CARE Transformer: Mobile-Friendly Linear Visual Transformer via Decoupled Dual Interaction

Yuan Zhou[1], Qingshan Xu[1], Jiequan Cui[1], Junbao Zhou[1], Jing Zhang[2], Richang Hong[3], Hanwang Zhang[1]
[1]Nanyang Technological University, [2]Beihang University, [3]Hefei University of Technology

A. BACKGROUND

(I) Dot-Product Attention

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{SoftMax}(\mathbf{Q}\mathbf{K}^{\top})\mathbf{V}$$

where $\mathbf{Q} = \mathbf{W}_1\mathbf{X}, \mathbf{K} = \mathbf{W}_2\mathbf{X}, \mathbf{V} = \mathbf{W}_3\mathbf{X}$ and $\mathbf{X}, \mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{\mathrm{n}\times\mathrm{d}}$

*****_Dot-product attention has **quadratic complexity** w.r.t. the length of input tokens, i.e.,_ $O(n^2 d)$.

A. BACKGROUND

(II) Linear Attention

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{SoftMax}(\mathbf{Q}\mathbf{K}^\top)\mathbf{V}$$

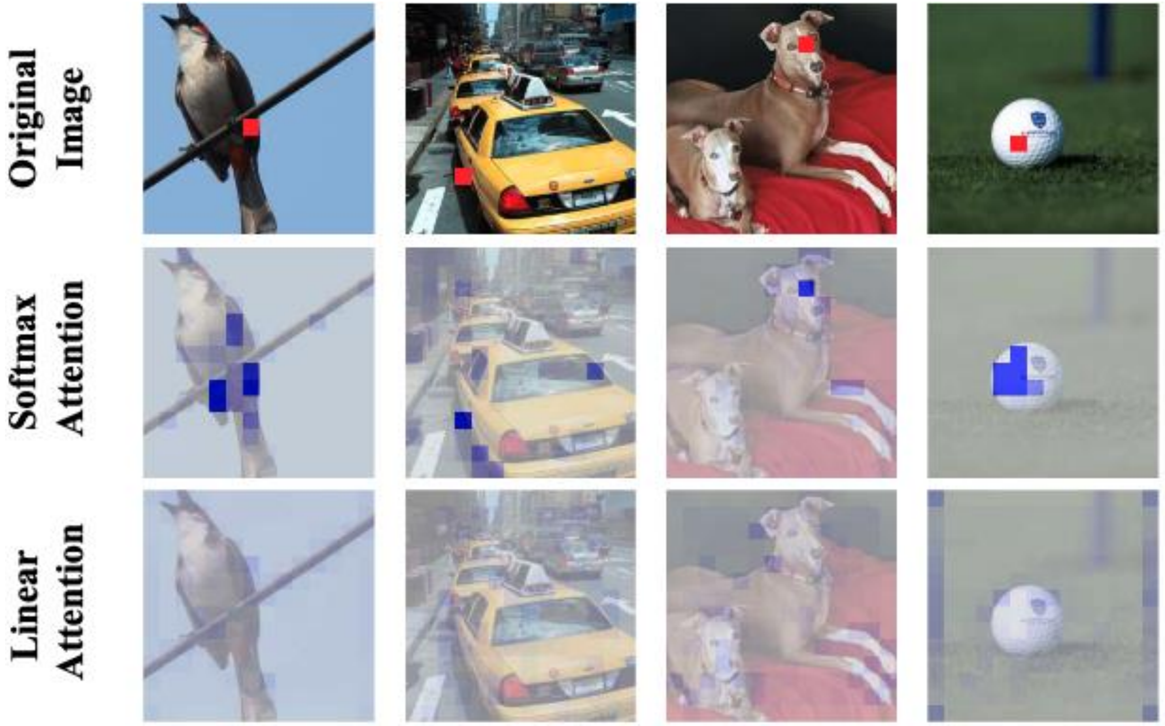
$$\xrightarrow{\text{Remove } SoftMax(\cdot)} \mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{Q}\mathbf{K}^\top\mathbf{V}$$

$$\xrightarrow{\text{Change direction}} \mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{Q}(\mathbf{K}^\top\mathbf{V})$$

These two steps change the quadratic complexity to the channel dimension, i.e., $O(nd^2)$*!*

(II) Linear Attention

One main drawback: Low entropy property[1][2] .



