A View of Mamba

杨泽康

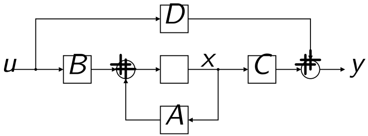
中科院计算技术研究所

2024 年 4 月 19 日

State Space Model

$\mu(t)$ 为输入变量，$x(t)$ 为隐变量，$y(t)$ 为输出变量，$\dot{x}(t)$ 为隐变量的梯度

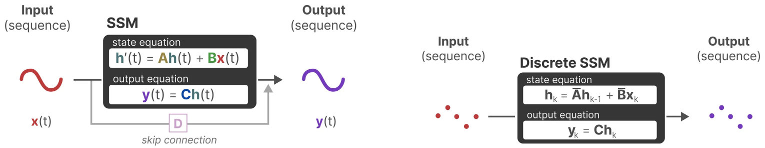
$$\dot{x}(t) = A(t)x(t) + B(t)\mu(t)$$
$$y(t) = C(t)x(t) + D(t)\mu(t)$$



System type	State-space model
Continuous time-invariant	$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$ $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)$
Continuous time-variant	$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t)$ $\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t)$
Explicit discrete time-invariant	$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k)$ $\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k)$
Explicit discrete time-variant	$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k)$ $\mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \mathbf{D}(k)\mathbf{u}(k)$
Laplace domain of continuous time-invariant	$s\mathbf{X}(s) - \mathbf{x}(0) = \mathbf{A}\mathbf{X}(s) + \mathbf{B}\mathbf{U}(s)$ $\mathbf{Y}(s) = \mathbf{C}\mathbf{X}(s) + \mathbf{D}\mathbf{U}(s)$
Z-domain of discrete time-invariant	$z\mathbf{X}(z) - z\mathbf{x}(0) = \mathbf{A}\mathbf{X}(z) + \mathbf{B}\mathbf{U}(z)$ $\mathbf{Y}(z) = \mathbf{C}\mathbf{X}(z) + \mathbf{D}\mathbf{U}(z)$

S4

Efficiently Modeling Long Sequences with Structured State Spaces
想通过 SSM 来处理长序列问题，采用的是 SSM 的一种连续时间不变形式



- 离散化：表达式中存在导数，没有办法直接处理离散的序列

$$\overline{A} = (I - \Delta/2 \cdot A)^{-1}(I + \Delta/2 \cdot A)$$
$$\overline{B} = (I - \Delta/2 \cdot A)^{-1}\Delta B$$
$$\overline{C} = C$$

S4

- 循环-卷积表示：等效成卷积形式，训练时采用卷积方式，推理时循环形式，加快训练和推理速度

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

$$
\begin{aligned}
y_k &= \overline{CA}^k \overline{B} u_0 + \overline{CA}^{k-1} \overline{B} u_1 + \dots + \overline{CAB} u_{k-1} + \overline{CB} u_k \\
y &= \overline{K} * u
\end{aligned}
$$

$$
\overline{K} \in \mathbb{R}^L = (\overline{CB}, \overline{CAB}, \dots, \overline{CA}^{L-1} \overline{B})
$$

卷积核大小固定为序列长度大小：傅里叶变换后进行加法

序列为因果关系：掩码

S4

- 基于 HiPPO 处理长序列：解决长距离依赖问题

<table>
<tr><th></th><th>x</th><th>Wxh</th><th>Hidden</th><th>Whh</th><th>Why</th><th>y</th></tr>
<tr><td rowspan="2">SNN</td><td>1</td><td>1xd’</td><td rowspan="2">d’</td><td rowspan="4">d’xd’</td><td>1x1</td><td>1</td></tr>
<tr><td>d</td><td>dxd’</td><td>1xd</td><td>d’</td></tr>
<tr><td rowspan="2">S4</td><td>1</td><td rowspan="2">1xd’</td><td>d’</td><td rowspan="2">1xd’</td><td>1</td></tr>
<tr><td>d</td><td>dxd’</td><td>d</td></tr>
</table>

