وابستگی‌های بلندمدت (Long-range Dependencies) در وظایف بخش‌بندی تصاویر زیست‌پزشکی (Biomedical Image Segmentation) تمرکز دارد. با بهره‌گیری از مدل‌های فضای حالت (State Space Models)، U-Mamba قادر است روابط ظریف و دقیق در تصاویر زیست‌پزشکی را به‌خوبی شناسایی و مدل‌سازی کند و به‌ویژه چالش‌های مربوط به مدل‌سازی دنباله‌ای بلندمدت را برطرف سازد. به‌طور مشابه، مدل SegMamba قابلیت‌های معماری Mamba را برای وظایف بخش‌بندی تصاویر پزشکی سه‌بعدی (3D Medical Image Segmentation) گسترش می‌دهد. با ترکیب معماری Mamba، مدل SegMamba قادر است وابستگی‌های متوالی را در داده‌های تصویری حجمی (Volumetric Medical Image Data) به‌طور مؤثر مدل‌سازی کند. مدل MambaMorph به مسئله‌ی ثبت تغییرپذیر تصاویر MR-CT (Deformable MR-CT Registration) می‌پردازد و این کار را از طریق ادغام یادگیری ویژگی متقابل (Contrastive Feature Learning) در چارچوب Mamba انجام می‌دهد. این رویکرد امکان ثبت دقیق‌تر و مقاوم‌تر تصاویر پزشکی را با استفاده از توانایی‌های مدل‌های فضای حالت فراهم می‌سازد. مدل VM-UNet از

ترکیب Vision Mamba UNet برای بخش‌بندی تصاویر پزشکی استفاده می‌کند. با ترکیب معماری Mamba با UNet، VM-UNet به عملکردی برتر در بخش‌بندی تصاویر پزشکی دست می‌یابد و از نقاط قوت هر دو معماری بهره‌مند می‌شود. مدل nnMamba مدل‌های فضای حالت را به حوزه‌های مختلف تحلیل تصاویر زیست‌پزشکی (Biomedical Image Analysis) از جمله بخش‌بندی تصاویر سه‌بعدی، طبقه‌بندی (Classification) و شناسایی نقاط شاخص (Landmark Detection) گسترش می‌دهد. با به‌کارگیری معماری Mamba، مدل nnMamba عملکرد و کارایی بهتری را در پردازش داده‌های پیچیده‌ی تصویری پزشکی نشان می‌دهد. مدل FD-Vision از معماری Mamba برای اصلاح نوردهی در تصاویر آندوسکوپی (Endoscopic Exposure Correction) بهره می‌برد و موجب افزایش پایداری در تحلیل تصاویر پزشکی می‌شود. مدل Weak-MambaUNet با ترکیب Mamba با معماری‌های CNN و ViT وظیفه‌ی بخش‌بندی تصاویر پزشکی مبتنی بر حاشیه‌گذاری ضعیف (Scribble-based Segmentation) را بهبود می‌دهد. این ترکیب امکان بهره‌برداری مؤثرتر از سیگنال‌های نظارت ضعیف (Weak Supervision Signals) را فراهم می‌کند و دقت بخش‌بندی را افزایش می‌دهد. مدل MedMamba اثربخشی معماری Mamba را به‌طور خاص در وظایف طبقه‌بندی تصاویر پزشکی (Medical Image Classification) بررسی کرده و توانایی آن را در پردازش داده‌های متنوع تصویربرداری پزشکی نشان می‌دهد. در نهایت، مدل LightM-UNet با استفاده از نسخه‌ی سبک‌تر معماری Mamba بخش‌بندی تصاویر پزشکی را با تأکید بر کارایی و سرعت بهینه می‌کند. مدل Large Window-based Mamba UNet نیز با بهره‌گیری از پنجره‌های بزرگ‌تر و روش‌های نوآورانه فراتر از مکانیزم‌های سنتی هم‌نهشتی (Convolutional) و خودتوجهی (Self-Attention)، عملکرد بخش‌بندی تصاویر پزشکی را به‌طور قابل‌توجهی بهبود می‌بخشد. مدل H-vMUNet یک معماری پیشرفته از نوع High-Order Vision Mamba UNet را برای وظایف بخش‌بندی تصاویر پزشکی (Medical Image Segmentation) معرفی می‌کند. این مدل با بهره‌گیری از وابستگی‌های مرتبه بالا (High-order Dependencies) و معماری Mamba، دقت و پایداری بخش‌بندی را به شکل چشمگیری بهبود می‌بخشد. مدل ProMamba در زمینه‌ی بخش‌بندی پولیپ (Polyp Segmentation) تخصص دارد و با به‌کارگیری تکنیک‌های مبتنی بر پرامپت (Prompt-based Techniques) در چارچوب Mamba به نتایج پیشرفته و قابل‌توجهی در این وظیفه‌ی خاص تصویربرداری پزشکی دست یافته است. مدل CMViM از خودرمزگذارهای ماسک‌شده‌ی متضاد (Contrastive Masked Vim Autoencoders) برای یادگیری بازنمایی چندوجهی سه‌بعدی (3D Multi-modal Representation Learning) در طبقه‌بندی بیماری آلزایمر (Alzheimer's Disease Classification) استفاده می‌کند. با تکیه بر معماری Mamba، این مدل توانایی یادگیری بازنمایی بهتری از داده‌های چندوجهی و در نتیجه بهبود دقت در طبقه‌بندی را نشان داده است. مدل Gamba ترکیبی از تکنیک پاشش گاوسی (Gaussian Splatting) و معماری Mamba برای بازسازی سه‌بعدی از نمای منفرد (Single-view 3D Reconstruction) ارائه می‌دهد. این ترکیب با بهره‌گیری از مزایای هر دو روش، به بازسازی دقیق‌تر و مقاوم‌تر منجر می‌شود. مدل ReMamber بر وظایف بخش‌بندی تصویر تمرکز دارد و از ماژول Mamba Twister برای بهبود عملکرد استفاده می‌کند، به‌طوری‌که نتایج بهتری نسبت به روش‌های سنتی در بخش‌بندی به‌دست می‌آورد. مدل MambaIR به‌عنوان یک خط مبنا (Baseline) ساده برای وظایف بازسازی تصویر (Image Restoration) عمل می‌کند و از مدل‌های فضای حالت بهره می‌برد. مدل T-Mamba با هدف بهبود دقت بخش‌بندی دندان‌ها در تصاویر سه‌بعدی (3D Tooth Segmentation) طراحی شده است. این مدل با ترکیب ویژگی‌های مبتنی بر فرکانس (Frequency-based Features) و وابستگی‌های طولانی‌مدت دروازه‌دار (Gated Long-range Dependencies) درون معماری Vision Mamba، دقت بخش‌بندی را به‌ویژه در شرایط دشوار

تصویربرداری مانند نویز، کنتراست پایین و وجود آرتیفکت‌ها افزایش می‌دهد. هرچند شبکه‌های عصبی کانولوشنی (CNNs) و ترنسفورمرها در بخش‌بندی تصاویر کاربرد گسترده‌ای دارند، اما در مدیریت وابستگی‌های بلندمدت به دلیل محدودیت‌های محلی یا پیچیدگی محاسباتی ضعف دارند. در بخش‌بندی سه‌بعدی دندان‌ها که برای تشخیص ارتودنسی حیاتی است چالش‌هایی مانند نویز، کنتراست پایین و آرتیفکت‌های موجود در تصاویر CBCT فرآیند را دشوار می‌کند. مدل T-Mamba با ادغام کدگذاری موقعیت اشتراکی (Shared Positional Encoding) و ویژگی‌های مبتنی بر فرکانس در معماری Vision Mamba به حفظ موقعیت‌های مکانی و بهبود ویژگی‌ها در حوزه‌ی فرکانس کمک می‌کند. این مدل از واحد انتخاب دروازه‌ای (Gate Selection Unit) استفاده می‌کند که دو ویژگی حوزه‌ی مکانی و یک ویژگی حوزه‌ی فرکانس را به‌صورت تطبیقی ترکیب می‌کند. T-Mamba نتایج پیشرفته‌ای در مجموعه‌داده‌های عمومی Tooth CBCT ارائه کرده و در شاخص‌های مختلف ارزیابی مانند IoU، SO، DSC، HD و ASSD به‌طور قابل‌توجهی از روش‌های پیشین عملکرد بهتری نشان داده است.

Avg.	Path-X (۱۶۳۸۴)	Pathfinder (۱۰۲۴)	Image (۱۰۲۴)	Retrieval (۴۰۰۰)	Text (۴۰۹۶)	ListOps (۲۰۴۸)	Model (Input length)
۵۳.۶۶	✗	۷۱.۴۰	۴۲.۴۴	۵۷.۴۶	۶۴.۲۷	۳۷.۳۶	Transformer
۴۶.۷۱	✗	۶۶.۶۳	۴۱.۴۶	۵۳.۳۹	۵۲.۹۸	۱۵.۸۲	Local Attention
۵۱.۰۳	✗	۷۱.۷۱	۴۴.۲۴	۵۹.۵۹	۶۳.۵۸	۱۷.۰۷	Sparse Trans.
۵۲.۸۸	✗	۶۹.۷۱	۴۲.۲۲	۵۶.۸۹	۶۲.۸۵	۳۵.۶۳	Longformer
۵۱.۱۴	✗	۷۶.۳۴	۳۸.۵۶	۵۲.۲۷	۵۳.۹۴	۳۵.۷۰	Linformer
۵۰.۵۶	✗	۶۸.۵۰	۳۸.۰۷	۵۳.۴۰	۵۶.۱۰	۳۷.۲۷	Reformer
۵۱.۲۳	✗	۶۷.۴۵	۴۱.۲۳	۵۳.۸۳	۶۱.۲۰	۳۳.۶۷	Sinkhorn Transformer
۵۲.۴۰	✗	۶۹.۴۵	۴۱.۶۱	۵۴.۶۷	۶۱.۶۸	۳۶.۹۹	Synthesizer
۵۴.۱۷	✗	۷۴.۸۷	۴۰.۸۳	۵۹.۲۹	۶۴.۰۲	۳۶.۰۵	BigBird
۵۰.۴۶	✗	۷۵.۳۰	۴۲.۳۴	۵۳.۰۹	۶۵.۹۰	۱۶.۱۳	Linear Trans.
۵۱.۸۱	✗	۷۷.۰۵	۴۲.۷۷	۵۳.۸۲	۶۵.۴۰	۱۸.۰۱	Performer
۵۱.۷۶	–	۷۱.۹۶	۵۱.۲۳	۸۳.۱۵	۶۷.۷۰	۳۶.۵۰	cosFormer
۵۱.۰۹	–	۷۰.۲۵	۴۷.۴۰	۸۶.۱۰	۶۴.۱۰	۳۸.۷۰	FLASH
۵۴.۴۲	✗	۷۷.۸۰	۳۸.۶۷	۵۹.۶۱	۶۵.۱۱	۳۵.۳۳	FNet
۵۷.۴۶	✗	۷۰.۹۴	۴۱.۵۸	۷۹.۵۶	۶۵.۵۲	۳۷.۱۵	Nyströmformer
۵۹.۳۷	✗	۷۷.۷۲	۴۷.۳۸	۷۹.۲۹	۶۴.۵۷	۳۷.۲۵	Luna-256
۶۱.۴۱	✗	۶۸.۷۸	۴۶.۰۵	۶۳.۹۹	۷۸.۶۹	۴۹.۵۳	H-Transformer-1D
۶۸.۰۲	✗	۹۱.۵۱	۸۸.۹۰	✗	۸۴.۰۸	۴۳.۶۰	CCNN
۸۰.۴۸	۸۸.۱۰	۸۶.۰۵	۸۷.۲۶	۸۷.۰۹	۷۶.۰۲	۵۸.۳۵	S4
۸۱.۱۸	۸۵.۶۰	۸۴.۷۰	۸۴.۹۰	۸۷.۶۰	۸۴.۶۰	۵۹.۷۰	DSSEXP
۸۱.۸۸	۸۷.۸۰	۸۴.۶۰	۸۵.۷۰	۸۷.۸۰	۸۴.۸۰	۶۰.۶۰	DSSSOFTMAX
۸۴.۸۹	۹۱.۹۵	۹۳.۰۶	۸۸.۱۹	۸۹.۴۶	۸۶.۱۸	۶۰.۴۷	S4D-LegS
۸۵.۶۶	۹۳.۸۱	۹۴.۴۱	۸۵.۸۰	۹۰.۹۷	۹۰.۱۹	۵۸.۷۶	Mega-chunk($\mathcal{O}(L)$)
۸۶.۰۹	۹۶.۳۵	۹۴.۲۰	۸۸.۶۵	۹۰.۹۰	۸۶.۸۲	۵۹.۶۰	TNN
۸۶.۲۱	۹۶.۱۰	۹۳.۰۰	۸۸.۲۴	۹۰.۹۷	۸۷.۹۰	۶۱.۰۴	LRU
۸۶.۹۱	۹۷.۵۰	۹۲.۹۲	۸۸.۶۹	۹۴.۲۳	۸۸.۱۴	۵۹.۹۵	HGRN
۸۷.۱۷	۹۷.۸۳	۹۵.۴۶	۸۷.۹۷	۹۱.۱۱	۸۹.۲	۶۱.۴۵	SGConv
۸۷.۳۲	۹۶.۶۶	۹۴.۸	۸۹.۵۰	۹۱.۲۰	۸۹.۰۲	۶۲.۷۵	Liquid-S4
۸۷.۴۶	۹۸.۵۸	۹۵.۳۳	۸۸.۰۰	۹۱.۴۰	۸۹.۳۱	۶۲.۱۵	S5
۸۸.۲۱	۹۷.۹۸	۹۶.۰۱	۹۰.۴۴	۹۱.۲۵	۹۰.۴۳	۶۳.۱۴	Mega ($\mathcal{O}(L^2)$)

جدول ۳: دقت آزمون در وظایف معیار LRA. علامت ✗ نشان می‌دهد که عملکرد مدل از حد تصادفی فراتر نرفته است. ارجاعات به مدل اصلی اشاره دارند. نتایج مدل‌هایی از Transformer تا Performer از مقاله‌ی Tay et al. (2020) استخراج شده‌اند. این جدول با استفاده از داده‌های مقاله‌ی HGRN نوشته Qin et al. (2023) و مقاله‌ی S5 نوشته Smith et al. (2022) گردآوری شده است، به‌گونه‌ای که نتایج در قالبی یکپارچه ارائه می‌شوند.

B۱۰۰	B۱۵	B۱۰	B۵	Model
	۱۱.۲	۱۱.۹	۱۳.۳	GPT (125M)
	۱۱.۱	۱۱.۸	۱۳.۳	Hyena-2 (153M)
	۹.۱	۹.۸	۱۱.۴	GPT (355M)
	۹.۲	۹.۸	۱۱.۳	Hyena-2 (355M)
۴.۵۶	–	–	–	Transformer (1000M)
۵.۰۷	–	–	–	LRU (1000M)
۴.۱۴	–	–	–	HGRN (1000M)

جدول ۴: نتایج بر روی مجموعه‌داده Pile: اندازه مدل‌ها و امتیازات Perplexity. جدول زیر مدل‌های زبانی مختلفی را نشان می‌دهد که بر روی تعداد متفاوتی از توکن‌ها (از ۵ میلیارد تا ۱۰۰ میلیارد) در مجموعه‌داده Pile آموزش داده شده‌اند. مقادیر پایین‌تر Perplexity (PPL) نشان‌دهنده عملکرد بهتر در وظایف مدل‌سازی زبان هستند. این نتایج از مقالات HGRN و Hyena Hierarchy اقتباس شده‌اند.