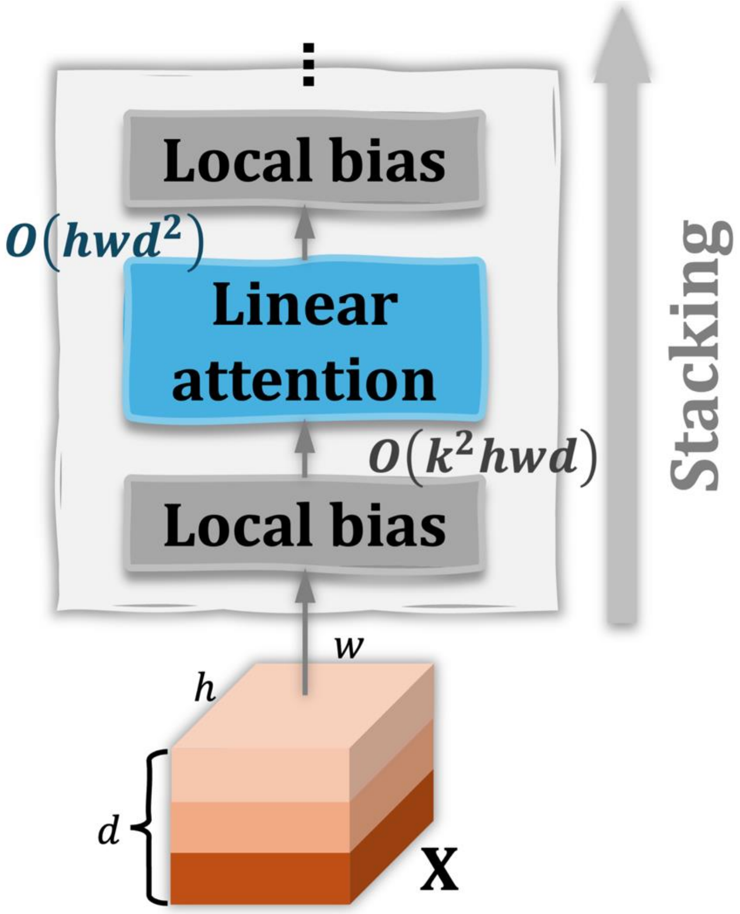
*The figure is borrowed from [1].

The non-linear function SoftMax serves as a global relation comparator, suppressing low-value similarities and highlighting high-value relationships!

[1] FLatten Transformer: Vision Transformer using Focused Linear Attention, ICCV' 23
[2] The Hedgehog & the Porcupine: Expressive Linear Attentions with Softmax Mimicry, ICLR' 24

A. BACKGROUND

(III) Stacked local enhancement[3]



$$\Omega = \underbrace{2k^2hwd}_{\text{Local inductive bias}} + \underbrace{\overbrace{4hwd^2}^{\text{Projection}} + \overbrace{2hwd^2}^{\textbf{QKV}\text{ Multiplication}}}_{\text{Long-range dependencies}}$$