相比 RNN 更显式的建模隐藏层状态对过去信息的保存，此时 A 表示了对过去信息的记忆方式，采用 HiPPO 方式初始化 A 能更好的记忆长距离信息

HiPPO Matrix

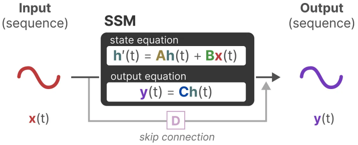
$$\text{HiPPO Matrix} \quad A_{nk} = \begin{cases} (2n+1)^{1/2}(2k+1)^{1/2} & \leftarrow \text{everything below the diagonal} \\ n+1 & \leftarrow \text{the diagonal} \\ 0 & \leftarrow \text{everything above the diagonal} \end{cases}$$

$$\begin{array}{cc} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 3 & 0 \\ 1 & 3 & 5 & 4 \end{bmatrix} & k \\ n & \end{array}$$

Mamba

相比 S4

- 将 Δ, B, C 变成输入依赖的，通过将输入通过全连接层预测得到,$\overline{A}$ 间接依赖输入（能对输入进行不同程度的记忆，之前 A, B, C 共享对不同输入的关注程度相同）
- 每个维度用一个 SSM,A,B,C 不再在不同维度共享



$$\overline{A} = (I - \Delta/2 \cdot A)^{-1}(I + \Delta/2 \cdot A)$$
$$\overline{B} = (I - \Delta/2 \cdot A)^{-1}\Delta B$$
$$\overline{C} = C$$

Algorithm 1 SSM (S4)

Input: $x : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
Output: $y : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
$A : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
▷ Represents structured $N \times N$ matrix
$B : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
$C : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
$\Delta : (\mathtt{D}) \leftarrow \tau_\Delta(\text{Parameter})$
$\overline{A}, \overline{B} : (\mathtt{D}, \mathtt{N}) \leftarrow \text{discretize}(\Delta, A, B)$
$y \leftarrow \text{SSM}(\overline{A}, \overline{B}, C)(x)$
▷ Time-invariant: recurrence or convolution
return y

Algorithm 2 SSM + Selection (S6)

Input: $x : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
Output: $y : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
$A : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
▷ Represents structured $N \times N$ matrix
$B : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow s_B(x)$
$C : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow s_C(x)$
$\Delta : (\mathtt{B}, \mathtt{L}, \mathtt{D}) \leftarrow \tau_\Delta(\text{Parameter} + s_\Delta(x))$
$\overline{A}, \overline{B} : (\mathtt{B}, \mathtt{L}, \mathtt{D}, \mathtt{N}) \leftarrow \text{discretize}(\Delta, A, B)$
$y \leftarrow \text{SSM}(\overline{A}, \overline{B}, C)(x)$
▷ Time-varying: recurrence (*scan*) only
return y

Mamba

带来的问题

- $\overline{A}, \overline{B}, \overline{C}$ 与输入相关，没有办法再提前计算卷积核并行训练减少反复读写显存这个瓶颈：并行扫描 + 核融合 + 重计算

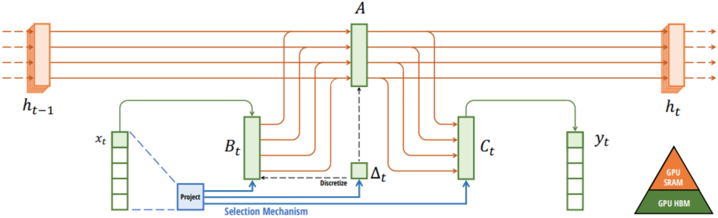
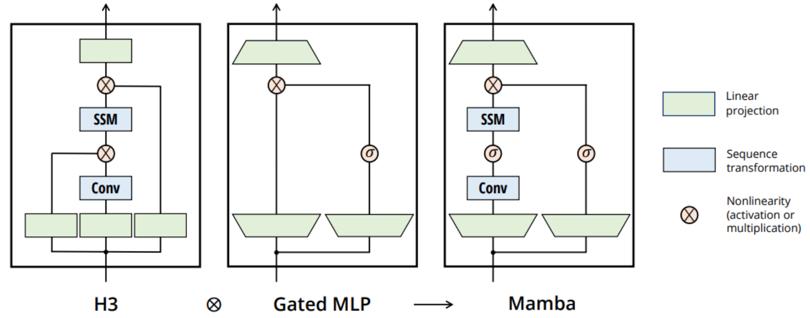


