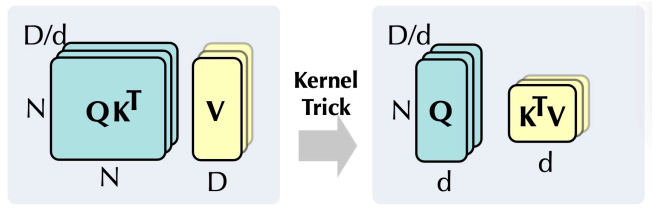
Mobile Attention: Mobile-Friendly Linear-Attention for Vision Transformers

Tsinghua University

Zhiyu Yao, Jian Wang, Haixu Wu, Jingdong Wang, Mingsheng Long#

ICML | 2024 Poster

Standard attention is not Mobile-Friendly

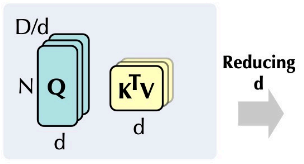


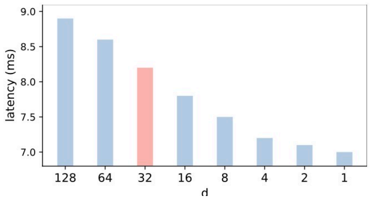
Standard Attention $O(N^2D)$ **Linear Attention** $O(NDd)$

Previous work: Standard attention in Transformers has a **quadratic complexity** with respect to the number of tokens

Linear-attention is emerging as a promising alternative with linear complexity

Key Insight: Reducing head dimensions will result in **lower latency** and **improved efficiency,** but leading too many heads

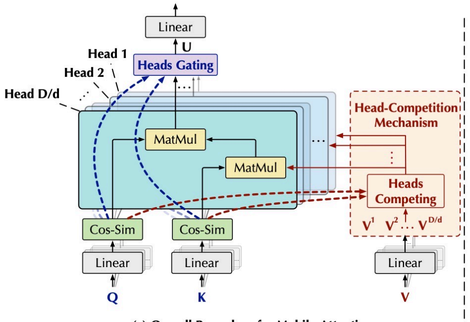




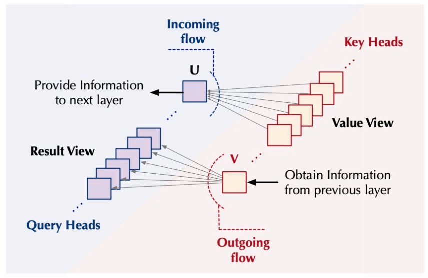
Linear Attention $O(NDd)$

Challenges: a small per-head dimension may cause some heads to struggle in learning valuable subspaces

Mobile-Attention with a Head-Competitive Mechanism



(a) Overall Procedure for Mobile-Attention



(b) Information Flow among Different Heads

➤ Incoming and Outgoing Flow

$$\mathbf{I}^h = \Phi\left(\mathbf{Q}^h\right) \sum_{j=1}^{M} \Phi\left(\mathbf{K}^j\right)^{\mathsf{T}},$$

① "I" Represent the capacity of incoming flow

$$\mathbf{O}^h = \Phi\left(\mathbf{K}^h\right) \sum_{i=1}^{M} \Phi\left(\mathbf{Q}^i\right)^{\mathsf{T}},$$

② "O" Represent the capacity of outgoing flow

➤ Head-Competitive Mechanism

$$\overline{\mathbf{I}}^h = \Phi(\mathbf{Q}^h) \sum_{h'=1}^{M} \frac{\Phi\left(\mathbf{K}^{h'}\right)^{\mathsf{T}}}{\mathbf{O}^{h'}},$$

$$\overline{\mathbf{O}} = \Phi(\mathbf{K}^h) \sum_{h'=1}^{M} \frac{\Phi\left(\mathbf{Q}^{h'}\right)^{\mathsf{T}}}{\mathbf{I}^{h'}},$$

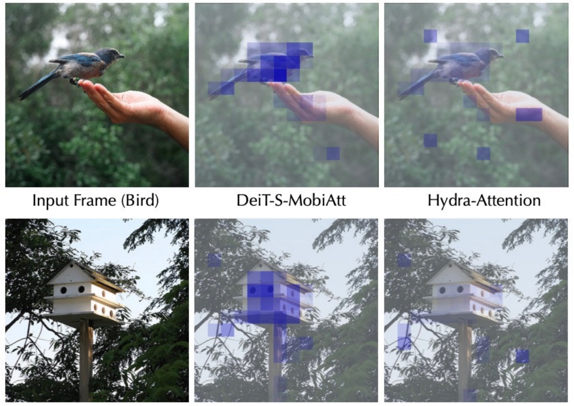
$$\overline{\mathbf{V}} = \mathrm{Softmax}\left(\overline{\mathbf{O}}\right) \odot \mathbf{V},$$

$$\mathbf{U}_t^h = \sigma\left(\overline{\mathbf{I}}_t^h\right) \frac{\Phi\left(\mathbf{Q}_t^h\right) \sum_{i=1}^{N} \Phi\left(\mathbf{K}_i^h\right)^{\mathsf{T}} \left(\overline{\mathbf{V}}_i^h\right)}{\Phi\left(\mathbf{Q}_t^h\right) \sum_{j=1}^{N} \Phi\left(\mathbf{K}_j^h\right)^{\mathsf{T}}},$$