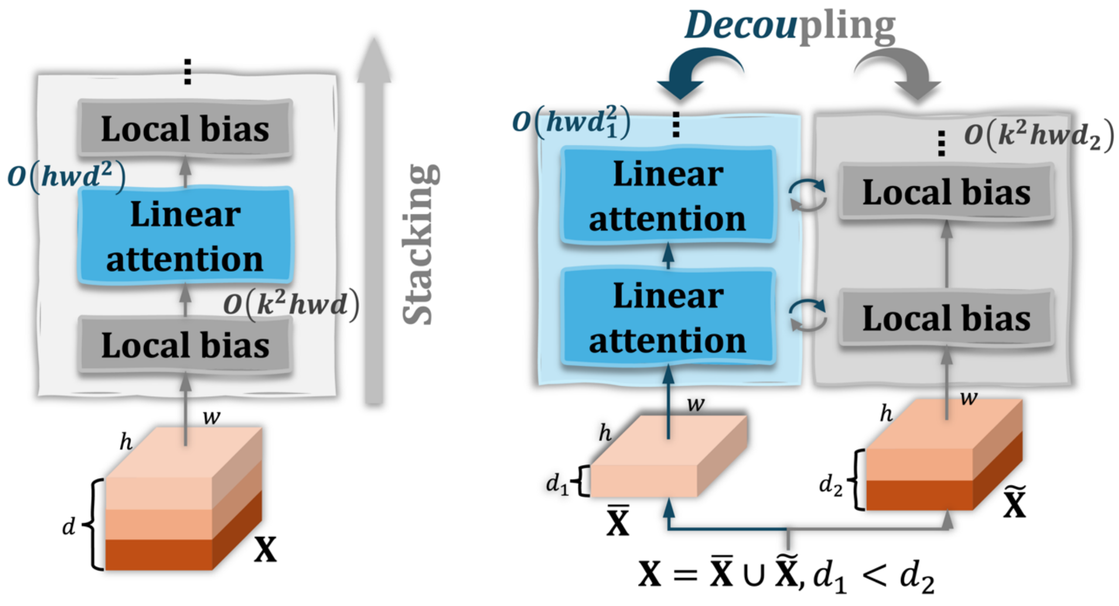
Two Drawbacks

- Input features need to undergo all the local and the global processors, leading to low efficiency.
- The flexibility of models is damaged, since it makes them inflexible in facilitating information exchange between local and global features, yielding unsatisfactory accuracy.

[3] Demystify Mamba in Vision: A Linear Attention Perspective, Han et al., NeurIPS' 2024

B. Methodology

(I) Stacking *V.S.* Decoupling



(a) Stacked local bias and global information

(b) Asymmetrical decoupling strategy

Two Solutions

- Feature decoupling can fully unleash the power of linear attention!
- Beyond S1, It is necessary to fully leverage interaction and complementarity between features.

$$\Omega = 2\lambda_2 k^2 hwd + 4\lambda_1 hwd^2 + 2\lambda_1 hwd^2 \text{ where } \lambda_1 = \left(\frac{d_1}{d}\right)^2 \text{ and } \lambda_2 = \frac{d_2}{d}$$

B. Methodology

(I)Asymmetrical Feature Decoupling: Divide and Conquer!

Proposition1. Linear attention has *quadratic complexity* to channel dimension, i.e., $O(hwd_1^2)$. Features should be decoupled in an asymmetrical way, meaning $d_1 + d_2 = d$ and $d_1 < d_2$, which further boosts the efficiency of models. ***The philosophy behind this design lies that long-range dependencies are useful but neighboring local information is more important!***

Proof of Proposition1. Setting $d_1 < d_2$ can further reduce the computation complexity of models. Letting $d_2 - d_1 = \Delta$, we have $d_1 = \frac{d-\Delta}{2}$ and $d_2 = \frac{d+\Delta}{2}$, due to $d_1 + d_2 = d$ and $d_2 - d_1 = \Delta$. Therefore, Equation (b) can be rewritten as follows:

$$\Omega(\Delta) = 2\lambda_2 k^2 hwd + 4\lambda_1 hwd^2 + 2\lambda_1 hwd^2 = k^2 hw(d+\Delta) + \frac{2}{3} hw(d-\Delta)^2$$