Figure 1: (**Overview**.) Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN, times batch size B and sequence length L) through clever alternate computation paths requiring time-invariance: the $(\Delta, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ parameters are constant across time. Our selection mechanism adds back input-dependent dynamics, which also requires a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU memory hierarchy.

Mamba

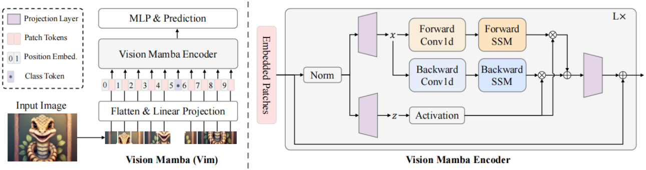
与现代神经网络结合，得到 Mamba 架构



Vision Mamba

Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model

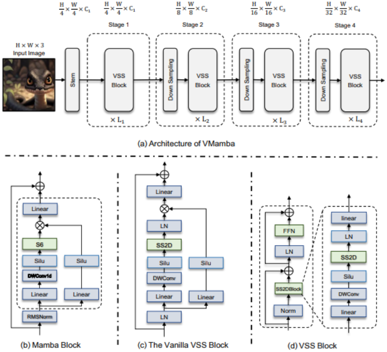
- Vision Transformer 结构
- 增加 backward 扫描分支，和 forward 扫描融合

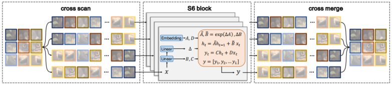


VMamba

VMamba: Visual State Space Model

- ConvNext 形式的 CNN 架构
- 图像 patch 展开后四个方向进行扫描 (左-> 右，右-> 左，上-> 下，下-> 上)
- 对比了几种不同的扫描方式