Top-1 Acc.	FLOPs	#Params.	Image Size	Method
		Convnets		
۷۷.۴	–	۴۵ M	224^2	ResNet-101
۸۱.۷	۸.۰ G	۳۹ M	224^2	RegNetY-8G
۷۸.۳	–	۶۰ M	224^2	ResNet-152
۸۲.۹	۱۶.۰ G	۸۴ M	224^2	RegNetY-16G
		Transformers		
۷۹.۸	۴.۶ G	۲۲M	224^2	DeiT-S
۸۱.۳	۴.۵ G	۲۹ M	224^2	Swin-T
۸۲.۹	۴.۲ G	۱۹ M	224^2	EffNet-B4
۸۲.۹	۴.۱ G	۲۲.۷ M	224^2	WaveViT-H-S*
۸۴.۳	۳.۹ G	۲۲.۲ M	224^2	SpectFormer-H-S
۸۴.۲	۳.۹ G	۲۲ M	224^2	SVT-H-S
۸۴.۵	۴.۱ G	۲۱.۷ M	224^2	SCT-H-S
۸۳.۶	۹.۹G	۳۰ M	456^2	EffNet-B5
۸۳.۰	۸.۷ G	۵۰ M	224^2	Swin-S
۸۴.۵	۹.۳ G	۴۵ M	224^2	CMT-B
۸۴.۵	۱۱.۷ G	۶۹ M	224^2	MaxViT-S
۸۴.۶	۹.۴ G	۴۸ M	224^2	iFormer-B*
۸۴.۸	۷.۲ G	۳۳ M	224^2	Wave-Vit-B*
۸۵.۱	۶.۳ G	۳۳.۱ M	224^2	SpectFormer-H-B*
۸۵.۲	۶.۳ G	۳۲.۸ M	224^2	SVT-H-B*
۸۵.۲	۶.۵ G	۳۲.۵ M	224^2	SCT-H-B*
۷۹.۵	–	۴۵ M	224^2	M2-ViT-b
۸۱.۸	۱۷.۵ G	۸۶ M	224^2	DeiT-B
۸۳.۵	۱۵.۴ G	۸۸ M	224^2	Swin-B
۸۳.۵	–	۵۰ M	224^2	M2-Swin-B
۸۴.۰	۱۹.۰ G	۴۳ M	224^2	EffNet-B6
۸۵.۰	۲۳.۴ G	۱۲۰ M	224^2	MaxViT-B
۸۵.۴	۲۰.۶ G	۸۶ M	224^2	VOLO-D3*
۸۵.۵	۱۴.۸ G	۵۷ M	224^2	Wave-ViT-L*
۸۵.۷	۱۲.۷ G	۵۴.۷ M	224^2	SpectFormer-H-L*
۸۵.۷	۱۲.۷ G	۵۴.۰ M	224^2	SVT-H-L*
۸۵.۹	۱۳.۴ G	۵۴.۱ M	224^2	SCT-H-L*
		SSMs		
۷۶.۱	–	۷ M	224^2	Vim-Ti
۷۷.۹	۳.۰ G	۷ M	224^2	PlainMamba-L1
۸۲.۲	۵.۶ G	۲۲ M	224^2	Vmamba-T
۸۱.۱	۳.۶ G	۱۸.۵ M	224^2	SiMBA-S(Monarch)
۸۱.۷	–	۲۴ M	224^2	Mamba-2D-S
۸۱.۷	۲.۴ G	۱۵.۳ M	224^2	siMBA-S(EinFFT)
۸۲.۷	۵.۷ G	۲۶ M	224^2	LocalVMamba-T
۸۲.۷	–	۲۰ M	224^2	Vim2-T
۸۴.۰	۵.۰ G	۲۶.۵ M	224^2	SiMBA-S(MLP)
۸۰.۵	–	۲۶ M	224^2	Vim-S
۸۱.۶	۸.۱ G	۲۵ M	224^2	PlainMamba-L2
۸۲.۶	۶.۳ G	۲۶.۹ M	224^2	SiMBA-B(Monarch)
۸۳.۰	–	۹۲ M	224^2	Mamba-2D-B
۸۳.۵	۵.۲ G	۲۲.۸ M	224^2	SiMBA-B(EinFFT)
۸۳.۵	۱۱.۲ G	۴۴ M	224^2	VMamba-S
۸۳.۷	۱۱.۴ G	۵۰ M	224^2	Local VMamba-S
۸۳.۷	–	۴۳ M	224^2	Vim2-S
۸۴.۷	۹.۰ G	۴۰.۰ M	224^2	SiMBA-B(MLP)
۷۸.۵	–	۸۸ M	224^2	Hyena Vit-B
۸۰.۴	–	۸۹ M	224^2	S4ND-ViT-B
۸۲.۳	۱۴.۴ G	۵۰ M	224^2	PlainMamba-L3
۸۳.۲	۱۸.۰ G	۷۵ M	224^2	VMamba-B
۸۳.۸	۱۰.۷ G	۴۲ M	224^2	SiMBA-L(Monarch)
۸۳.۹	–	۷۴ M	224^2	Vim2-B
۸۴.۴	۹.۶ G	۳۶.۶ M	224^2	SimBA-L(EinFFT)

جدول ۵: نتایج SOTA بر روی مجموعه‌داده ImageNet-1K. این جدول عملکرد مدل‌های مختلف بینایی ماشین را بر روی مجموعه‌داده ImageNet-1K برای وظایف تشخیص تصویر نشان می‌دهد. علامت * نشان می‌دهد که مدل با استفاده از روش Token Labeling برای رمزگذاری پچ‌ها آموزش اضافی دیده است. مدل‌های بینایی بر اساس مقدار GFLOPs به سه دسته تقسیم شده‌اند: کوچک (Small)، پایه (Base) و بزرگ (Large). بازه‌های GFLOPs به ترتیب عبارت‌اند از: کوچک (کمتر از ۵)، پایه (بین ۵ تا ۱۰)، و بزرگ (بین ۱۰ تا ۳۰). این جدول از مقاله SiMBA اقتباس شده است.

HyenaDNA (1.6M)	NT (2.5B, 850 genomes)	NT (2.5B)	NT (500M)	Model
M۶.۱	B۵.۲	B۵.۲	M۵۰۰	**Params**
۱	۸۵۰	۳،۲۰۲	۱	**of Genomes**
۶۲.۶	۵۸.۰	۵۹.۳	۵۳.۵	Enhancer
۵۵.۷	۴۷.۴	۵۰.۰	۴۸.۵	Enhancer types
۸۱.۷	۸۱.۴	۷۷.۶	۷۳.۷	H3
۵۷.۱	۵۵.۹	۴۴.۵	۳۵.۸	H3K4me1
۵۳.۹	۳۲.۶	۳۰.۰	۲۸.۱	H3K4me2
۶۱.۲	۴۲.۱	۲۸.۱	۲۶.۳	H3K4me3
۶۵.۱	۵۷.۵	۵۰.۸	۴۶.۲	H3K9ac
۶۶.۳	۵۵.۰	۴۷.۱	۳۷.۷	H3K14ac
۶۵.۳	۶۳.۲	۵۳.۳	۴۶.۷	H3K36me3
۷۱.۶	۶۴.۲	۵۹.۲	۵۷.۷	H3K79me3
۷۹.۶	۸۲.۲	۷۸.۹	۷۶.۲	H4
۶۳.۷	۵۰.۱	۴۲.۳	۳۴.۴	H4ac
۹۶.۵	۹۷.۴	۹۶.۶	۹۵.۴	Promoter all
۹۶.۶	۹۷.۷	۹۶.۹	۹۵.۶	Promoter non-TATA
۹۶.۷	۹۶.۴	۹۵.۸	۹۴.۸	Promoter TATA
۹۶.۶	۹۹.۰	۹۸.۵	۹۶.۵	Splice acceptor
۹۷.۳	۹۸.۴	۹۸.۲	۹۷.۲	Splice donor
۹۷.۹	۹۸.۳	۹۷.۸	۹۷.۲	Splice all

جدول ۶: ارزیابی مدل Nucleotide Transformer (NT). برای مجموعه‌داده‌های Enhancer و Epigenetic Marks، معیار عملکرد از ضریب همبستگی متیو (Matthews Correlation Coefficient – MCC) استفاده شده است. برای مجموعه‌داده‌های Promoter و Splice Site، معیار عملکرد از امتیاز F1-Score بهره گرفته شده است.

۳.۴ حوزه های پزشکی

مدل‌های مبتنی بر Mamba در خط مقدم تحقیقات پیشرفته در حوزه‌های ژنومیک و طراحی دارو قرار دارند. در حوزه زیست‌شناسی، پژوهشگران از Mamba برای تحلیل توالی‌های ژنومی، رمزگشایی تنوع‌های ژنتیکی و درک بیماری‌های ارثی استفاده می‌کنند. در حالی‌که مدل‌های مبتنی بر Nucleotide Transformers و ترنسفورمرهای BERT پیش‌تر مورد بررسی قرار گرفته‌اند، مدل‌های مبتنی بر فضای حالت (SSMs) در مدل‌سازی توالی‌های ژنومی نتایج امیدوارکننده‌ای نشان داده‌اند. در حوزه شیمی، Mamba نقش مهمی در اکتشاف مولکولی و طراحی داروهای جدید ایفا می‌کند. مدل زبانی شیمیایی (Chemical Language Model – CLM) قادر است مولکول‌های متنوع و زیست‌فعال (bio-active) تولید کند. به‌تازگی، مدل S4 برای بهبود عملکرد هاCLM مورد استفاده قرار گرفته و توانسته است بر محدودیت‌های ترنسفورمرهای سنتی غلبه کند. از رمز ژنتیکی انسان تا ترکیبات شیمیایی نوین، Mamba در حال شکل‌دادن به آینده‌ی پزشکی دقیق و نوآوری دارویی است.

- حوزه زیستی (ژنوم): Genomics علمی است که به مطالعه‌ی ساختار، عملکرد، تکامل، نقشه‌برداری و ژنوم ویرایش‌های موجودات زنده می‌پردازد. ژنوم انسان تقریباً شامل 3.1 billion base pairs است. مطالعه‌ی ژنوم به درک ریسک ابتلا به بیماری‌های شایع مانند cancer و diabetes کمک می‌کند. در سال‌های اخیر، تلاش‌های متعددی برای پردازش توالی‌های طولانی ژنومی انجام شده است؛ از جمله استفاده از مدل‌های Nucleotide Transformer و ترنسفورمرهای مبتنی بر BERT برای توالی‌یابی DNA. با این حال، ترنسفورمرها در حوزه‌ی ژنومیکس با پیچیدگی محاسباتی بالا و محدودیت در اندازه‌ی پنجره‌ی توجه

(attention window) مواجه‌اند که توانایی آن‌ها را در مدل‌سازی وابستگی‌های بلندمدت کاهش می‌دهد. مدل HyenaDNA نشان داده است که مدل‌های فضای حالت (State Space Models – SSMs) می‌توانند در مدل‌سازی توالی‌های ژنومی عملکردی بهتر از ترنسفورمرها ارائه دهند. در جدول (۷) که از مقاله‌ی HyenaDNA اقتباس شده است، دقت Top-1 مدل‌های مختلف بر روی مجموعه‌داده‌های ژنومی نمایش داده شده است. همچنین، عملکرد مدل‌ها بر روی ۱۸ مجموعه‌داده‌ی ژنومی دیگر در مقاله‌ی Nucleotide Transformer جدول (۶) گزارش شده است.

- حوزه شیمی (طراحی دارو): طراحی مولکول‌ها از ابتدا مستلزم پیمایش در فضای شیمیایی عظیمی است که اندازه‌ی آن می‌تواند در حدود 10^{60} باشد. در این راستا، مدل زبانی شیمیایی (Chemical Language Model – CLM) به‌عنوان ابزاری قدرتمند برای تولید مولکول‌های زیست‌فعال (bio-active) و طراحی ساختارهای مولکولی جدید معرفی شده است. مدل‌های CLM در ابتدا بر پایه‌ی شبکه‌های بازگشتی LSTM پیاده‌سازی می‌شدند، اما در ادامه نسخه‌های مبتنی بر ترنسفورمرها نیز توسعه یافتند. با این حال، پیچیدگی درجه دوم ($O(N^2)$) در ترنسفورمرها ظرفیت آن‌ها برای کاوش در فضای شیمیایی را محدود می‌کند. در پژوهش‌های اخیر، نویسندگان با انطباق مدل S4 و به‌کارگیری آن به‌عنوان یک مدل زبانی مولد (Generative CLM) موفق شده‌اند مولکول‌هایی معتبر، متنوع و زیست‌فعال طراحی کنند. این دستاورد گامی مهم در جهت بهبود فرایند طراحی دارو و افزایش کارایی مدل‌های مولد شیمیایی محسوب می‌شود.

HyenaDNA	GPT	DNABERT	CNN	Dataset
۸۵.۱	۸۰.۱	۶۶.۹	۶۹.۰	Mouse Enhancers
۹۱.۳	۸۸.۸	۹۲.۵	۷۸.۶	Coding vs Intergenomic
۹۶.۶	۹۵.۶	۹۶.۵	۹۰.۳	Human vs Worm
۴۷.۲	۷۰.۵	۷۴.۰	۶۹.۵	Human Enhancers Cohn
۸۹.۲	۸۳.۵	۸۵.۷	۶۸.۹	Human Enhancers Ensembl
۹۳.۸	۹۱.۵	۸۸.۱	۹۳.۳	Human Regulatory
۹۶.۶	۸۷.۷	۸۵.۶	۸۴.۶	Human Nontata Promoters
۸۰.۹	۷۳.۰	۷۵.۱	۶۸.۰	Human OCR Ensembl

جدول ۷: دقت Top-1 (به درصد) بر روی مجموعه‌داده‌های ژنومی در مدل‌های از پیش آموزش‌داده‌شده‌ی HyenaDNA، DNABERT و Transformer (GPT)، به همراه مدل CNN پایه (از ابتدا آموزش‌داده‌شده).