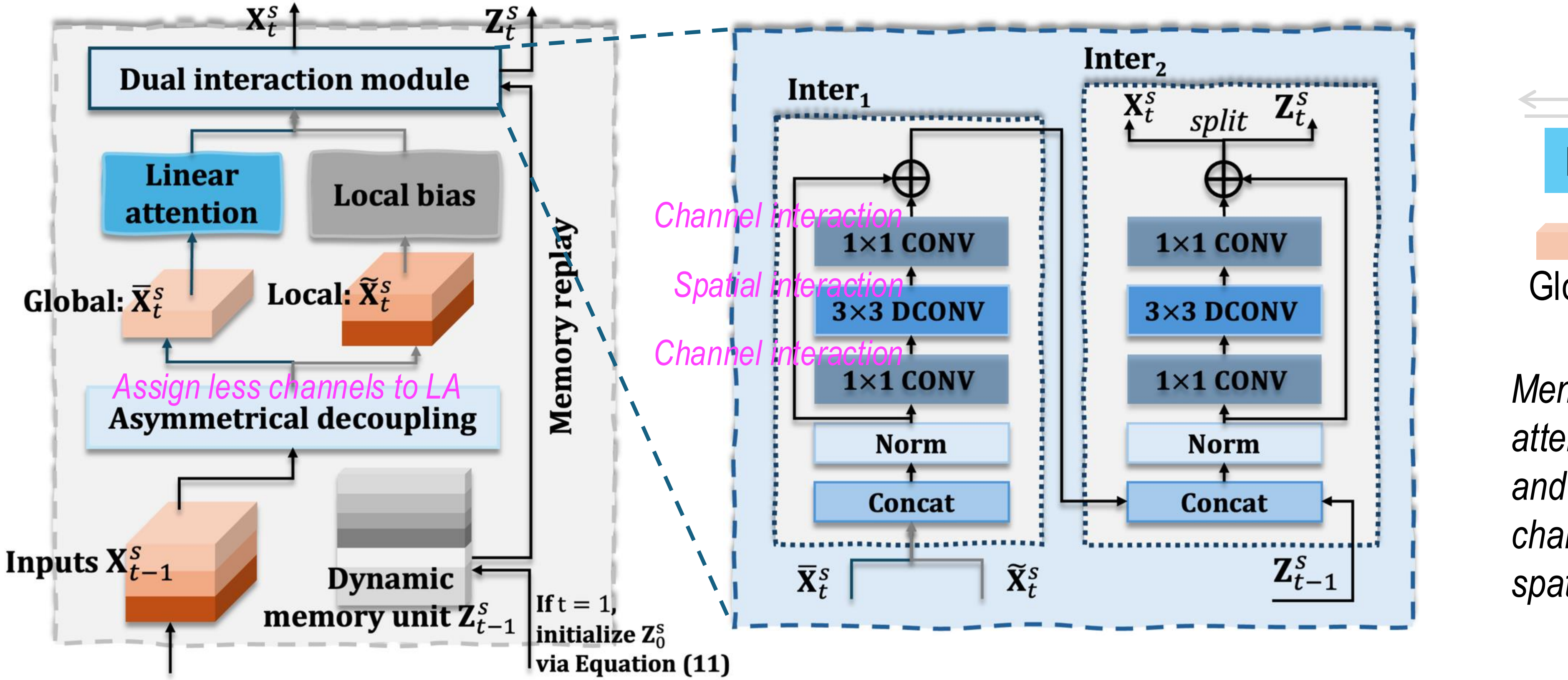
Also,

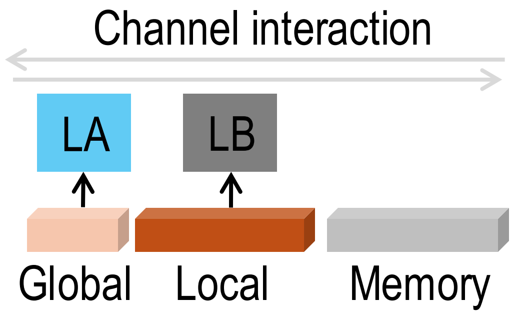
$$\Omega(\Delta_1) - \Omega(\Delta_2) = \frac{2}{3} hw(\Delta_1 - \Delta_2)\left(\frac{2}{3}k^2 + \Delta_1 + \Delta_2 - 2d\right)$$