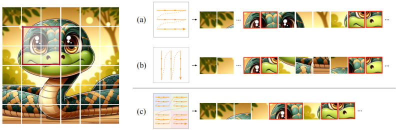


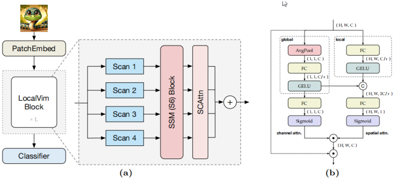


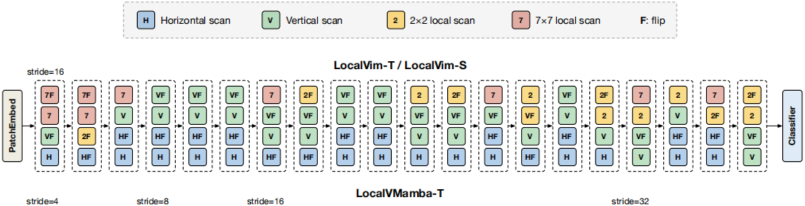
Local Mamba

LocalMamba: Visual State Space Model with Windowed Selective Scan

- 提出了四种扫描方式，对四种扫描方式进行融合
- 进行神经架构搜索，搜索四种扫描方式的最佳组合



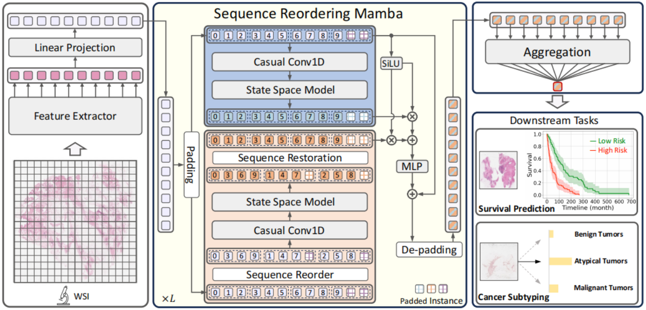




MambaMIL

MambaMIL: Enhancing Long Sequence Modeling with Sequence Reordering in Computational Pathology

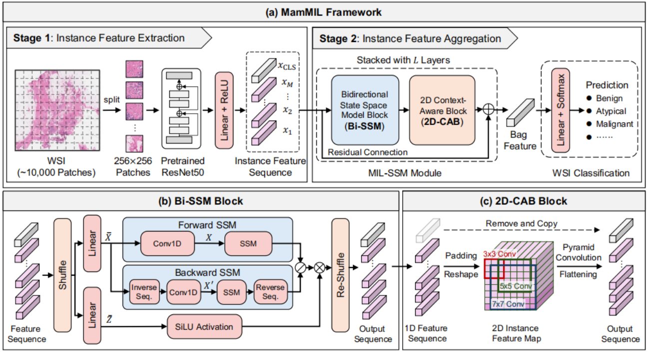
- 添加一个对序列重新排序后通过 Mamba 的分支
- 重新排序为 (近似) 上-> 下扫描



MamMIL

MamMIL: Multiple Instance Learning for Whole Slide Images with State Space Models

- 结构与 Vision Mamba 一样
- 增加一个 shuffle 和 reshuffle 操作



Fusion Mamba

FusionMamba: Efficient Image Fusion with State Space Model

- 提出 FSSM 模块，由另一个模态的数据去产生 B, C, Δ

Algorithm 1 SSM Block

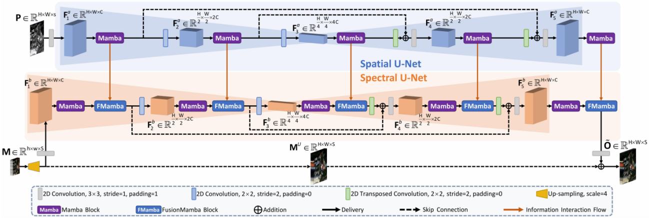
Input: x : (HW, C)
Output: y : (HW, C)

A : (C, N) ← $\textbf{Parameter}_A$
/* A represents C sets of structured N × N matrices [12] */
B : (HW, N) ← $\textbf{Linear}_B(x)$
C : (HW, N) ← $\textbf{Linear}_C(x)$
Δ : (HW, C) ← $\log(1 + \exp(\textbf{Linear}_\Delta(x) + \textbf{Parameter}_\Delta))$
/* $\textbf{Parameter}_\Delta$ is a bias vector with a size of C */
$\overline{\text{A}}$: (HW, C, N) ← exp(Δ ⊗ A)
$\overline{\text{B}}$: (HW, C, N) ← Δ ⊗ B
y ← SSM($\overline{\text{A}}$, $\overline{\text{B}}$, C)(x)
/* SSM represents Eq. 3 implemented by selective scan [11] */
return y

Algorithm 2 FSSM Block

Input: x^a, x^b : (HW, C)
Output: y^a : (HW, C)