۴.۴ حوزه‌ی ویدئو

روش‌های مدرن پردازش ویدئو معمولاً بر روی کلیپ‌های کوتاه ویدئویی (به‌طور میانگین بین ۵ تا ۱۰ ثانیه) عمل می‌کنند. با این حال، درک بلندمدت ویدئو (Long-term Video Understanding) نیازمند پردازش کلیپ‌های طولانی‌تر و مدل‌سازی وابستگی‌های بلندمدت در ویدئوها است. مدل‌های Object Transformers با هدف مدل‌سازی وابستگی‌های بلندمدت در ویدئوها معرفی شده‌اند. با این حال، تلاش‌های اخیر مانند ViS4mer نشان داده‌اند که معماری‌های مبتنی بر Mamba نیز می‌توانند برای درک بلندمدت ویدئو، از جمله وظایف طبقه‌بندی ویدئوهای طولانی، مورد استفاده قرار گیرند. مدل Structured State-Space Sequence (S4) به‌عنوان یک راه‌حل نویدبخش برای مدل‌سازی وابستگی‌های پیچیده‌ی فضایی-زمانی در ویدئوهای بلندمدت شناخته شده است. پیچیدگی خطی این مدل آن را به

گزینه‌ای جذاب تبدیل کرده است؛ با این حال، یک محدودیت در آن وجود دارد: این مدل تمام توکن‌های تصویری را به‌طور مساوی در نظر می‌گیرد که می‌تواند بر کارایی و دقت تأثیر منفی بگذارد. برخلاف روش‌های پیشین مبتنی بر ماسک برای کاهش توکن‌ها، مدل Selective S4 (S5) از یک تولیدکننده‌ی ماسک سبک‌وزن استفاده می‌کند که به‌طور انتخابی توکن‌های تصویری مهم را برمی‌گزینند. این رویکرد باعث مدل‌سازی کارآمدتر و دقیق‌تر وابستگی‌های بلندمدت در ویدئو می‌شود. نکته‌ی مهم این است که ما با بهره‌گیری از راهنمایی مدل S4 به‌روزشده با مومنتوم، از انجام محاسبات سنگین Dense Self-Attention اجتناب می‌کنیم. با این وجود، مشابه سایر روش‌های کاهش توکن، خطر حذف اشتباهی توکن‌های تصویری مهم وجود دارد. برای افزایش پایداری و حفظ زمینه‌ی زمانی، رویکردی نوین با نام Long-Short Masked Contrastive Learning (LSMCL) پیشنهاد شده است. این روش به مدل اجازه می‌دهد وابستگی‌های زمانی طولانی‌تر را با استفاده از ورودی‌های ویدئویی کوتاه‌تر پیش‌بینی کند. ارزیابی‌های گسترده‌ی LSMCL بر روی مجموعه‌داده‌های چالش‌برانگیز درک ویدئوی بلندمدت شامل LVU، COIN و Breakfast نشان می‌دهد که این رویکرد در مقایسه با مدل S4 پیشرفته‌ی قبلی، تا ۹.۶ ٪ دقت بالاتری کسب کرده است، در حالی که میزان مصرف حافظه را تا ۲۳ ٪ کاهش داده است. VideoMamba یک رویکرد مبتنی بر مدل‌های فضای حالت (State Space Model – SSM) است که به‌طور خاص برای درک کارآمد ویدئو طراحی شده است. این مدل دو چالش اصلی در داده‌های ویدئویی، یعنی redundancy محلی و وابستگی‌های سراسری (global dependencies) را به‌طور هم‌زمان برطرف می‌کند. VideoMamba با سازگار کردن نوآورانه‌ی معماری Mamba برای حوزه‌ی ویدئو، این هدف را محقق می‌سازد. در حالی‌که شبکه‌های عصبی کانولوشنی سه‌بعدی (3D CNNs) و ترنسفورمرهای ویدئویی در مدل‌سازی بلندمدت و پردازش ویدئوهای با وضوح بالا محدودیت دارند، VideoMamba با بهره‌گیری از عملگر خطی ذاتی در Mamba این محدودیت‌ها را برطرف می‌کند. این عملگر دارای پیچیدگی خطی بوده و امکان مدل‌سازی بلندمدت را فراهم می‌کند؛ ویژگی‌ای که برای درک ویدئوهای طولانی و با جزئیات بالا ضروری است. VideoMamba بدون نیاز به مجموعه‌داده‌های بسیار بزرگ، از طریق روش Self-Distillation به مقیاس‌پذیری در حوزه‌ی بینایی دست یافته است. همچنین، این مدل در زمینه‌های چندوجهی (Multi-Modal) عملکردی پایدار دارد و می‌تواند با داده‌هایی از جنس‌های مختلف (مانند تصویر، صوت یا متن) ترکیب شود. با ترکیب این مزایا، VideoMamba توانسته است درک ویدئو را در مجموعه‌داده‌های مختلف از جمله Kinetics (K400)، Something-Something V2 (SSV2)، Breakfast، Long-form Video Understanding (LVU) و COIN به سطحی بی‌سابقه برساند. جدول‌های ۸،۹،۱۰ و همچنین جدول ۱۵ نتایج پیشرفته‌ی این مدل را در مجموعه‌داده‌های مذکور نشان می‌دهند. SpikeMba یک روش نوآورانه برای مسئله‌ی Temporal Video Grounding است . یکی از وظایف کلیدی در درک محتوای ویدئویی. این وظیفه شامل شناسایی دقیق بازه‌ها یا لحظات خاصی در یک ویدئو است که با یک پرس‌وجوی متنی خاص مطابقت دارند. به‌عنوان مثال، در ویدئویی از مسابقه‌ی فوتبال، اگر پرسش «گل زدن مسی» داده شود، مدل باید دقیقاً لحظه‌ای را بیابد که مسی گل می‌زند. روش‌های موجود معمولاً در استخراج ارتباطات دقیق بین چند وجه مختلف داده مانند فریم‌های ویدئو، صوت و متن دچار ضعف‌اند. SpikeMba با ترکیب شبکه‌های عصبی اسپایک‌محور (Spiking Neural Networks – SNNs) و مدل‌های فضای حالت (SSMs) این مشکل را برطرف می‌کند. در این چارچوب، مؤلفه‌ای به نام Contextual Moment Reasoner (CMR) نقش کلیدی در حفظ اطلاعات زمینه‌ای و بررسی ارتباط معنایی بین لحظات ویدئویی ایفا می‌کند. ViViM یا Video Vision Mamba مدلی است که برای وظایف بخش‌بندی اشیای پزشکی در ویدئوها طراحی شده است. این مدل با تکیه بر معماری Mamba، توانایی مدل‌سازی توالی و وابستگی‌های زمانی را بهبود می‌بخشد و دقت در بخش‌بندی اشیای ویدئویی را به

شکل قابل‌توجهی افزایش می‌دهد. ViViM در کاربردهای پزشکی، مانند تحلیل ویدئوهای آندوسکوپی یا جراحی، کارایی بالایی از خود نشان داده و مسیر تازه‌ای در تلفیق مدل‌های بینایی و سیستم‌های مبتنی بر State Space گشوده است.

K400 Top-5	K400 Top-1	FLOPs (G)	Param (M)	Input Size	Extra Data	iso.	Model	Arch.
۹۳.۹	۷۹.۸	$234 \times 3 \times 10$	۶۰	80×224^2		✗	$\text{SlowFast}_{R101+NL}$	CNN
۹۲.۳	۷۶.۰	$6 \times 3 \times 10$	۴	16×224^2		✗	X3D-M	
۹۴.۶	۸۰.۴	$194 \times 3 \times 10$	۲۰	16×312^2		✗	X3D-XL	
۹۳.۶	۷۸.۸	$88 \times 3 \times 4$	۲۸	32×224^2	IN-1K	✗	Swin-T	Trans.
۹۴.۵	۸۰.۶	$88 \times 3 \times 4$	۸۸	32×224^2	IN-1K	✗	Swin-B	
۹۵.۵	۸۲.۷	$282 \times 3 \times 4$	۸۸	32×224^2	IN-21K	✗	Swin-B	
۹۴.۴	۸۰.۲	$70 \times 1 \times 5$	۳۷	32×224^2		✗	MViTv1-B	CNN+Trans.
۹۴.۶	۸۱.۰	$64 \times 1 \times 5$	۳۵	16×224^2		✗	MViTv2-S	
۹۴.۷	۸۰.۸	$42 \times 1 \times 4$	۲۱	16×224^2	IN-1K	✗	UniFormer-S	
۹۵.۱	۸۲.۰	$97 \times 1 \times 4$	۵۰	16×224^2	IN-1K	✗	UniFormer-B	
۹۵.۴	۸۳.۰	$259 \times 3 \times 4$	۵۰	32×224^2	IN-1K	✗	UniFormer-B	
–	۷۹.۲	$1040 \times 1 \times 1$	۱۲۱	64×224^2	IN-21K	✓	STAM	Trans.
۹۴.۷	۸۰.۷	$2380 \times 3 \times 1$	۱۲۱	96×224^2	IN-21K	✓	TimeSformer-L	
۹۴.۷	۸۱.۳	$3992 \times 3 \times 4$	۳۱۱	16×224^2	IN-21K	✓	ViViT-L	
۹۵.۲	۸۱.۱	$959 \times 3 \times 10$	۳۱۱	16×224^2	IN-21K	✓	Mformer-HR	
۹۳.۵	۷۸.۱	$17 \times 3 \times 4$	۷	16×224^2	IN-1K	✓	VideoMamba-Ti	SSM
۹۳.۹	۷۸.۸	$34 \times 3 \times 4$	۷	32×224^2	IN-1K	✓	VideoMamba-Ti	
۹۴.۸	۸۰.۳	$202 \times 3 \times 4$	۷	64×224^2	IN-1K	✓	VideoMamba-Ti	
۹۴.۸	۸۰.۸	$68 \times 3 \times 4$	۲۶	16×224^2	IN-1K	✓	VideoMamba-S	
۹۵.۲	۸۱.۵	$135 \times 3 \times 4$	۲۶	32×224^2	IN-1K	✓	VideoMamba-S	
۹۵.۶	۸۲.۷	$395 \times 3 \times 4$	۲۶	64×224^2	IN-1K	✓	VideoMamba-S	
۹۵.۴	۸۱.۹	$202 \times 3 \times 4$	۷۴	16×224^2	IN-1K	✓	VideoMamba-M	
۹۵.۷	۸۲.۴	$403 \times 3 \times 4$	۷۴	32×224^2	IN-1K	✓	VideoMamba-M	
۹۶.۱	۸۳.۳	$2368 \times 3 \times 4$	۷۴	64×224^2	IN-1K	✓	VideoMamba-M	
–	۸۱.۱	$282 \times 3 \times 4$	۸۸	32×224^2	IN-1K	✗	BEVT-B_{800e}	Trans.
۹۴.۹	۸۱.۳	$180 \times 3 \times 7$	۸۷	16×224^2		✓	ST-MAE-B_{2400e}	
۹۳.۸	۷۹.۰	$57 \times 3 \times 5$	۸۲	16×224^2		✓	$\text{VideoMAE-S}_{2400e}$	
۹۵.۱	۸۱.۵	$180 \times 3 \times 5$	۸۷	16×224^2		✓	$\text{VideoMAE-B}_{1600e}$	
۹۷.۰	۸۵.۷	$180 \times 3 \times 5$	۸۷	8×224^2	CLIP-400M	✓	UMT-B_{800e}	
۹۵.۴	۸۲.۰	$101 \times 3 \times 4$	۷۴	8×224^2	CLIP-400M	✓	$\text{VideoMamba-M}_{800e}$	SSM
۹۵.۹	۸۳.۴	$202 \times 3 \times 4$	۷۴	16×224^2	CLIP-400M	✓	$\text{VideoMamba-M}_{800e}$	
۹۶.۲	۸۳.۹	$403 \times 3 \times 4$	۷۴	32×224^2	CLIP-400M	✓	$\text{VideoMamba-M}_{800e}$	
۹۶.۹	۸۵.۰	$2368 \times 3 \times 4$	۷۴	64×224^2	CLIP-400M	✓	$\text{VideoMamba-M}_{800e}$	

جدول ۸: مقایسه‌ی مدل VideoMamba با روش‌های پیشرفته‌ی روز (State-of-the-Art) در مجموعه‌داده‌ی Kinetics-400 مربوط به صحنه‌ها. عبارت iso. نشان‌دهنده‌ی معماری ایزوتروپیک (بدون لایه‌های downsampling) است. روش Masked Modeling نیز برای مدل Mamba کاربرد دارد، اما به‌دلیل ناهمخوانی در معماری، منجر به هم‌ترازی ضعیف‌تر می‌شود. این جدول از مقاله‌ی VideoMamba اقتباس شده است.

SSV2 Top-5	SSV2 Top-1	FLOPs (G)	Param (M)	Input Size	Extra Data	iso.	Model	Arch.
۸۷.۶	۶۳.۱	$106 \times 3 \times 1$	۵۳	32×224^2	K400	✗	SlowFast_{R101}	CNN
۸۹.۳	۶۴.۵	$75 \times 1 \times 1$	۲۱	16×224^2	IN-1K	✗	CT-Net_{R50}	
۹۱.۶	۶۵.۳	$75 \times 1 \times 1$	۲۶	16×224^2	IN-1K	✗	TDN_{R50}	
۹۲.۷	۹۶.۶	$88 \times 3 \times 1$	۸۹	32×224^2	K400	✗	Swin-B	Trans.
۸۹.۲	۶۴.۷	$71 \times 3 \times 1$	۳۷	16×224^2	K400	✗	MViTv1-B	CNN+Trans.
۹۰.۸	۶۷.۱	$170 \times 3 \times 1$	۳۷	32×224^2	K400	✗	MViTv1-B	
۹۱.۴	۶۸.۲	$65 \times 3 \times 1$	۳۵	16×224^2	K400	✗	MViTv2-S	
۹۲.۷	۷۰.۵	$225 \times 3 \times 1$	۵۱	32×224^2	K400	✗	MViTv2-B	
۹۱.۴	۶۷.۷	$42 \times 3 \times 1$	۲۰	16×224^2	IN-21K+K400	✗	UniFormer-S	
۹۲.۸	۷۰.۴	$97 \times 3 \times 1$	۵۰	16×224^2	IN-21K+K400	✗	UniFormer-B	
–	۶۲.۵	$1703 \times 3 \times 1$	۱۲۱	16×224^2	IN-21K	✓	TimeSformer-HR	Trans.
۸۹.۸	۶۵.۴	$3992 \times 3 \times 4$	۳۱۱	16×224^2	IN-21K+K400	✓	ViViT-L	
۹۱.۲	۶۸.۱	$1185 \times 3 \times 1$	۳۱۱	16×336^2	IN-21K+K400	✓	Mformer-HR	
۸۹.۱	۶۵.۱	$9 \times 3 \times 2$	۷	8×224^2	IN-1K	✓	VideoMamba-Ti	SSM
۸۹.۶	۶۶.۰	$17 \times 3 \times 1$	۷	16×224^2	IN-1K	✓	VideoMamba-Ti	
۹۰.۰	۶۶.۲	$28 \times 3 \times 2$	۷	16×224^2	IN-1K	✓	VideoMamba-Ti	
۹۰.۴	۶۶.۶	$34 \times 3 \times 2$	۲۶	8×224^2	IN-1K	✓	VideoMamba-S	
۹۰.۹	۶۷.۶	$68 \times 3 \times 2$	۲۶	16×224^2	IN-1K	✓	VideoMamba-S	
۹۱.۲	۶۸.۱	$112 \times 3 \times 2$	۲۶	16×224^2	IN-1K	✓	VideoMamba-S	
۹۱.۰	۶۷.۳	$101 \times 3 \times 4$	۷۴	8×224^2	IN-1K	✓	VideoMamba-M	
۹۱.۴	۶۸.۳	$202 \times 3 \times 4$	۷۴	16×224^2	IN-1K	✓	VideoMamba-M	
۹۱.۶	۶۸.۴	$303 \times 3 \times 4$	۷۴	16×224^2	IN-1K	✓	VideoMamba-M	
–	۷۰.۶	$321 \times 3 \times 1$	۸۸	32×224^2	IN-1K+K400	✗	BEVT-B_{800e}	Trans.
۹۰.۳	۶۶.۸	$57 \times 3 \times 2$	۸۲	16×224^2		✓	$\text{VideoMAE-B}_{2400s}$	
۹۲.۴	۷۰.۸	$180 \times 3 \times 2$	۸۷	16×224^2		✓	$\text{VideoMAE-B}_{2400e}$	
۹۲.۶	۷۰.۸	$180 \times 3 \times 2$	۸۷	8×224^2	CLIP-400M	✓	UMT-B_{800e}	
۹۲.۶	۷۰.۲	$101 \times 3 \times 2$	۷۴	8×224^2	CLIP-400M	✓	$\text{VideoMamba-M}_{800e}$	SSM
۹۲.۷	۷۱.۰	$202 \times 3 \times 2$	۷۴	16×224^2	CLIP-400M	✓	$\text{VideoMamba-M}_{800e}$	
۹۲.۹	۷۱.۴	$303 \times 3 \times 2$	۷۴	16×288^2	CLIP-400M	✓	$\text{VideoMamba-M}_{800e}$	

جدول ۹: مقایسه‌ی مدل VideoMamba با روش‌های پیشرفته‌ی روز (State-of-the-Art) در مجموعه‌داده‌ی SthSth V2 مرتبط با پردازش‌های زمانی. عبارت .iso به معنای معماری ایزوتروپیک (بدون لایه‌های downsampling) است. روش Masked Modeling نیز برای مدل Mamba قابل استفاده است و عملکرد بهتری نسبت به VideoMAE دارد. این جدول از مقاله‌ی VideoMamba اقتباس شده است.