Letting $\Delta_1 > 0$ and $\Delta_2 = 0$, we have $\Omega(\Delta_1) - \Omega(\Delta_2) < 0$ as $\Delta_1 < d$ and the kernel size generally obeys $k_2 \ll d$, thereby proving that the asymmetrical setting ($\Delta_1 > 0$) has less complexity compared to the symmetrical scenario ($\Delta_2 = 0$).

B. Methodology

(II) CARE: Decoupled Dual-Interactive Linear Attention



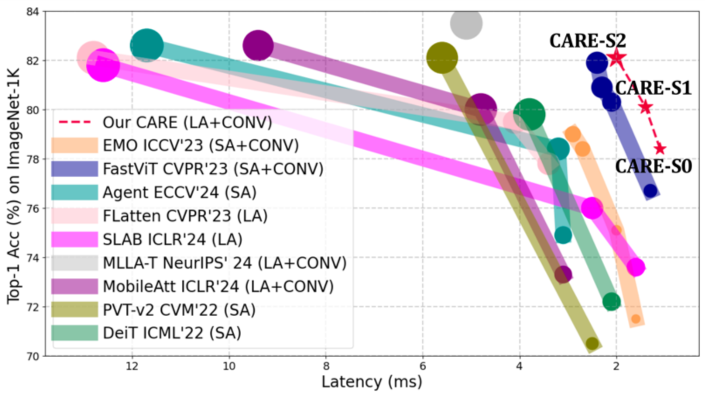


Memory is not fed to the linear attention or the local bias learner, and just updated by 1x1 CONVs in channel domain and 3x3 CONVs in spatial domain.

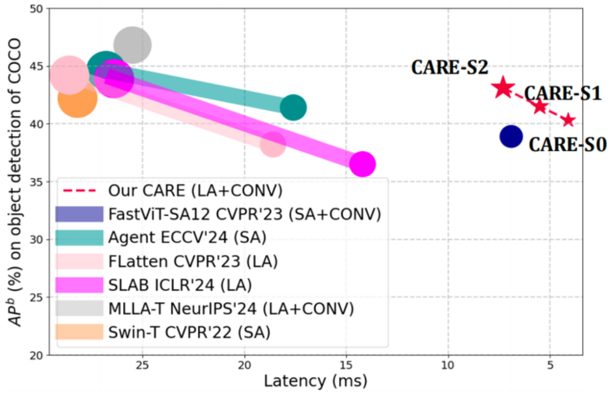
(a) CARE Block (b) Dual Interaction Module (c) Memory update