A : (C, N) ← $\textbf{Parameter}_A$
/* A represents C sets of structured N × N matrices [12] */
B : (HW, N) ← $\textbf{Linear}_B(x^b)$
C : (HW, N) ← $\textbf{Linear}_C(x^b)$
Δ : (HW, C) ← $\log(1 + \exp(\textbf{Linear}_\Delta(x^b) + \textbf{Parameter}_\Delta))$
/* $\textbf{Parameter}_\Delta$ is a bias vector with a size of C */
$\overline{\text{A}}$: (HW, C, N) ← exp(Δ ⊗ A)
$\overline{\text{B}}$: (HW, C, N) ← Δ ⊗ B
y^a ← SSM($\overline{\text{A}}$, $\overline{\text{B}}$, C)(x^a)
/* SSM represents Eq. 3 implemented by selective scan [11] */
return y^a

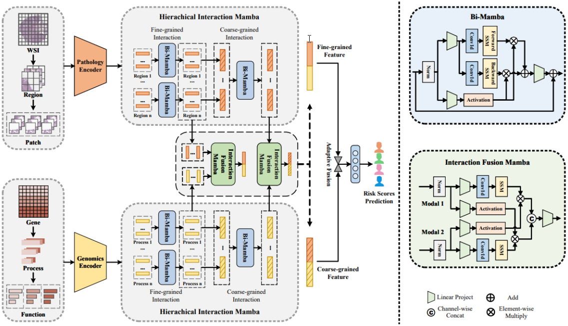


SurvMamba

SurvMamba: State Space Model with Multi-grained Multi-modal Interaction for Survival Prediction

- 不同模态交叉 gated
- 拼接融合特征

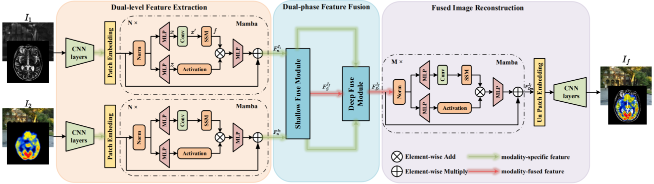
存在问题：这种融合方式需要图像序列的长度和基因序列的长度一致才能融合，而且序列相应位置由对应关系融合才有意义

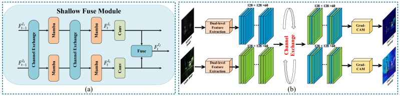


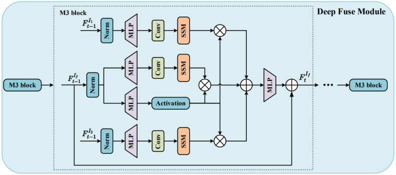
MambaDFuse

MambaDFuse: A Mamba-based Dual-phase Model for Multi-modality Image Fusion

- 浅层融合： 交换通道
- 深层融合： 浅层融合的作 gate



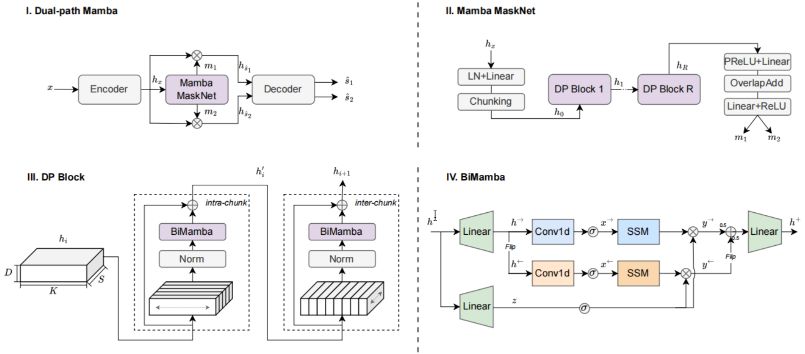




Frame Title

Dual-path Mamba: Short and Long-term Bidirectional Selective Structured State Space Models for Speech Separation

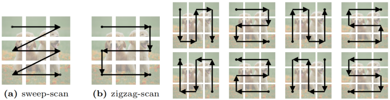
- 分成多个 chunk，intra-chunk 和 inter-chunk 交替
- 前向和反向 Mamba 结合