User(↓)		Metadata(↑)				Content(↑)			Backbone	e2e	Method
View	Like	Year	Wtr.	Genre	Dir.	Scene	Speak	Rel.			
۴.۴۶	۰.۳۲	۳۶.۱۰	۳۸.۵۰	۵۱.۹۰	۴۷.۳۰	۵۴.۹۰	۳۷.۹۰	۵۲.۸۰	S3D	✗	VideoBERT
۳.۵۵	۰.۲۳	۳۹.۱۰	۳۴.۵۰	۵۴.۶۰	۵۱.۲۰	۵۶.۹۰	۳۹.۴۰	۵۳.۱۰	ResNet	✗	Object Trans.
۳.۸۳	۰.۳۱	۳۹.۱۶	۴۲.۲۶	۵۲.۷۰	۵۶.۰۷	۶۲.۷۹	۳۷.۳۱	۵۲.۳۸	ViT-L	✗	LST
۳.۹۳	۰.۳۱	۴۱.۲۵	۴۸.۲۱	۴۹.۴۵	۵۸.۸۷	۶۰.۴۶	۳۸.۸۰	۵۰.۰۰	ViT-L	✗	Performer
۳.۸۶	۰.۲۹	۴۳.۳۵	۴۷.۰۲	۵۵.۷۹	۵۵.۱۴	۶۶.۲۷	۳۹.۳۹	۵۰.۰۰	ViT-L	✗	Orthoformer
۳.۶۳	۰.۲۶	۴۴.۷۵	۴۸.۸۰	۵۴.۷۱	۶۲.۶۱	۶۷.۴۴	۴۰.۷۹	۵۷.۱۴	ViT-L	✗	ViS4mer
۲.۹۰	۰.۲۶	۴۸.۲۳	۵۲.۹۸	۶۵.۲۴	۶۷.۲۹	۷۰.۷۳	۴۰.۴۳	۶۲.۵۰	VM-Ti	✓	VideoMamba$_{f32}$

جدول ۱۰: مقایسه‌ی مدل‌ها با روش‌های پیشرفته‌ی روز (State-of-the-Art) بر روی مجموعه‌داده‌ی LVU. عبارت e2e به معنی روش‌های انتها به انتها (end-to-end) است که بدون استفاده از استخراج ویژگی‌های از پیش محاسبه‌شده عمل می‌کنند. اختصارات Rel.، Dir. و Wtr. به ترتیب به معنای «روابط» (Relation)، «کارگردان» (Director) و «نویسنده» (Writer) هستند. این جدول از مقاله‌ی Video-Mamba اقتباس شده است.

۵.۴ حوزه‌ی داده‌های جدولی (Tabular Domain)

اگرچه شبکه‌های عصبی کانولوشنی (CNN) و ترنسفورمرها عملکرد قابل‌توجهی بر روی داده‌های جدولی از خود نشان داده‌اند، اما اجرای آن‌ها معمولاً به منابع محاسباتی قابل‌توجه، پیش‌پردازش داده و تنظیمات دقیق (tuning) نیاز دارد. افزون بر این، این مدل‌ها در یادگیری تدریجی ویژگی‌ها (Incremental Feature Learning) که در آن ویژگی‌ها به‌صورت متوالی به مجموعه داده اضافه می‌شوند، ممکن است با چالش مواجه شوند. مدل MambaTab نشان داده است که می‌تواند یادگیری تدریجی ویژگی‌ها را به‌صورت کارآمد مدل‌سازی کند و در عین حال، عملکردی قابل‌مقایسه با CNNها و ترنسفورمرها ارائه دهد، با این تفاوت که از نظر تعداد پارامترها به‌مراتب کارآمدتر است. برخلاف مدل‌های سنتی، MambaTab قادر است خود را با مجموعه‌داده‌هایی که در طول زمان ویژگی‌های جدیدی به آن‌ها افزوده می‌شود، بدون نیاز به بازآموزی کامل سازگار کند. همچنین این مدل به حداقل پیش‌پردازش نیاز دارد و از منابع محاسباتی کمتری استفاده می‌کند. در این پژوهش، رویکردی مبتنی بر مدل فضای حالت ساختاریافته (Structured State-Space Model - SSM) برای داده‌های جدولی توسعه داده شده است که با عنوان MambaTab شناخته می‌شود. مدل‌های SSM توانایی بالایی در استخراج بازنمایی‌های مؤثر از داده‌هایی با وابستگی‌های بلندمدت دارند. MambaTab از معماری Mamba یکی از گونه‌های نوظهور SSM برای یادگیری نظارت‌شده‌ی انتهابه‌انتها (End-to-End Supervised Learning) بر روی داده‌های جدولی استفاده می‌کند. در مقایسه با مدل‌های

مرجع پیشرفته (State-of-the-Art)، MambaTab عملکرد بهتری را با تعداد پارامتر کمتر و نیاز حداقلی به پیش‌پردازش ارائه می‌دهد؛ موضوعی که در آزمایش‌های تجربی بر روی مجموعه‌داده‌های متنوع به‌صورت تجربی تأیید شده است. کارایی، مقیاس‌پذیری، قابلیت تعمیم و دقت بالای پیش‌بینی در MambaTab، این مدل را به یک راه‌حل سبک، آماده‌به‌کار (Out-of-the-Box) و مناسب برای کاربردهای گسترده در داده‌های جدولی تبدیل کرده است.

۶.۴ حوزه‌ی صوت و گفتار (Audio and Speech Domain)

در حوزه‌ی پردازش گفتار، شبکه‌های عصبی بازگشتی (Recurrent Neural Networks - RNNs) به‌طور سنتی انتخاب اصلی برای انجام وظایف مختلف بوده‌اند. با این حال، ظهور ترنسفورمرهای مبتنی بر مکانیزم توجه (Attention-based Transformers) موجب تلاش‌هایی برای به‌کارگیری آن‌ها در مجموعه‌داده‌های گفتاری شد. اگرچه شبکه‌های مبتنی بر توجه در مدل‌سازی وابستگی‌های بلندمدت عملکرد قابل‌توجهی دارند، اما در مدیریت وابستگی‌های محلی معمولاً با چالش مواجه می‌شوند و نیاز به ترکیب با شبکه‌های کانولوشنی دارند. این ترکیب در مدل‌هایی نظیر Conformer و BranchFormer به‌خوبی مشاهده می‌شود. یکی از تلاش‌های پیشگامانه در به‌کارگیری مدل‌های فضای حالت (State Space Models - SSMs) در وظایف گفتاری، مدل SaShiMi است. این مدل با توجه به ناپایداری عددی S4 در وظایف خودرگرسیو، ماتریس حالت را به‌صورت Hurwitz تعریف می‌کند تا برای تولید سیگنال‌های صوتی مناسب شود. مدل Multi-Head State Space Model (MH-SSM) یک لایه‌ی چندسری SSM را به‌عنوان جایگزینی برای لایه‌ی توجه در ترنسفورمرها معرفی می‌کند و مدل ترکیبی حاصل را StateFormer می‌نامد. StateFormer توانسته است عملکردی در سطح مدل‌های پیشرفته (State-of-the-Art) در مجموعه‌داده‌هایی مانند LibriSpeech ارائه دهد. مدل SP-Mamba به ناکارآمدی ترنسفورمرهای مبتنی بر توجه در مدل‌سازی سیگنال‌های گفتاری بلند اشاره می‌کند، زیرا این مدل‌ها از پیچیدگی محاسباتی درجه دوم (Quadratic Complexity) رنج می‌برند. در عوض، SP-Mamba از معماری دوطرفه‌ی Mamba بهره می‌برد که هم ویژگی‌های حوزه‌ی زمان و هم حوزه‌ی فرکانس را ترکیب می‌کند. این مدل، شبکه‌ی Bidirectional LSTM (BLSTM) را در ماژول زمان‌فرکانس با مکانیزم توجه چندسری جایگزین کرده است. ارزیابی این مدل بر روی مجموعه‌داده‌هایی مانند LibriSpeech و WHAM، موفقیت آن را در بهبود مدل‌سازی سیگنال‌های گفتاری تأیید می‌کند. به‌طور هم‌زمان، مدل Dual-Path Mamba (DPMamba) سیگنال‌های گفتاری بلند را به بخش‌های کوچک‌تر تقسیم کرده و در هر دو جهت زمانی، مدل Mamba را بر روی آن‌ها اعمال می‌کند. DPMamba عملکردی برتر از ترنسفورمرها در وظایفی مانند جداسازی گفتار نشان داده است، که این موضوع با ارزیابی بر روی مجموعه‌داده‌هایی مانند WSJ0-2mix تأیید شده است. علاوه بر این، مدل Multichannel Long-Term Streaming Neural Speech Enhancement برای بهبود گفتار گویندگان ایستا و متحرک، نوعی نسخه‌ی مبتنی بر Mamba از معماری SpatialNet را پیشنهاد می‌کند که جایگزین مکانیزم خودتوجهی (Self-Attention) شده است. این رویکرد نشان‌دهنده‌ی انعطاف‌پذیری بالای مدل‌های فضای حالت (SSMs) در بهبود وظایف مرتبط با گفتار است.

۷.۴ حوزه‌ی سری‌های زمانی (Time Series Domain)

در حوزه‌ی مدل‌سازی داده‌های سری زمانی، رویکردهای سنتی عمدتاً بر پایه‌ی مدل‌های آماری مانند ARIMA (AutoRegressive Integrated Moving Average) برای پیش‌بینی و تحلیل استوار بوده‌اند. با این حال، پیشرفت‌های اخیر باعث تغییر رویکرد به‌سمت استفاده از ترنسفورمرها (Transformers) که در ابتدا برای پردازش زبان طبیعی (NLP) طراحی شده بودند در حوزه‌ی سری‌های زمانی شده است. مدل‌هایی مانند AutoFormer، Informer، FEDFormer، TFT، WaveNet و PatchTST برای سازگارسازی معماری ترنسفورمر با داده‌های سری زمانی توسعه یافته‌اند. با وجود اثربخشی این مدل‌ها، بسیاری از آن‌ها با دو چالش اساسی روبه‌رو هستند: نخست، پیچیدگی محاسباتی مکانیزم توجه (Attention Complexity) و دوم، توانایی محدود در ثبت وابستگی‌های بلندمدت که ذاتاً در داده‌های سری زمانی وجود دارد. برای غلبه بر این چالش‌ها، پژوهش‌های اخیر به‌سمت ادغام مدل‌های فضای حالت (State Space Models - SSMs) از جمله S4 در تحلیل سری‌های زمانی حرکت کرده‌اند. رویکردهای پیشرفته‌ی کنونی در مدل‌سازی سری‌های زمانی شامل مدل‌های نوآورانه‌ای چون Timemachine، SiMBA و MambaMix هستند. این مدل‌ها با بهره‌گیری از قدرت مدل‌های فضای حالت، وابستگی‌ها و الگوهای زمانی موجود در داده‌های سری زمانی را به‌صورت کارآمد ثبت می‌کنند. برای ارائه‌ی مقایسه‌ای جامع میان ترنسفورمرها و مدل‌های فضای حالت در تحلیل سری‌های زمانی، نتایج معیارهای بنچمارک از هر دو حوزه گردآوری شده‌اند که در جدول 5.2.5 (اقتباس‌شده از مقاله‌ی SiMBA) آورده شده است. چنین تحلیل‌های مقایسه‌ای به پژوهشگران کمک می‌کند تا بینش‌های ارزشمندی درباره‌ی نقاط قوت و ضعف هر رویکرد به‌دست آورند و مسیر پیشرفت‌های آتی در مدل‌سازی داده‌های سری زمانی را هموار سازند.

MTGNN		Pyraformer		Autoformer		FEDFormer		DLinear		ETSFormer		PatchTST		Crossformer		TimesNet		Simba		Models
MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	Metric
																				ETTm1
0.446	0.379	0.510	0.543	0.475	0.505	0.419	0.379	0.372	0.345	0.398	0.375	0.377	0.339	0.395	0.349	0.375	0.338	0.360	0.324	96
0.430	0.470	0.655	0.754	0.523	0.621	0.453	0.426	0.389	0.380	0.410	0.408	0.392	0.376	0.411	0.405	0.387	0.374	0.382	0.363	192
0.437	0.473	0.655	0.754	0.587	0.613	0.459	0.506	0.413	0.413	0.438	0.435	0.417	0.408	0.431	0.432	0.411	0.410	0.405	0.395	336
0.499	0.553	0.724	0.905	0.561	0.671	0.490	0.543	0.453	0.474	0.462	0.499	0.461	0.499	0.463	0.487	0.450	0.478	0.437	0.451	720
																				ETTm2
0.299	0.261	0.495	0.438	0.340	0.255	0.287	0.207	0.292	0.193	0.280	0.189	0.273	0.189	0.292	0.208	0.267	0.187	0.263	0.177	96
0.328	0.265	0.673	0.730	0.340	0.281	0.328	0.269	0.362	0.284	0.319	0.253	0.314	0.252	0.332	0.263	0.309	0.249	0.306	0.245	192
0.374	0.361	0.845	1.201	0.372	0.339	0.366	0.325	0.422	0.369	0.357	0.314	0.357	0.318	0.369	0.337	0.351	0.321	0.343	0.304	336
0.379	0.436	1.451	3.625	0.432	0.432	0.436	0.421	0.522	0.554	0.414	0.414	0.430	0.440	0.430	0.429	0.404	0.408	0.399	0.400	720
																				ETTh1
0.517	0.515	0.681	0.664	0.459	0.449	0.415	0.376	0.400	0.386	0.479	0.494	0.408	0.385	0.428	0.384	0.402	0.384	0.395	0.379	96
0.522	0.552	0.681	0.691	0.518	0.514	0.448	0.420	0.432	0.437	0.504	0.538	0.432	0.431	0.452	0.438	0.429	0.436	0.424	0.432	192
0.597	0.609	0.681	0.693	0.512	0.514	0.507	0.506	0.460	0.481	0.504	0.574	0.457	0.471	0.459	0.495	0.447	0.493	0.444	0.473	336
0.597	0.609	0.894	0.963	0.512	0.514	0.507	0.506	0.516	0.519	0.535	0.562	0.483	0.497	0.501	0.522	0.500	0.521	0.469	0.483	720
																				ETTh2
0.454	0.354	0.597	0.645	0.388	0.346	0.397	0.358	0.387	0.333	0.391	0.340	0.376	0.343	0.391	0.347	0.374	0.340	0.339	0.290	96
0.457	0.355	0.688	0.748	0.388	0.352	0.429	0.384	0.437	0.379	0.437	0.391	0.419	0.394	0.447	0.414	0.414	0.370	0.360	0.343	192
0.540	0.516	0.747	0.907	0.486	0.482	0.487	0.496	0.541	0.594	0.479	0.485	0.453	0.448	0.465	0.449	0.452	0.452	0.406	0.376	336
0.576	0.532	0.783	0.963	0.511	0.515	0.474	0.463	0.657	0.831	0.497	0.500	0.483	0.464	0.505	0.479	0.468	0.462	0.431	0.407	720
																				Electricity
0.316	0.193	0.473	0.386	0.311	0.201	0.287	0.203	0.304	0.213	0.291	0.187	0.268	0.159	0.288	0.185	0.300	0.198	0.253	0.165	96
0.318	0.216	0.447	0.473	0.338	0.231	0.329	0.214	0.314	0.209	0.329	0.212	0.296	0.195	0.312	0.211	0.302	0.198	0.262	0.173	192
0.348	0.260	0.447	0.473	0.338	0.231	0.329	0.214	0.301	0.209	0.329	0.212	0.296	0.195	0.312	0.211	0.300	0.198	0.277	0.188	336
0.398	0.290	0.445	0.376	0.338	0.254	0.329	0.229	0.330	0.245	0.324	0.212	0.317	0.215	0.332	0.223	0.320	0.220	0.305	0.214	720
																				Traffic
0.437	0.660	0.468	0.867	0.388	0.613	0.366	0.587	0.396	0.650	0.392	0.607	0.319	0.583	0.329	0.591	0.321	0.593	0.268	0.468	96
0.438	0.649	0.467	0.869	0.382	0.612	0.373	0.604	0.373	0.598	0.396	0.621	0.333	0.591	0.345	0.607	0.336	0.617	0.317	0.413	192
0.437	0.643	0.467	0.869	0.382	0.616	0.373	0.621	0.370	0.639	0.396	0.621	0.332	0.599	0.348	0.622	0.350	0.642	0.339	0.529	336
0.437	0.639	0.473	0.891	0.408	0.660	0.382	0.626	0.394	0.645	0.396	0.632	0.341	0.601	0.348	0.620	0.350	0.640	0.297	0.564	720
																				Weather
0.329	0.230	0.556	0.622	0.311	0.266	0.292	0.226	0.287	0.217	0.281	0.197	0.230	0.171	0.251	0.191	0.220	0.172	0.219	0.176	96
0.329	0.230	0.574	0.739	0.336	0.267	0.309	0.237	0.297	0.229	0.297	0.237	0.293	0.211	0.282	0.219	0.261	0.219	0.260	0.222	192
0.398	0.354	0.753	1.004	0.395	0.359	0.359	0.308	0.353	0.283	0.353	0.298	0.321	0.277	0.332	0.287	0.306	0.280	0.297	0.275	336
0.371	0.409	0.934	1.420	0.428	0.419	0.428	0.403	0.381	0.345	0.288	0.352	0.367	0.365	0.378	0.368	0.359	0.365	0.349	0.350	720