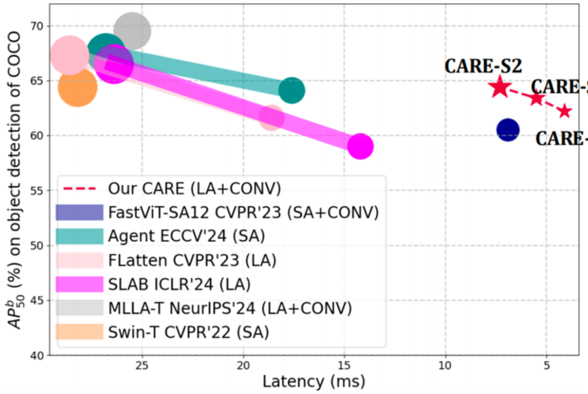
C. Experimental Results

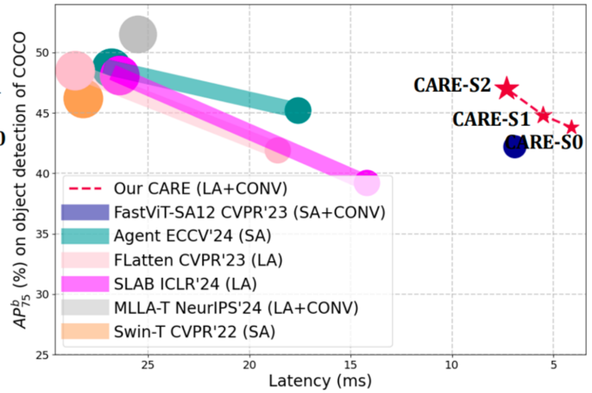
ImageNet-1K classification



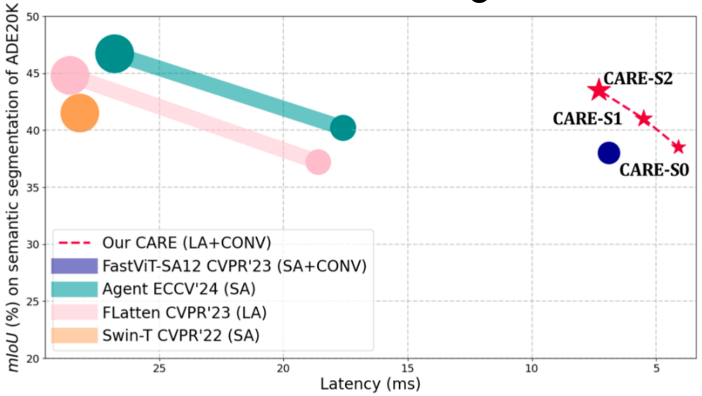
COCO object detection



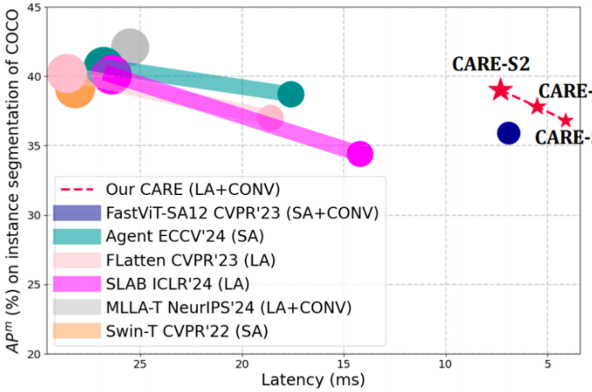


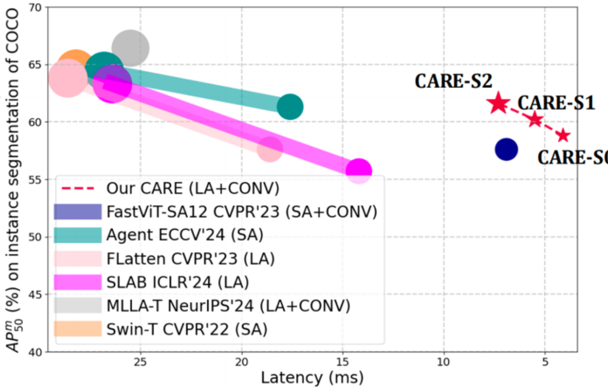


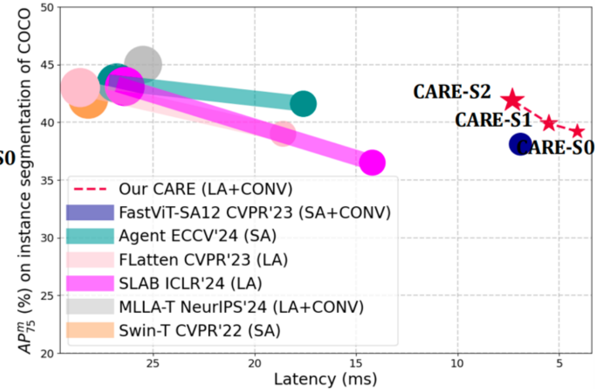
ADE20K semantic segmentation



COCO instance segmentation







Thanks!!!