① Contrasting the capacity of incoming flow for final result tokens as 1

② Making the outgoing flow of value tokens compete with each other under this fixed sum situation

Attention Visualization



Input Frame (Bird) | DeiT-S-MobiAtt | Hydra-Attention

Input Frame (Birdhouse) | DeiT-S-MobiAtt | Hydra-Attention

ImageNet-1K Classification

Model	Params(M)	GMACs	CoreML(ms)	A100 (ms)	Pixel 6 (ms)	Top-1 Acc(%)
MobileNetV2 (Sandler et al., 2018)	3.5	0.30	0.9	5.0	25.3	71.8
MobileViT-XS (Mehta & Rastegari, 2021)	2.3	0.70	7.3	11.7	64.4	74.8
EdgeViT-XXS (Chen et al., 2022)	4.1	0.60	2.4	11.3	30.9	74.4
EfficientNet-B0 (Tan & Le, 2019)	5.3	0.40	1.4	10.0	29.4	77.1
ConvNeXt-T (Liu et al., 2022a)	29.0	4.50	83.7	28.8	340.5	82.1
Swin-T (Liu et al., 2021)	29.0	4.50	97.3	22.0	-	81.3
DeiT-T (Touvron et al., 2021)	5.7	1.25	4.5	7.1	66.6	72.2
DeiT-T-MobiAtt	5.7	**1.22**	**3.8**	5.9	53.9	**73.3**
DeiT-S (Touvron et al., 2021)	22.0	4.60	9.0	15.5	218.2	79.8
DeiT-S-MobiAtt	22.0	**4.20**	**7.2**	13.3	175.7	**80.0**
DeiT-B (Touvron et al., 2021)	86.3	17.56	18.2	-	-	83.4
DeiT-B-MobiAtt	86.3	**17.03**	**13.3**	-	-	**84.2**
PVT-v2-b0 (Wang et al., 2022)	3.7	0.60	78.4	17.6	-	70.5
PVT-v2-b0-MobiAtt	**3.5**	**0.56**	**57.3**	15.0	-	**71.5**
PVT-v2-b2 (Wang et al., 2022)	25.4	4.00	101.0	36.2	-	82.1
PVT-v2-b2-MobiAtt	**21.1**	**3.80**	**65.6**	33.7	-	**82.6**
PVT-v2-b3 (Wang et al., 2022)	45.2	-	114.5	230.9	-	83.3
PVT-v2-b3-MobiAtt	**39.0**	-	**89.1**	210.1	-	**84.0**
EfficientFormerV2-S0 (Li et al., 2022a)	3.5	0.40	0.9	6.6	20.8	75.7
EfficientformerV2-S0-MobiAtt	3.5	0.37	**0.7**	5.5	16.2	**76.0**
EfficientFormerV2-S2 (Li et al., 2022a)	12.6	1.25	1.6	14.5	57.2	81.6
EfficientformerV2-S2-MobiAtt	12.6	**1.22**	**1.2**	13.1	48.9	**82.1**
EfficientFormerV2-L (Li et al., 2022a)	26.1	2.56	2.7	22.5	117.7	83.3
EfficientformerV2-L-MobiAtt	26.1	**2.50**	**2.2**	20.3	97.4	**83.7**

Compared with Other Linear Attention

Model	Complexity	GMACs	CoreML(ms)	Top-1 Acc (%)
Hydra-DeiT-S (Bolya et al., 2022)	$\mathcal{O}(ND)$	4.10	7.0	73.5
Castling-DeiT-S (You et al., 2023)	$\mathcal{O}(ND^2)$	4.52	9.4	79.8
DeiT-S (Touvron et al., 2021)	$\mathcal{O}(N^2D)$	4.60	9.0	79.8
DeiT-S-MobiAtt w/ vanilla design	$\mathcal{O}(ND^2)$	-	8.1	79.0
DeiT-S-MobiAtt* w/ SE (Hu et al., 2018)	$\mathcal{O}(ND^2)$	-	7.3	78.3
DeiT-S-MobiAtt* w/ GLU (Shazeer, 2020)	$\mathcal{O}(ND^2)$	-	7.3	77.5
DeiT-S-MobiAtt w/o Head-competing	$\mathcal{O}(ND)$	4.18	7.2	76.4
DeiT-S-MobiAtt	$\mathcal{O}(ND)$	4.20	**7.2**	**80.0**