جدول ۱۱: نتایج پیش‌بینی بلندمدت چندمتغیره. در این جدول، طول‌های پیش‌بینی $T\{96, 192, 336, 720\}$ برای تمامی مجموعه‌داده‌ها با پنجره‌ی جست‌وجوی ۹۶ به‌کار رفته‌اند. بهترین نتایج با حروف پررنگ و نتایج دوم با خط زیرین مشخص شده‌اند. این جدول از مقاله‌ی SiMBA اقتباس شده است

۸.۴ سیستم‌های توصیه‌گر

سیستم‌های توصیه‌گر می‌توانند به‌صورت گراف‌های heterophilic مدل‌سازی شوند (در مقابل گراف‌های homophilous که در آن گره‌ها تمایل دارند به گره‌های هم‌کلاس خود متصل شوند). شبکه‌های عصبی گرافی (GNNs) معمولاً بر روی گراف‌های homophilic کار می‌کنند، در حالی‌که GraphMamba نشان داده است که می‌تواند در گراف‌های heterophilic عملکرد مطلوبی در وظایفی مانند پیش‌بینی امتیاز محصولات داشته باشد (گره‌ها در اینجا نمایانگر محصولاتی مانند کتاب، موسیقی، ویدیو یا DVD هستند، در حالی که یال‌ها محصولاتی را که معمولاً با یکدیگر خریداری می‌شوند، به هم متصل می‌کنند). مدل GraphMamba با استفاده از مراحلی مانند neighborhood tokenization، token ordering، selective scan، و رمزگذاری‌های محلی، مکانی و ساختاری، مدل‌های SSM را برای کار بر روی گراف‌های heterophilic سازگار می‌کند. مدل DenseSSM شکاف بین SSMs و Transformers را پر می‌کند و مدل‌های زبانی بزرگ کارآمدتری با عملکرد بهبودیافته ارائه می‌دهد. مدل‌های زبانی بزرگ (LLMs) که بر پایه‌ی معماری Transformer ساخته شده‌اند، با محدودیت‌های محاسباتی و حافظه مواجه‌اند. DenseSSM رویکردی نوآورانه است که جریان اطلاعات بین لایه‌های SSM را از طریق ادغام انتخابی حالت‌های پنهان لایه‌های کم‌عمق در لایه‌های عمیق‌تر تقویت می‌کند. مدل‌های فضای حالت (State Space Models, SSMs) از پیچیدگی محاسباتی کمتری برخوردارند، اما هنوز به‌طور کامل به عملکرد Transformerها نرسیده‌اند. این مدل ضمن حفظ اطلاعات دقیق و جزئی که برای خروجی نهایی حیاتی هستند، موازی‌سازی در آموزش و کارایی در استنتاج را نیز حفظ می‌کند.

۹.۴ حوزه‌ی گراف

Graph-Mamba یک روش نوآورانه است که یک بلوک Mamba را با سازوکار انتخاب گره ترکیب می‌کند تا مدل‌سازی وابستگی‌های بلندبرد در شبکه‌های گرافی را بهبود بخشد. هدف این پژوهش، پرداختن به چالش مدل‌سازی وابستگی‌های دوربرد در داده‌های گرافی با استفاده از روش‌های کارآمد است. هسته‌ی اصلی Graph-Mamba، بلوک Graph-Mamba Block (GMB) است که سازوکار انتخابی ماژول Mamba را با رویکرد اولویت‌بندی گره‌ها ترکیب می‌کند. بلوک Mamba به دلیل کارایی بالای خود در مدل‌سازی وابستگی‌های بلندبرد در داده‌های ترتیبی شناخته شده است، در حالی‌که سازوکار انتخاب گره، گره‌ها را به‌صورت خاص‌گرافی اولویت‌بندی و بازچینش می‌کند. Graph-Mamba با ترکیب سازوکار انتخاب Mamba و راهبردهای مبتنی بر گراف، استدلال آگاه از زمینه را تقویت می‌کند. با تدوین راهبردهای اولویت‌بندی و بازچینش گره به‌صورت گراف‌محور، Graph-Mamba به بهبود قابل‌توجهی در استدلال زمینه‌محور و عملکرد پیش‌بینی دست می‌یابد. شایان توجه است که این مدل در وظایف پیش‌بینی بلندبرد گرافی عملکردی فراتر از روش‌های پیشرفته‌ی روز دارد، در حالی‌که تنها بخشی از هزینه‌ی محاسباتی (از نظر FLOPs و مصرف حافظه‌ی GPU) را نیاز دارد. نسخه‌ی Spatio-Temporal Graph Mamba (STG-Mamba) نیز به‌صورت موازی توسعه یافته است تا از Mamba برای مدل‌سازی داده‌های گرافی فضایی-زمانی بهره گیرد.

۱۰.۴ سیستم‌های چندوجهی

در حوزه‌ی سیستم‌های چندوجهی که داده‌های دیداری و زبانی را با یکدیگر ادغام می‌کنند، تلاش‌های اخیر بر بهبود معماری‌های سنتی از طریق ماژول‌های نوآورانه برای پردازش مؤثر انواع داده‌های متنوع متمرکز شده است. مدل VL-Mamba نسخه‌ی گسترش‌یافته‌ای از معماری پایه‌ی Mamba است که با افزودن یک ماژول اتصال چندوجهی اختصاصی طراحی‌شده برای وظایف چندوجهی، داده‌های دیداری و زبانی را به‌طور هم‌زمان پردازش می‌کند. این ماژول شامل یک بخش vision selection scan و دو لایه‌ی خطی است که به ادغام مؤثر اطلاعات دیداری و متنی کمک می‌کند. عملکرد VL-Mamba به‌صورت دقیق بر روی مجموعه‌داده‌های معیار چندوجهی استاندارد مانند VQA (Visual Question Answering)، GQA (Visual Grounding Question Answering)، و SQA (Science Question Answering) ارزیابی شده است. تحلیل‌های مقایسه‌ای نشان می‌دهند که VL-Mamba سطح عملکردی مشابه با مدل‌های زبانی و دیداری پیشرفته‌ی روز، از جمله مدل‌های زبانی بزرگ (LLMs) مانند Lava-1.5 و LavaVA-Phi دارد. نکته‌ی قابل توجه آن است که VL-Mamba چالش کارایی ناشی از پیچیدگی محاسباتی درجه دوم شبکه‌های Transformer را که معمولاً در وظایف پایین‌دستی استفاده می‌شوند، برطرف می‌سازد. در پژوهشی مشابه، مدل Cobra با گسترش معماری Mamba از طریق افزودن اطلاعات دیداری به‌وسیله‌ی یک رمزگذار تصویر معرفی شده است. هدف از این کار، ایجاد یک مدل زبانی بزرگ چندوجهی (MLLM) کارآمد است. Cobra با ادغام داده‌های دیداری از طریق رمزگذار تصویر و بهره‌گیری از یک دستورالعمل آموزشی دقیق طراحی‌شده، به‌عنوان مدلی قدرتمند در حوزه‌ی زبان چندوجهی ظاهر شده است. ارزیابی‌های انجام‌شده بر روی مجموعه‌داده‌های معیار دیداری-زبانی، از جمله VQA و GQA، نشان می‌دهند که عملکرد Cobra با مدل‌های Transformer پیشرفته مانند Lava-Phi رقابت‌پذیر است. تحلیل جامع ارائه‌شده در جدول ۱۳ (اقتباس‌شده از مقاله‌ی VL-Mamba) بینش‌های ارزشمندی درباره‌ی کارایی مدل‌های Transformer و SSM در حوزه‌ی چندوجهی فراهم می‌آورد و مسیر پیشرفت‌های آتی را در این حوزه‌ی در حال رشد روشن می‌سازد.

۱۱.۴ یادگیری تقویتی

یادگیری تقلیدی (Behavior Cloning, BC) نقش مهمی در یادگیری تقویتی برخط (Reinforcement Learning, RL) ایفا می‌کند، زیرا مستقیماً نگاشت بین حالت‌ها و کنش‌ها را از داده‌های در دسترس می‌آموزد. با این حال، BC در شرایطی که نمونه‌های کافی از رفتارهای متخصص در دسترس نباشند، با چالش‌هایی مواجه است. برای رفع این محدودیت، رویکردی به نام return-conditioned BC معرفی شده است. مدل Decision Transformer (DT) یک رویکرد انقلابی است که یادگیری تقویتی را به‌صورت یک مسئله‌ی مدل‌سازی دنباله‌ای در نظر می‌گیرد. این مدل از return-conditioned BC برای بهبود عملکرد RL بهره می‌گیرد. DT معماری Transformer را در زمینه‌ی یادگیری تقویتی به کار می‌گیرد تا بتواند وابستگی‌های پیچیده میان دنباله‌ای از حالت‌ها، کنش‌ها و پاداش‌ها را ثبت کند. هسته‌ی اصلی DT در شبکه‌ی causal self-attention آن نهفته است که به‌طور مؤثری دنباله‌های شامل حالت، کنش و پاداش را مدل می‌کند. در راستای ثبت بهتر وابستگی‌های زمانی و الگوهای پیچیده‌ی ذاتی در وظایف مدل‌سازی دنباله‌ای، مدل De-cision Mamba (DMamba) تمرکز خود را بر مدل‌سازی وابستگی‌های زمانی و الگوهای پیچیده قرار داده است. برخلاف DT، مدل DMamba به جای استفاده از سازوکار causal self-attention، از چارچوب Mamba بهره می‌گیرد. در این مدل، Mamba به‌عنوان یک ماژول ترکیب توکن‌ها (token mixing module) درون معماری متداول Transformer به‌کار می‌رود که شامل شبکه‌های پیش‌خور (feed-forward networks) و نرمال‌سازی لایه‌ای (layer normalization) است. ارزیابی‌های گسترده نشان داده‌اند که DMamba عملکردی برتر از DT در مجموعه‌داده‌های D4RL و Atari دارد. این مدل توانایی بالاتری در درک وابستگی‌های زمانی پیچیده از خود نشان داده است و در نتیجه کارایی بیشتری را در وظایف RL به نمایش گذاشته است. شایان ذکر است که مدل Meta-RL، اگرچه نام آن Mamba را شامل می‌شود، اما رویکرد متفاوتی دارد. این مدل به‌جای اینکه یک مدل فضای حالت باشد، یک استراتژی یادگیری فرا-تقویتی (meta-reinforcement learning) را پیشنهاد می‌کند. هدف Meta-RL بهبود عملکرد RL از طریق یادگیری میان چندین وظیفه‌ی مختلف است.

۵ نتایج پیشرفته (State of the Art Results)

در این بخش، نتایج به‌دست‌آمده از منابع متعدد گردآوری شده و تحلیل آن‌ها مورد بحث قرار می‌گیرد. در ابتدا، نتایج مربوط به معیار Long Range Arena (LRA) برای پردازش دنباله‌های بلند ارائه می‌شود، زیرا LRA شامل وظایف پردازش داده‌های متنی مانند ListOps و همچنین وظایف دیداری مانند PathFinder است. افزون بر این، چهار معیار متنی شامل GLUE، Pile و Wikitext نیز در این تحلیل گنجانده شده و نتایج مربوط به آن‌ها از چندین مقاله‌ی مرتبط با مدل‌های SSM گزارش شده است. همچنین، معیارهای دیداری از جمله طبقه‌بندی تصاویر در مجموعه‌داده‌ی ImageNet و بخش‌بندی نمونه‌ها با استفاده از مجموعه‌داده‌ی MS COCO نیز در این ارزیابی لحاظ شده‌اند. در نهایت، هفت مجموعه‌داده‌ی معیار مربوط به سری‌های زمانی نیز برای مقایسه‌ی عملکرد مدل‌ها مورد استفاده قرار گرفته‌اند.

۱.۵ جزئیات مجموعه‌داده‌ها

۱.۱.۵ LRA

در این بخش، توضیحات و جزئیات تکمیلی مربوط به هر یک از مجموعه‌داده‌های LRA که توسط (Tay et al.، ۲۰۲۰) معرفی شده‌اند و همچنین مجموعه‌داده‌ی Speech Commands که توسط (Warden et al.، ۲۰۱۸) ارائه گردیده است، آورده می‌شود. مراحل پیش‌پردازش داده‌ها مطابق با دستورالعمل‌های بیان‌شده توسط (Gu et al.، ۲۰۲۱) و انجام شده است برای تکمیل بحث، در اینجا نیز ارائه می‌شوند.

- ListOps: این مجموعه‌داده نسخه‌ای گسترش‌یافته از داده‌هایی است که توسط (Nangia et al ., 2018) معرفی شده و با نام ListOps شناخته می‌شود. هدف آن، ارزیابی عملیات‌های ریاضی تو در تو مانند min و max بر روی عملوندهای صحیح بین صفر تا نه است. عبارات به‌صورت پیشوندی و درون براکت‌ها نوشته می‌شوند. هدف، محاسبه‌ی نتیجه‌ی عددی عبارت ریاضی است. هر کاراکتر به‌صورت یک بردار one-hot با ۱۷ مقدار یکتا (شامل عملگرها و براکت‌ها) کدگذاری می‌شود. طول دنباله‌ها متغیر است و تا حداکثر ۲۰۰۰ توکن با شاخص ویژه پر می‌شود. یک توکن ویژه برای پایان دنباله افزوده می‌شود. این مجموعه‌داده شامل ۱۰ کلاس مختلف است که نتیجه‌ی عددی عبارات را نشان می‌دهند و دربرگیرنده‌ی ۹۶۰۰۰ داده‌ی آموزشی، ۲۰۰۰ داده‌ی اعتبارسنجی و ۲۰۰۰ داده‌ی آزمایشی است. هیچ نرمال‌سازی‌ای بر روی داده‌ها اعمال نشده است.

- Text : این مجموعه‌داده برگرفته از داده‌ی احساسات iMDB است که توسط (Maas et al ., 2011) معرفی شده است. هدف، طبقه‌بندی نقدهای فیلم به دو دسته‌ی مثبت یا منفی است. نقدها به‌صورت دنباله‌ای از توکن‌های عددی با ۱۲۹ مقدار یکتا که نشان‌دهنده‌ی کاراکترها هستند کدگذاری شده‌اند. دنباله‌ها تا طول حداکثر ۴۰۹۶ توکن پر می‌شوند. این مجموعه شامل ۲۵۰۰۰ داده‌ی آموزشی و ۲۵۰۰۰ داده‌ی آزمایشی است و مجموعه‌ی اعتبارسنجی ندارد. هیچ نرمال‌سازی‌ای بر روی داده‌ها انجام نشده است.