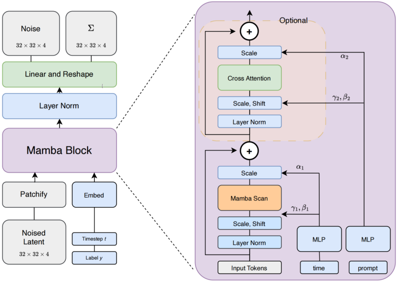
ZigMa

ZigMa: A DiT-style Zigzag Mamba Diffusion Model

- 基于 DiT 修改，时间信息影响 SSM 模块，prompt 信息对应 CrossAttention
- 不同的扫描方式



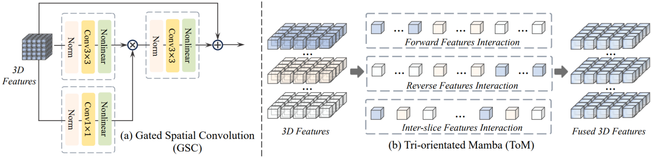
(a) sweep-scan (b) zigzag-scan (c) zigzag-scan with 8 schemes

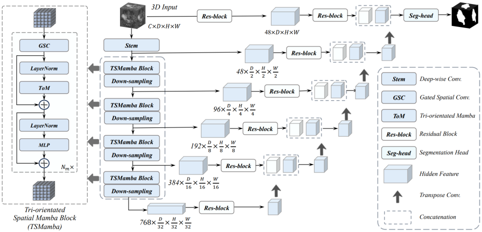


SegMamba

SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation

- 将 3D 图像按不同方式展平成 1D 序列





The Hidden Attention of Mamba Models

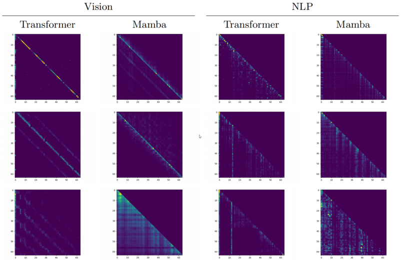
The Hidden Attention of Mamba Models

$$y = \tilde{\alpha}x, \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix} = \begin{bmatrix} C_1\bar{B}_1 & 0 & \cdots & 0 \\ C_2\bar{A}_2\bar{B}_1 & C_2\bar{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ C_L\Pi_{k=2}^{L}\bar{A}_k\bar{B}_1 & C_L\Pi_{k=3}^{L}\bar{A}_k\bar{B}_2 & \cdots & C_L\bar{B}_L \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix},$$

$$\tilde{Q}_i := S_C(\hat{x}_i),\ \tilde{K}_j := \mathrm{ReLU}(S_\Delta(\hat{x}_j)S_B(\hat{x}_j)),\ \tilde{H}_{i,j} := \exp\Big(\sum_{\substack{k=j+1 \\ S_\Delta(\hat{x}_k)>0}}^{i} S_\Delta(\hat{x}_k)\Big)A \tag{18}$$

Eq. 17 can be further simplified to:

$$\tilde{\alpha}_{i,j} \approx \tilde{Q}_i\tilde{H}_{i,j}\tilde{K}_j \tag{19}$$



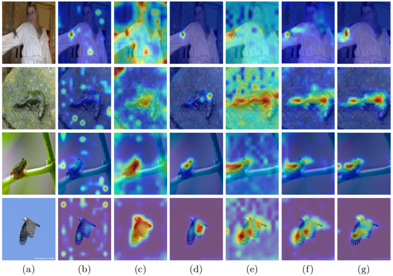


Fig. 5: Qualitative results for the different explanation methods for the ViT-small and the Mamba-small models. (a) the original image, (b) the aggregated Raw-Attention of ViT-Small, (c) Attention Rollout for ViT-Small, (d) Transformer-Attribution for ViT-Small, (e) the Raw-Attention of Mamba-Small, (f) Attention-Rollout of Mamba-Small and (g) the Mamba-Attribution method for the Mamba-Small model.

Thanks!