- Retrieval: این مجموعه‌داده بر اساس پیکره‌ی متنی ACL Anthology Network معرفی‌شده توسط (Radev et al ., 2009) ساخته شده است. هدف، تشخیص این است که آیا دو نقل‌قول متنی از نظر معنایی معادل‌اند یا خیر. نقل‌قول‌ها به‌صورت دنباله‌ای از توکن‌های عددی کدگذاری می‌شوند و هر جفت از نقل‌قول‌ها به‌طور جداگانه فشرده‌سازی شده و سپس وارد لایه‌ی نهایی طبقه‌بندی می‌شود. سرِ رمزگشا از نمایش برداری حاصل برای انجام طبقه‌بندی استفاده می‌کند. کاراکترها به‌صورت one-hot با ۹۷ مقدار یکتا کدگذاری می‌شوند. دنباله‌های جفتی ممکن است طول‌های متفاوتی داشته باشند و حداکثر طول آن‌ها ۴۰۰۰ توکن است. این مجموعه شامل ۱۴۷۰۸۶ جفت داده‌ی آموزشی، ۱۸۰۹۰ داده‌ی اعتبارسنجی و ۱۷۴۳۷ داده‌ی آزمایشی است. هیچ نرمال‌سازی‌ای اعمال نشده است.

- Image : این مجموعه‌داده از داده‌های CIFAR-10 معرفی‌شده توسط (Krizhevsky et al ., 2009) و استفاده می‌کند. هدف، طبقه‌بندی تصاویر 32×32 در مقیاس خاکستری از مجموعه‌ی CIFAR-10 در ده کلاس مختلف است. تصاویر به‌صورت بردارهای تک‌بعدی با طول ۱۰۲۴ نمایش داده می‌شوند. مجموعه شامل ۴۵۰۰۰ داده‌ی آموزشی، ۵۰۰۰ داده‌ی اعتبارسنجی و ۱۰۰۰۰ داده‌ی آزمایشی است. مقادیر RGB به مقادیر خاکستری تبدیل شده و برای داشتن میانگین صفر و واریانس واحد نرمال‌سازی می‌شوند. این نرمال‌سازی بر روی کل مجموعه اعمال می‌گردد.

- Pathfinder : این مجموعه‌داده مبتنی بر چالش Pathfinder است که توسط (Linsley et al ., 2018) و معرفی شده است. هدف، تشخیص این است که آیا بین نقاط آغاز و پایان در یک تصویر 32×32 در مقیاس خاکستری، مسیر پیوسته‌ای از خط وجود دارد یا خیر. این مجموعه شامل دو کلاس است که وجود یا عدم وجود مسیر معتبر را نشان می‌دهند. دنباله‌ها طول ثابتی برابر با ۱۰۲۴ دارند. مجموعه شامل ۱۶۰۰۰۰ داده‌ی آموزشی، ۲۰۰۰۰ داده‌ی اعتبارسنجی و ۲۰۰۰۰ داده‌ی آزمایشی است. داده‌ها در بازه‌ی [−۱، ۱] نرمال‌سازی شده‌اند.

- Path-X : این مجموعه نسخه‌ی گسترش‌یافته‌ی چالش Pathfinder است که در آن تصاویر دارای ابعاد 128×128 پیکسل هستند، بنابراین دنباله‌ها شانزده برابر طولانی‌ترند. سایر ویژگی‌ها مشابه چالش اصلی Pathfinder باقی مانده‌اند.

۲.۱.۵ مجموعه‌داده‌های معیار سری‌زمانی چندمتغیره

ارزیابی مدل SSM بر روی هفت مجموعه‌داده‌ی معیار استاندارد که معمولاً برای پیش‌بینی سری‌های زمانی چندمتغیره (Multivariate Time Series Forecasting) به‌کار می‌روند، انجام شده است و عملکرد قدرتمند آن را در مقایسه با طیفی از مدل‌های پیشرفته نشان می‌دهد. در این ارزیابی از هفت مجموعه‌داده‌ی معیار که به‌طور گسترده در حوزه‌ی پیش‌بینی سری‌های زمانی چندمتغیره استفاده می‌شوند، بهره گرفته شده است. این مجموعه‌داده‌ها حوزه‌های مختلفی را پوشش می‌دهند، از جمله برق (Electricity)، آب‌وهوا (Weather)، ترافیک (Traffic)، و چهار مجموعه‌داده از حوزه‌ی پیش‌بینی سری‌های زمانی انرژی (Energy Time Series Forecasting) با نام‌های ETTh1، ETTh2، ETTm1 و ETTm2.

۳.۱.۵ مجموعه‌داده‌های درک ویدیو

برای ارزیابی عملکرد مدل VideoMamba در درک ویدیوهای کوتاه‌مدت و بلندمدت، آزمایش‌هایی بر روی شش مجموعه‌داده‌ی متنوع انجام شده است:

- درک ویدیوهای کوتاه‌مدت: توانایی مدل VideoMamba در وظایف مرتبط با صحنه و زمان با استفاده از دو مجموعه‌داده‌ی پرکاربرد ارزیابی شد:
  - Kinetics-400: این مجموعه‌داده شامل ویدیوهایی با میانگین طول ۱۰ ثانیه است که بر روی کنش‌های مرتبط با صحنه تمرکز دارند.
  - Something-Something V2: این ویدیوها دارای میانگین مدت‌زمان ۴ ثانیه بوده و بر کنش‌های مرتبط با زمان تأکید دارند.

- درک ویدیوهای بلندمدت: برای درک ویدیوهای بلندمدت، ارزیابی‌های دقیقی بر روی سه مجموعه‌داده‌ی جامع انجام شد:
  - Breakfast: این مجموعه‌داده شامل ۱۷۱۲ ویدیو است که ۱۰ فعالیت پیچیده‌ی آشپزی را در مجموع ۷۷ ساعت پوشش می‌دهد.

- COIN: این مجموعه شامل ۱۱۸۲۷ ویدیو در ۱۸۰ وظیفه‌ی رویه‌ای است، با میانگین مدت‌زمان ۲:۳۶ دقیقه که محتوای متنوعی را برای ارزیابی فراهم می‌کند.
- Long-form Video Understanding (LVU): این معیار شامل حدود ۳۰۰۰۰ کلیپ فیلم است که هر یک بین ۱ تا ۳ دقیقه طول دارند. این مجموعه ۹ وظیفه را در سه دسته‌ی اصلی شامل درک محتوا، پیش‌بینی فراداده و تعامل کاربر پوشش می‌دهد.

با انجام ارزیابی‌ها بر روی این مجموعه‌داده‌ها، هدف ما سنجش جامع عملکرد مدل VideoMamba در جنبه‌های مختلف درک ویدیو است؛ از جمله تشخیص صحنه (scene recognition)، استدلال زمانی (temporal reasoning) و درک بلندمدت محتوای ویدیویی پیچیده.

۴.۱.۵ مجموعه‌داده‌های مدل‌های زبانی بزرگ چندوجهی

ما یک ارزیابی جامع از مدل خود را در میان طیف گسترده‌ای از ۸ مجموعه‌داده‌ی معیار انجام دادیم:

۱. **VQA-v2** : این مجموعه‌داده بر ارزیابی توانایی مدل‌ها در درک و استدلال درباره‌ی تصاویر و پرسش‌های همراه آن‌ها تمرکز دارد.

۲. **GQA** : برای ارزیابی درک فضایی و استنتاج چندمرحله‌ای در تصاویر واقعی طراحی شده است.

۳. **ScienceQA-IMG** : شامل پرسش‌های چندگزینه‌ای چندوجهی درباره‌ی موضوعات علمی است و بر استدلال عقل سلیم تأکید دارد.

۴. **TextVQA** : شامل پرسش‌هایی مرتبط با متون درون تصاویر است و توانایی مدل را در شناسایی نویسه‌های نوری (OCR) و استنتاج بررسی می‌کند.

۵. **POPE** : یک معیار برای ارزیابی خطاهای شناسایی اشیاء است که شامل یک وظیفه‌ی طبقه‌بندی دوتایی برای تعیین وجود یا عدم وجود اشیاء می‌باشد.

۶. **MME** : توانایی‌های ادراکی و شناختی مدل را می‌سنجد، از جمله OCR، شناسایی اشیاء، استدلال عقل سلیم، محاسبات عددی، ترجمه‌ی متون و استدلال مبتنی بر کد.

۷. **MMBench** : شامل ۳۰۰۰ پرسش تک‌گزینه‌ای در ۲۰ بُعد مختلف است و از استراتژی CircularEval برای ارزیابی دقیق استفاده می‌کند. در این ارزیابی، پیش‌بینی‌های مدل با پاسخ‌های ChatGPT تطبیق داده می‌شوند.

۸. **MM-Vet** : شامل ۱۶ وظیفه‌ی نوظهور از قابلیت‌های دیداری و زبانی (VL) است؛ از جمله شناسایی، دانش، OCR، آگاهی فضایی، تولید زبان و ریاضیات.

هر یک از این مجموعه‌داده‌ها چالش‌های منحصربه‌فردی را ارائه می‌دهند و جنبه‌های مختلف درک چندوجهی را ارزیابی می‌کنند؛ از شناسایی ابتدایی اشیاء گرفته تا استدلال پیچیده و وظایف استنتاجی میان تصویر و متن.

۲.۵ نتایج

۱.۲.۵ LRA

در این بخش، جدول نتایج از مقالات HGRN و S5 اقتباس و در جدول زیر ترکیب شده است. این نتایج نشان می‌دهند که مدل‌های Transformer از جمله Local Attention، Sparse Attention، LongFormer، LinFormer، Reformer، Sinkhorn، Synthesizer، BigBird، Linear Transformer، Performer، CosFormer، FNet، NystromFormer، Luna، H-Transformer-ID و CCNN عملکرد چندان مطلوبی در معیارهای LRA ندارند. در بسیاری از وظایف مانند ListOps، Text، Retrieval، Image، Pathfinder و Path-X، این مدل‌ها دچار افت عملکرد می‌شوند؛ دلیل اصلی این مسئله، پیچیدگی محاسباتی درجه دوم در مکانیزم توجه (Attention) و نبودِ سوگیری استقرایی (Inductive Bias) در معماری Transformer است. در مقابل، مدل‌های فضای حالت (State Space Models, SSMs) عملکرد بهتری نسبت به ترنسفورمرها در معیارهای LRA دارند. در میان مدل‌های SSM، دو مدل S5 و Mega بالاترین عملکرد را در تمامی وظایف ارائه داده‌اند. مدل S4، که یکی از مدل‌های پیشگام در این حوزه محسوب می‌شود، توسط نسخه‌های بهبود یافته‌ی خود مانند DSS، S4ND و Liquid-S4 پشت سر گذاشته شده است. مدل‌های TNN، HGRN و LRU عملکردی متوسط دارند اما همچنان از S4 و گونه‌های آن بهتر عمل می‌کنند. با این حال، علت دقیق اینکه چرا مدل‌های S5 و Mega در میان سایر SSMها بهترین عملکرد را دارند، هنوز به‌طور کامل مشخص نیست. برای درک بهتر این مسئله، می‌توان این مدل‌ها را از دیدگاه قابلیت تبیین (Explainability) مورد تحلیل قرار داد، که انجام چنین پژوهشی به کارهای آینده موکول می‌شود.

۲.۲.۵ حوزه‌ی زبانی

در این بخش، نتایج سه معیار اصلی در حوزه‌ی زبان ارائه می‌شوند که شامل GLUE، WikiText و Pile هستند. جدول (۲) نتایج معیار GLUE را نشان می‌دهد که ترکیبی از شش وظیفه‌ی اصلی است: MNLI، QNLI، QQP، SST2، MRPC و CoLA. این جدول به چهار دسته تقسیم می‌شود: مدل‌های مبتنی بر توجه (Attention-based Transformers)، شبکه‌های مبتنی بر MLP، ترنسفورمرهای مبتنی بر FFT و مدل‌های فضای حالت (SSMs). نتایج نشان می‌دهند که مدل‌های SSM در مقایسه با سه معماری دیگر دارای شکاف عملکردی هستند، در حالی که شبکه‌های مبتنی بر MLP و ترنسفورمرهای مبتنی بر FFT امتیازات بهتری نسبت به SSMها کسب کرده‌اند. در میان تمام مدل‌ها، ترنسفورمرهای مبتنی بر توجه بالاترین عملکرد را در این معیار دارند. تنها استثنا، شبکه‌ی عصبی Toeplitz Neural Network (TNN) است که از تمام مدل‌ها، از جمله ترنسفورمرهای مبتنی بر توجه، عملکرد بهتری دارد. به‌طور مشابه، معیار بعدی در ارزیابی حوزه‌ی زبان، مجموعه‌داده‌ی WikiText است. جدول (۱) امتیازهای Perplexity مربوط به معماری‌ها و مدل‌های مختلف را نشان می‌دهد. این جدول از مقالات TNN و HGRN اقتباس و نتایج در آن ترکیب شده‌اند. امتیاز Perplexity توانایی مدل را در پیش‌بینی واژه‌ی بعدی می‌سنجد؛ هرچه مقدار آن کمتر باشد، عملکرد مدل بهتر است. این جدول شامل نتایج مجموعه‌داده‌های اعتبارسنجی و آزمون است و سه گروه معماری مدل‌های مبتنی بر توجه، شبکه‌های مبتنی بر MLP و مدل‌های فضای حالت (SSM) را مقایسه می‌کند. نتایج نشان می‌دهند که مدل‌های مبتنی بر توجه همچنان عملکرد برتری نسبت به سایر معماری‌ها دارند، با این حال مدل HGRN از میان مدل‌های SSM توانسته است از همه‌ی مدل‌های دیگر بهتر عمل کرده و پایین‌ترین مقدار Perplexity را به‌دست آورد. آخرین معیار در حوزه‌ی زبان، مجموعه‌داده‌ی Pile است. Pile یک مجموعه‌داده‌ی بسیار بزرگ به حجم ۸۲۵ گیگابایت است که از ترکیب چندین مجموعه‌داده‌ی کوچک‌تر تشکیل شده است،

از جمله PubMed Central، ArXiv، GitHub، FreeLaw Project، Stack Exchange، US Patent and Trademark Office، PubMed، Ubuntu IRC، HackerNews، YouTube، PhilPapers و NIH ExPorter. نتایج به‌دست‌آمده از جدول (۱.۴) و سایر منابع نشان می‌دهند که مدل HGRN تنها مدل از نوع SSM است که عملکردی قابل مقایسه با ترنسفورمرهای مبتنی بر توجه دارد به‌ویژه در شرایطی که این مدل‌ها بر روی مجموعه‌داده‌های Pile آموزش دیده باشند.

۳.۲.۵ حوزه‌ی تصویر

ارزیابی انجام‌شده بر روی مجموعه‌داده‌ی ImageNet-1K عملکرد معماری‌های مختلف در حوزه‌ی بینایی را در دسته‌بندی‌های گوناگون و با پیچیدگی‌های محاسباتی متفاوت نشان می‌دهد، همان‌گونه که در جدول (۵) گزارش شده است. در میان شبکه‌های کانولوشنی (CNNs)، دو مدل RegNetY-8G و RegNetY-16G به‌ترتیب با دقت‌های ۸۱.۷٪ و ۸۲.۹٪ برجسته هستند. این نتایج اثربخشی مدل‌های فضای حالت (SSM) را در استخراج نمایش‌های تصویری تأیید می‌کنند؛ به‌ویژه مدل‌های LocalVMamba-S و PlainMamba-L3 که عملکردی رقابتی از خود نشان داده‌اند. افزون بر این، مدل‌های ترنسفورمری مانند Swin-T و EffNet-B4 نیز عملکرد قابل‌توجهی دارند و دقت‌هایی به‌ترتیب ۸۱.۳٪ و ۸۲.۹٪ کسب کرده‌اند، در حالی که RegNetY-8G و RegNetY-16G همچنان به‌عنوان بهترین مدل‌های مبتنی بر کانولوشن شناخته می‌شوند. در میان ترنسفورمرها، مدل‌های SpectFormer-H-B، SVT-H-B و SCT-H-B عملکرد چشمگیری از خود نشان داده‌اند و به‌ترتیب به دقت‌های ۸۵.۱٪، ۸۵.۲٪ و ۸۵.۲٪ در معیار Top-1 Accuracy دست یافته‌اند.

در میان مدل‌های SSM، مدل LocalVMamba-S عملکرد قوی‌ای با دقت ۸۳.۷٪ داشته است و مدل PlainMamba-L3 نیز با دقت ۸۳.۲٪ نتایجی رقابتی ارائه کرده است. همچنین مدل‌های Vim-Ti و VMamba-T نیز با دقت‌های ۷۶.۱٪ و ۸۲.۲٪ عملکرد قابل‌توجهی دارند. در میان مدل‌های SSM، مدل SiMBA-S (MLP) به‌عنوان یکی از گزینه‌های قدرتمند ظاهر شده است و به دقت ۸۴.۰٪ دست یافته است، در حالی که دو مدل SiMBA-B (MLP) و SiMBA-L (EinFFT) نیز به‌ترتیب دقت‌های ۸۴.۷٪ و ۸۳.۹٪ را در معیار Top-1 Accuracy به‌دست آورده‌اند. این نتایج نشان می‌دهند که رویکردهای متنوع معماری، از CNN تا Transformer و SSM، در سطوح مختلف پیچیدگی محاسباتی قادر به ارائه‌ی عملکردی مؤثر در وظایف شناسایی تصویر هستند. جدول (۱۲) عملکرد پیشرفته‌ی مدل‌های فضای حالت ساختاریافته (Structured State Space Models, SSMs) را بر روی مجموعه‌داده‌ی ImageNet-1K برای وظایف طبقه‌بندی تصویر نشان می‌دهد. این مدل‌های SSM بر اساس میزان پیچیدگی محاسباتی آن‌ها، که بر حسب GFLOPs اندازه‌گیری می‌شود، در سه گروه کوچک (Small)، پایه (Base) و بزرگ (Large) دسته‌بندی شده‌اند. از میان مدل‌های برجسته می‌توان به HyenaViT-B، S4ND-ViT-B، Vim-S، HGRN-S، PlainMamba-L3، Mamba-2D-B، VMamba-S، LocalVMamba-S، SiMBA-L (Monarch)، ViM2-B و SiMBA-L (EinFFT) اشاره کرد که دقت Top-1 آن‌ها در بازه‌ی ۷۸.۵٪ تا ۸۴.۵٪ قرار دارد. این مدل‌ها ضمن برخورداری از مقادیر متفاوتی از پارامترها و پیچیدگی‌های محاسباتی، عملکردی رقابتی از خود نشان می‌دهند و کارایی بالای مدل‌های SSM را در وظایف طبقه‌بندی تصویر با بهره‌وری محاسباتی بهبودیافته به نمایش می‌گذارند.

Top-1 acc. (%)	FLOPs	Param.	Image Size	Method
		SSMs		
78.5	–	88M	224^2	HyenaViT-B
80.4	–	89M	224^2	S4ND-ViT-B
72.29	–	6.4M	–	TNN-T
79.20	–	23.4M	–	TNN-S
76.1	–	7M	224^2	Vim-Ti
80.5	–	26M	224^2	Vim-S
74.40	–	6.1M	–	HGRN-T
80.09	–	23.7M	–	HGRN-S
77.9	3.0G	7M	224^2	PlainMamba-L1
81.6	8.1G	25M	224^2	PlainMamba-L2
82.3	14.4G	50M	224^2	PlainMamba-L3
81.7	–	24M	224^2	Mamba-2D-S
83.0	–	92M	224^2	Mamba-2D-B
82.2	5.6G	24M	224^2	VMamba-T
83.5	11.2G	44M	224^2	VMamba-S
83.2	18.0G	75M	224^2	VMamba-B
82.7	5.7G	26M	224^2	LocalVMamba-T
83.7	11.4G	50M	224^2	LocalVMamba-S
81.1	3.6G	16.3M	224^2	SiMBA-S (Monarch)
82.6	6.3G	28.9M	224^2	SiMBA-B (Monarch)
83.8	10.7G	40.0M	224^2	SiMBA-L (Monarch)
82.7	–	10M	224^2	ViM2-T
83.7	–	43M	224^2	ViM2-S
83.9	–	74M	224^2	ViM2-B
81.7	2.4G	15.3M	224^2	SiMBA-S (EinFFT)
83.5	5.2G	22.8M	224^2	SiMBA-B (EinFFT)
84.4	9.6G	16.6M	224^2	SiMBA-L (EinFFT)
84.0	5.0G	26.5M	224^2	SiMBA-S (MLP)
84.7	9.0G	40.0M	224^2	SiMBA-B (MLP)

جدول ۱۲: عملکرد مدل‌های SSM در معیار ImageNet-1K. این جدول عملکرد مدل‌های مختلف SSM را برای وظایف شناسایی تصویر بر روی مجموعه‌داده‌ی ImageNet-1K نشان می‌دهد. مدل‌های بینایی در سه گروه بر اساس میزان GFLOPs تقسیم شده‌اند: کوچک (Small)، پایه (Base) و بزرگ (Large). بازه‌های GFLOP عبارت‌اند از: کوچک ($\text{GFLOPs} < 5$)، پایه ($5 \leq \text{GFLOPs} < 10$) و بزرگ ($10 \leq \text{GFLOPs} < 30$).

۴.۲.۵ کاربردهای چندوجهی

مدل VL-Mamba عملکرد قابل‌توجهی را در میان معیارهای مختلف از خود نشان داده است، همان‌گونه که در جدول (۱۳) گزارش شده است. با استفاده از داده‌های آموزشی چندوجهی مشابه، VL-Mamba نسبت به مدل‌های SQAI (به‌ترتیب ۶۵.۴ در مقابل ۶۱.۲)، VQAT (۴۸.۹ در مقابل ۴۷.۵) و MME (۱۳۶۹.۶ در مقابل ۱۲۸۸.۹) عملکرد بهتری دارد. شایان توجه است که با وجود برخورداری از توکن‌های پیش‌آموزش‌یافته‌ی کمتر (۶۲۷ میلیارد) در مقایسه با مدل MobileVLM که از ستون فقرات MobileLLaMA با ۱.۳ تریلیون توکن استفاده می‌کند، مدل VL-Mamba همچنان عملکردی برتر دارد. برای نمونه، در مقایسه با مدل LLaVA-Phi که از مدل زبانی Phi-2-2.7B بهره می‌برد، VL-Mamba در معیارهای VQA-v2 (به‌ترتیب ۷۶.۶ در مقابل ۷۱.۴)، MME (۱۳۶۹ در مقابل ۱۳۵۵.۱) و MM-Vet (۳۲.۶ در مقابل ۲۸.۹) نتایج بهتری کسب کرده است. این یافته‌ها کارایی بالای VL-Mamba را در وظایف یادگیری چندوجهی نشان می‌دهند و بر پتانسیل بالای بهره‌گیری از مدل‌های فضای حالت (State Space Models) در چنین کاربردهایی تأکید می‌کنند.

MM-Vet	MMB	MME	POPE	VQA^T	SQA^I	GQA	VQA^v2	IT	PT	LLM	Method
22.4	–	1293.8	85.3	42.5	61.0	41.0	41.0	–	129M	Vicuna-13B	BLIP-2
–	23	581.7	–	–	–	32.2	–	5K	5M	Vicuna-7B	MiniGPT-4
26.2	36	–	78.9	50.1	60.5	49.2	–	1.2M	129M	Vicuna-7B	InstructBLIP
25.6	–	1212.8	–	50.7	63.1	49.5	–	1.2M	129M	Vicuna-13B	InstructBLIP
–	58.8	–	–	–	–	–	77.4	5.5M	600K	Vicuna-13B	Shikra
24.6	48.3	1292.3	–	–	–	–	–	–	–	LLaMA-7B	Otter
–	49.4	967.3	–	–	–	–	–	102K	2.1M	LLaMA-7B	mPLUG-Owl
–	48.2	–	–	25.9	–	38.4	50.9	1M	353M	LLaMA-7B	IDEFICS-9B
	–	54.5	–	30.9	–	45.2	60.0	1M	353M	LLaMA-65B	IDEFICS-80B
–	38.2	–	–	63.8	67.1	59.3	78.8	5.0M	1.4B	Qwen-7B	Qwen-VL
–	60.6	1487.5	–	61.5	68.2	57.5	78.2	5.0M	1.4B	Qwen-7B	Qwen-VL-Chat
30.5	64.3	1510.7	85.9	58.2	66.8	62.0	78.5	665K	558K	Vicuna-7B	LLaVA-1.5
35.4	67.7	1531.3	85.9	61.3	71.6	63.3	80.0	665K	558K	Vicuna-13B	LLaVA-1.5
28.9	59.8	1335.1	85.0	48.6	68.4	–	71.4	665K	558K	Phi-2-2.7B	LLaVA-Phi
–	59.6	1288.9	84.9	47.5	61.2	59.0	–	665K	558K	MobileLLaMA-2.7B	MobileVLM-3B
–	–	–	88.0	46.0	–	58.5	75.9	–	–	Mamba-2.8B	Cobra
32.6	57.0	1369.6	84.4	48.9	65.4	56.2	76.6	665K	558K	Mamba LLM-2.8B	VL-Mamba

جدول ۱۳: مقایسه با روش‌های پیشرفته (SoTA) در ۸ معیار ارزیابی. نام معیارها به‌دلیل محدودیت فضا به‌صورت اختصاری آورده شده‌اند: VQA-v2؛ GQA؛ SQAI: ScienceQA-IMG؛ VQAT: TextVQA؛ POPE؛ MME؛ MMB: MMBench؛ MM-Vet . نمادهای PT و IT به‌ترتیب بیانگر تعداد نمونه‌ها در مراحل پیش‌آموزش (Pretraining) و تنظیم دستورالعمل (Instruction Tuning) هستند. این جدول از مقاله‌ی VL-Mamba اقتباس شده است.

۵.۲.۵ پیش‌بینی سری‌های زمانی

ارزیابی مدل SSM که بر روی هفت مجموعه‌داده‌ی معیار استاندارد متداول در حوزه‌ی پیش‌بینی سری‌های زمانی چندمتغیره (Multi-variate Time Series Forecasting) انجام شده است، عملکرد قدرتمند آن را در مقایسه با طیفی از مدل‌های پیشرفته‌ی روز نشان می‌دهد. در این ارزیابی، مدل SiMBA با چندین مدل پیشرفته‌ی دیگر مقایسه شده است، از جمله روش‌های مبتنی بر Transformer مانند PatchTST، CrossFormer، FEDFormer، ETSFormer، PyraFormer و AutoFormer. همچنین، مقایسه‌هایی با مدل‌های مبتنی بر CNN مانند TimeNet، رویکردهای مبتنی بر گراف مانند MTGNN، و مدل‌های مبتنی بر MLP مانند DLinear نیز انجام شده است. عملکرد مدل SiMBA و سایر مدل‌های مقایسه‌ای با استفاده از معیارهای متداول در وظایف پیش‌بینی سری‌های زمانی سنجیده شده است. این معیارها معمولاً شامل Mean Squared Error (MSE) و Mean Absolute Error (MAE) هستند که معیارهای استانداردی برای اندازه‌گیری دقت مدل‌های پیش‌بینی محسوب می‌شوند. نتایج نشان دادند که SiMBA در تمام مجموعه‌داده‌های ارزیابی‌شده، از نظر معیارهای MSE و MAE عملکردی بهتر از سایر مدل‌های پیشرفته‌ی موجود داشته است. این امر نشان می‌دهد که SiMBA توانایی بالایی در شناسایی الگوهای زمانی و انجام پیش‌بینی‌های دقیق در وظایف مختلف سری زمانی و حوزه‌های گوناگون دارد. عملکرد برتر SiMBA نشان‌دهنده‌ی سازگاری و کارایی بالای آن در مواجهه با چالش‌های متنوع پیش‌بینی سری‌های زمانی است و آن را به یکی از مدل‌های شاخص در این زمینه تبدیل کرده است. به‌طور کلی، ارزیابی گسترده‌ی SiMBA در میان مجموعه‌داده‌های مختلف و مقایسه با مدل‌های پیشرفته‌ی روز، عملکرد قدرتمند و مؤثر آن را در حل طیف وسیعی از وظایف پیش‌بینی سری‌های زمانی تأیید می‌کند. یافته‌ها نشان می‌دهند که SiMBA مدلی امیدبخش برای کاربردهای عملی در حوزه‌های مختلف است که نیاز به پیش‌بینی‌های دقیق و قابل اعتماد سری‌های زمانی دارند.

۶.۲.۵ حوزه‌ی ویدیو

جدول (۸) مقایسه‌ای از عملکرد مدل‌های مختلف بر روی مجموعه‌داده‌ی مرتبط با صحنه Kinetics-400 ارائه می‌دهد. این جدول شامل معماری‌هایی مانند شبکه‌های کانولوشنی (CNN)، ترنسفورمرها (Transformers) و مدل‌های فضای حالت (State Space Models, SSM) است و ویژگی‌هایی همچون نوع داده‌های ورودی، تعداد پارامترها و میزان عملیات ممیز شناور (FLOPs) را نیز نشان می‌دهد. برای هر نوع معماری، مدل‌های مختلفی فهرست شده‌اند که مشخص می‌کنند آیا این مدل‌ها به‌صورت نظارت‌شده (Supervised) یا خودنظارتی (Self-Supervised) آموزش دیده‌اند. مدل‌ها بر اساس معیارهای دقت Top-1 و Top-5 بر روی مجموعه‌داده‌ی Kinetics-400 مورد ارزیابی قرار گرفته‌اند. علاوه بر این، در جدول مشخص شده است که آیا معماری مدل‌ها از نوع ایزوتروپیک (بدون لایه‌های Downsampling) است، و اینکه آیا از داده‌های اضافی یا مدل‌های معلم پیش‌آموزش‌یافته استفاده شده است یا خیر. مدل‌های CNN نظارت‌شده مانند SlowFast، X3D-M و X3D-XL عملکرد رقابتی بالایی از خود نشان داده‌اند. مدل‌های ترنسفورمری مانند Swin-T و Swin-B نیز عملکرد قدرتمندی دارند، به‌ویژه زمانی‌که با داده‌های اضافی از مجموعه‌ی IN-21K آموزش دیده‌اند. مدل‌های ترکیبی CNN+Transformer مانند MViTv1-B و UniFormer-B نیز نتایج امیدوارکننده‌ای ارائه کرده‌اند. در میان مدل‌های فضای حالت (SSM)، مدل VideoMamba عملکردی رقابتی از خود نشان داده است، به‌ویژه نسخه‌ی VideoMamba-M که در معیارهای دقت Top-1 و Top-5 از سایر مدل‌ها پیشی گرفته است، خصوصاً هنگامی که با داده‌های اضافی از مجموعه‌ی CLIP-400M آموزش دیده است. جدول (۹) مقایسه‌ای از عملکرد مدل‌های مختلف بر روی مجموعه‌داده‌ی زمانی SthSth V2 ارائه می‌دهد. این جدول شامل

معماری‌هایی مانند شبکه‌های کانولوشنی (CNN)، ترنسفورمرها (Transformers) و مدل‌های فضای حالت (State Space Models, SSM) است و اطلاعات مربوط به مدل‌ها، ویژگی‌های داده‌ی ورودی، تعداد پارامترها و میزان عملیات ممیز شناور (FLOPs) را در بر می‌گیرد. مدل‌های CNN مانند SlowFast_{R101}، CT-Net_{R50} و TDN_{R50} عملکرد متوسطی بر روی مجموعه‌داده‌ی SthSth V2 دارند. مدل ترنسفورمری Swin-B عملکردی رقابتی نسبت به مدل‌های CNN ارائه می‌دهد. مدل‌های ترکیبی CNN+Transformer مانند MViTv1-B و UniFormer-B نتایج امیدوارکننده‌ای نشان داده‌اند، به‌طوری‌که UniFormer-B از نظر دقت Top-1 عملکرد بهتری نسبت به سایر مدل‌ها دارد. مدل‌های ترنسفورمری مانند TimeSformer-HR و ViViT-L نیز عملکرد قابل‌قبولی بر روی این مجموعه‌داده ارائه می‌دهند. مدل‌های فضای حالت (SSM) مانند VideoMamba عملکردی رقابتی از خود نشان داده‌اند، به‌ویژه مدل VideoMamba-M که بالاترین دقت Top-1 را در میان مدل‌های SSM به‌دست آورده است. مدل‌های ترنسفورمری خودنظارتی مانند $\text{VideoMAE-B}_{2400e}$ و UMT-B_{800e} نیز عملکرد مطلوبی دارند، به‌ویژه زمانی که از مدل معلم پیش‌آموزش‌یافته‌ی CLIP-400M استفاده می‌کنند. مدل $\text{VideoMamba-M}_{800e}$ که با داده‌های CLIP-400M آموزش دیده است، بالاترین دقت Top-1 را در میان تمام مدل‌های SSM کسب کرده است. جدول (۱۰) عملکرد روش‌های مختلف را در مجموعه‌داده‌ی LVU (Large-scale Video Understanding) با استفاده از معیارهای ارزیابی گوناگون مرتبط با محتوا، فراداده (Metadata) و تعامل کاربر مورد مقایسه قرار می‌دهد. روش‌هایی مانند VideoBERT، Object Trans.، LST، Performer، Orthoformer، ViS4mer و VideoMamba_{f32} از معماری‌ها و ستون‌فقرات متفاوتی برای درک محتوای ویدیویی استفاده می‌کنند. مدل VideoMamba_{f32} بالاترین امتیازها را در معیارهای مرتبط با محتوا، به‌ویژه در زمینه‌ی درک صحنه، کسب کرده است. درک فراداده شامل ویژگی‌هایی مانند کارگردان، ژانر، نویسنده و سال انتشار است. مدل VideoMamba_{f32} در اغلب معیارهای مرتبط با فراداده عملکردی بهتر از سایر روش‌ها دارد که نشان‌دهنده‌ی توانایی آن در استخراج اطلاعات مفید از فراداده‌های ویدیویی است. برخی روش‌ها مانند VideoMamba_{f32} از رویکردهای انتها به انتها (End-to-End, e2e) استفاده می‌کنند که نیازی به استخراج جداگانه‌ی ویژگی‌ها ندارند. این رویکرد یکپارچه به درک بهتر و استفاده‌ی مؤثر از هر دو منبع محتوا و فراداده برای پیش‌بینی تعامل کاربر کمک می‌کند. جدول (۱۵) مقایسه‌ای از عملکرد روش‌های مختلف بر روی مجموعه‌داده‌های Breakfast (BF) و COIN ارائه می‌دهد، با تمرکز بر رویکردهای انتها به انتها (End-to-End, e2e) و معماری‌های مختلف ستون‌فقرات (Backbone). روش‌های موجود مانند Timeception، VideoGraph و GHRM از جمله رویکردهایی هستند که از معماری‌های ستون‌فقرات متفاوت و مجموعه‌داده‌های پیش‌آموزش‌یافته برای دستیابی به عملکرد مطلوب در مجموعه‌داده‌های BF و COIN استفاده کرده‌اند. در میان این روش‌ها، مدل Distant Supervision بالاترین عملکرد را نشان می‌دهد؛ این مدل از TimeSformer با مکانیزم توجه (Attention) و پیش‌آموزش مبتنی بر پایگاه دانش (Knowledge Base, KB) بهره می‌برد. مدل Turbo_{f32} یک رویکرد انتها به انتها (e2e) با استفاده از ستون‌فقرات VideoMAE-B است که عملکردی رقابتی را در هر دو مجموعه‌داده ارائه می‌دهد. مدل‌های VideoMamba_{f32} و VideoMamba_{f64} از نسخه‌های مختلف ستون‌فقرات VideoMamba شامل Ti، S و M استفاده کرده و دقت بالایی را در هر دو مجموعه‌داده به‌دست آورده‌اند. نسخه‌های VideoMamba_{f32} و VideoMamba_{f64} که از ستون‌فقرات با پیش‌آموزش ماسک‌دار (با علامت †مشخص شده‌اند) استفاده می‌کنند، عملکردی حتی بالاتر دارند؛ به‌ویژه مدل VideoMamba_{f32} که بالاترین دقت را در هر دو مجموعه‌داده به‌دست آورده است. ارزیابی نسخه‌های مختلف ستون‌فقرات VideoMamba (شامل Ti، S و M) با دقت‌های متفاوت (f32 و f64) نشان می‌دهد که دقت بالاتر (f64) معمولاً منجر به بهبود عملکرد می‌شود. در میان تمام نسخه‌ها، مدل

VideoMamba-M با دقت f32 و ستون‌فقرات پیش‌آموزش‌یافته‌ی ماسک‌دار، بالاترین دقت را در مجموعه‌داده‌ی BF کسب کرده است، در حالی‌که مدل VideoMamba-M با دقت f64 بهترین عملکرد را در مجموعه‌داده‌ی COIN از خود نشان داده است.

Prophet		ARIMA		DeepAR		LSTMa		Reformer		LogTrans		Informer†		Informer		S4		Methods
MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	Metric
																		$ETTh_1$
0.275	0.115	0.284	0.108	0.280	0.107	0.272	0.114	0.389	0.222	0.259	0.103	0.246	0.092	0.247	0.098	**0.191**	**0.061**	24
0.330	0.168	0.424	0.175	0.327	0.162	0.358	0.193	0.445	0.284	0.328	0.167	0.322	0.161	0.319	0.158	0.220	0.079	48
1.820	0.549	0.759	0.401	0.552	0.445	0.402	0.239	0.708	0.496	0.371	0.230	0.375	0.210	0.346	0.183	0.258	0.104	168
1.820	1.549	0.959	0.468	0.552	0.445	0.698	0.590	1.124	1.860	0.393	0.230	0.369	0.215	0.387	0.222	0.229	0.080	336
3.253	2.735	0.766	0.659	0.707	0.658	0.768	0.683	1.436	2.112	0.463	0.273	0.421	0.257	0.435	0.269	0.271	0.116	720
																		$ETTh_2$
0.381	0.199	0.445	3.554	0.263	0.098	0.307	0.155	0.437	0.263	0.255	0.102	0.241	0.099	**0.240**	**0.093**	0.234	0.095	24
0.462	0.304	0.474	3.190	0.341	0.163	0.341	0.163	0.545	0.458	0.348	0.169	0.317	0.159	**0.314**	**0.155**	0.346	0.191	48
1.068	2.145	0.595	2.800	0.414	0.255	0.514	0.385	0.879	1.029	0.422	0.246	0.390	0.235	0.389	0.232	0.333	0.167	168
2.543	2.096	0.738	2.573	0.687	0.604	0.606	0.558	1.223	1.668	0.437	0.267	0.423	0.258	0.417	0.263	0.361	0.189	336
4.664	3.355	1.044	2.878	0.580	0.429	0.681	0.640	1.721	2.303	0.493	0.303	0.424	0.285	0.431	0.277	0.358	0.187	720
																		$ETTm_1$
0.290	0.120	0.206	0.090	0.243	0.091	0.233	0.121	0.228	0.095	0.202	0.065	0.160	0.034	0.137	0.030	**0.117**	**0.024**	24
0.305	0.133	0.306	0.179	0.362	0.219	0.411	0.305	0.390	0.249	0.220	0.078	0.194	0.066	0.203	0.069	0.174	0.051	48
0.703	0.194	0.641	0.379	0.498	0.364	0.442	0.287	0.767	0.629	0.386	0.149	0.324	0.142	0.372	0.194	0.229	0.086	96
0.574	0.452	0.558	0.462	0.795	0.948	0.584	0.524	1.245	1.108	0.572	0.411	0.548	0.409	0.554	0.401	0.327	0.160	288
1.174	2.747	0.697	0.639	1.352	2.437	0.873	1.064	1.528	1.793	0.702	0.598	0.665	0.519	0.644	0.512	0.466	0.292	672
																		Weather
0.433	0.302	0.355	0.219	0.274	0.128	0.254	0.131	0.401	0.231	0.279	0.136	0.256	0.119	**0.251**	**0.117**	0.254	0.125	24
0.536	0.445	0.409	0.273	0.343	0.190	0.324	0.243	0.423	0.283	0.356	0.206	0.316	0.185	**0.318**	**0.178**	0.305	0.181	48
1.142	2.441	0.599	0.503	0.451	0.294	0.444	0.341	0.634	0.654	0.439	0.309	0.404	0.269	0.398	0.266	0.333	0.198	168
2.468	2.451	0.994	0.530	0.644	0.588	0.554	0.454	1.093	1.792	0.484	0.330	0.431	0.302	0.416	0.297	0.417	0.300	336
1.144	3.859	0.943	1.062	0.596	0.499	0.809	0.866	1.534	2.087	0.499	0.388	0.471	0.361	0.466	0.359	0.375	0.245	720
																		ECL
0.595	0.524	0.764	0.879	**0.357**	**0.204**	0.539	0.493	0.884	0.971	0.429	0.280	0.368	0.238	0.359	0.239	0.350	0.222	48
1.273	2.275	0.873	1.032	0.436	0.315	0.655	0.723	1.587	1.671	0.529	0.454	0.514	0.442	0.503	0.447	0.421	0.331	168
3.079	2.246	0.836	1.116	0.519	0.414	0.896	1.212	2.196	3.528	0.563	0.514	0.552	0.501	0.528	0.489	0.422	0.328	336
4.145	4.243	0.933	1.251	0.595	0.563	0.966	1.511	4.047	4.891	0.609	0.558	0.578	0.543	0.571	0.540	0.494	0.428	720
4.264	6.901	0.982	1.370	0.683	0.657	1.006	1.545	5.105	7.019	0.645	0.624	0.638	0.594	0.608	0.582	0.497	0.432	960

جدول ۱۴: نتایج پیش‌بینی سری‌زمانی تک‌متغیره با دنباله‌های بلند بر روی چهار مجموعه‌داده (پنج حالت). این جدول از مقاله‌ی S4 اقتباس شده است.

COIN Top-1	BF Top-1	Pretraining Dataset	Neck Type	Backbone	e2e	Method
–	71.3	IN-1K+K400	Conv.	3D-ResNet	✗	Timeception
–	69.5	IN-1K+K400	Conv.+Atten.	I3D	✗	VideoGraph
–	75.5	IN-1K+K400	Graph Conv.	I3D	✗	GHRM
90.0	**89.9**	IN-21K+HTM	Atten. w/ KB	TimeSformer	✗	Distant Supervision
88.4	88.2	IN-21K+K600	SSM	Swin-B	✗	ViS4mer
82.3	86.8	K400	–	VideoMAE-B	✓	Turbo_{f32}
87.5	91.3	K400+HTM-AA	–	VideoMAE-B	✓	Turbo_{f64}
86.2	94.3	K400	–	VideoMamba-Ti	✓	VideoMamba_{f32}
87.0	94.3	K400	–	VideoMamba-Ti	✓	VideoMamba_{f64}
88.4	95.3	K400	–	VideoMamba-S	✓	VideoMamba_{f32}
88.7	97.4	K400	–	VideoMamba-S	✓	VideoMamba_{f64}
88.3	94.8	K400	–	VideoMamba-M	✓	VideoMamba_{f32}
89.5	95.8	K400	–	VideoMamba-M	✓	VideoMamba_{f64}
89.6	**97.9**	K400	–	VideoMamba-M†	✓	$^{\dagger}\text{VideoMamba}_{f32}$
90.4	96.9	K400	–	VideoMamba-M†	✓	$^{\dagger}\text{VideoMamba}_{f64}$

جدول ۱۵: مقایسه با جدیدترین روش‌های پیشرفته در مجموعه‌داده‌های Breakfast و COIN. عبارت "e2e" به روش‌های انتها به انتها اشاره دارد که نیازی به استخراج ویژگی‌های جداگانه ندارند.
علامت "†" نشان‌دهنده‌ی ستون‌فقرات با پیش‌آموزش ماسک‌دار است. این جدول از مقاله‌ی VideoMamba اقتباس شده است.

۶ نتیجه‌گیری

این مقاله مروری جامع بر مدل‌های فضای حالت (State Space Models, SSMs) برای پردازش توالی‌ها ارائه داده است شامل بررسی نحوه‌ی تکامل آن‌ها از شبکه‌های بازگشتی (RNNs) تا رقابت کنونی آن‌ها با ترنسفورمرها در حوزه‌های مختلفی مانند زبان، بینایی، سری‌های زمانی، ویدیو و صوت. اگرچه هنوز در برخی وظایف مانند کپی‌کردن و بازیابی اطلاعات از زمینه، ترنسفورمرها عملکرد بهتری نسبت به SSMها دارند، اما مدل‌های فضای حالت توانسته‌اند فاصله‌ی عملکردی خود را با ترنسفورمرهای پیشرفته کاهش دهند. در این مطالعه، مدل‌های SSM به سه دسته‌ی اصلی تقسیم شدند: Structured (ساختاریافته)، Gated (دروازه‌دار) و Recurrent (بازگشتی). در هر دسته، مدل‌های بنیادی شناسایی شده و نوآوری‌های کلیدی آن‌ها مورد بحث قرار گرفته است. علاوه بر این، عملکرد مدل‌های SSM در بنچمارک‌های مختلف گردآوری و با ترنسفورمرهای پیشرفته در حوزه‌های گوناگون مقایسه شد. این مقایسه، چشم‌اندازهای پژوهشی تازه‌ای را در زمینه‌ی SSMها گشوده و شکاف عملکردی آن‌ها با ترنسفورمرها را در چندین حوزه کاهش داده است. به عنوان مثال، مدل SiMBA با ترکیب معماری Transformer و Mamba، به نتایج پیشرفته‌ای در مجموعه‌داده‌های استاندارد سری‌های زمانی و بینایی دست یافته است. به همین ترتیب، سایر مدل‌های بنیادی SSM نیز می‌توانند با ترنسفورمرها ترکیب شوند تا عملکردی فراتر از مدل‌های پیشرفته‌ی فعلی ارائه دهند. با این حال، مقیاس‌پذیری مدل‌های SSM به اندازه‌های بزرگ شبکه همچنان یک چالش باز محسوب می‌شود به‌ویژه در مدل Mamba که در مقیاس بزرگ با مشکلات پایداری روبه‌رو است. پایداری مدل‌های فضای حالت در مقیاس‌های بزرگ، به‌خصوص در بینایی ماشین، هنوز یک مسئله‌ی تحقیقاتی حل‌نشده است. یکی دیگر از زمینه‌های تحقیقاتی باز، بهبود توانایی مدل‌ها در انجام وظایف پیشرفته‌ی یادگیری درون‌متنی (In-Context Learning) است هرچند که ترنسفورمرها و مدل‌های فضای حالت در برخی از این وظایف آموزش دیده‌اند، اما هنوز فرصت‌های پژوهشی قابل‌توجهی برای توسعه‌ی آن‌ها وجود دارد.