

Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles

Peizhen Li

Faculty of Science and Engineering
Macquarie University

Feb 16, 2024

Outline

- 1 Multiscale Vision Transformers**
 - Multiscale Feature Hierarchies
 - Multi Head Pooling Attention
 - Multiscale Transformer Networks
- 2 Improved Multiscale Vision Transformers**
 - Decomposed Relative Positional Embedding
 - Residual Pooling Connection
- 3 A Hierachical Vision Transformer without Bells-and-Whistles**
 - Implementation Details
- 4 Reflection & Future Work**

Multiscale Feature Hierarchies

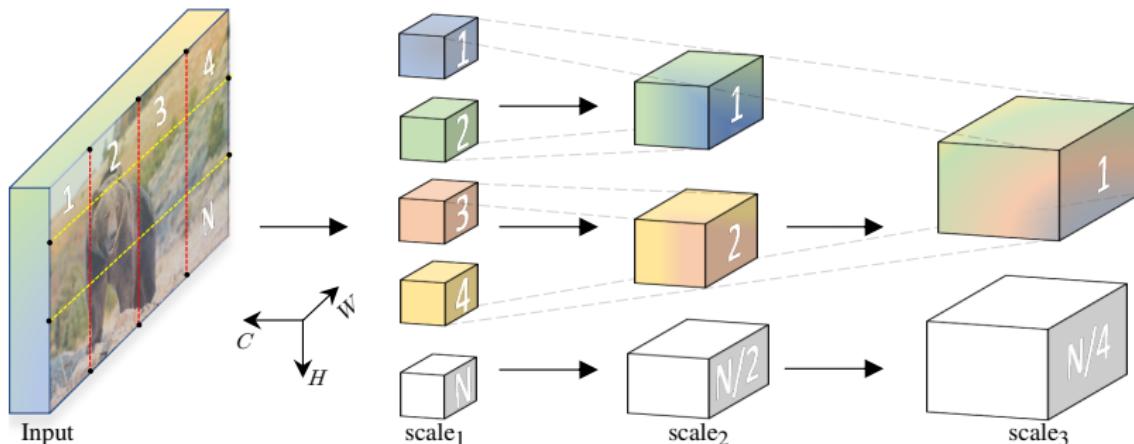
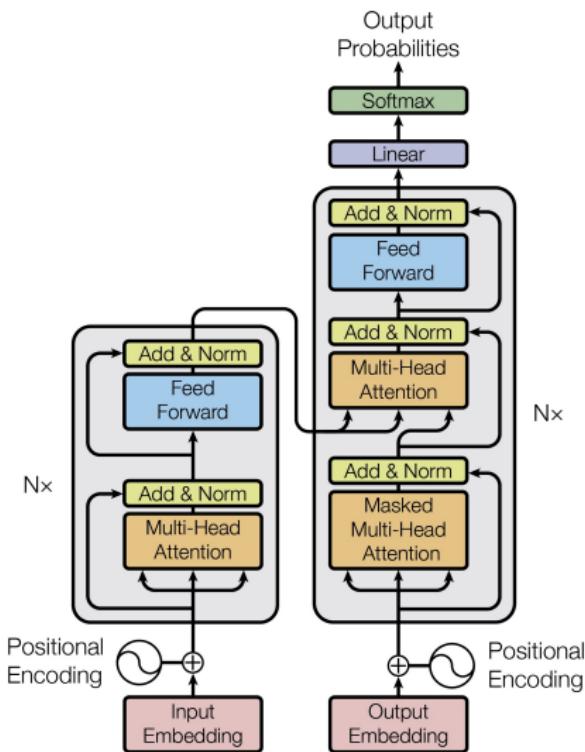


Figure: Several resolution-channel *scale* stages of MViT¹.

¹ICCV-2021

Transformer Revisit



Things to Think About

- A stack of **identical** blocks (Single Scale)
- Computational complexity of canonical self-attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{\tilde{d}}}\right)V \quad (1)$$

where

$$X \in \mathbb{R}^{n \times d}, Q = XW_Q, K = XW_K, V = XW_V$$

$$W_Q, W_K, W_V \in \mathbb{R}^{d \times \tilde{d}}$$

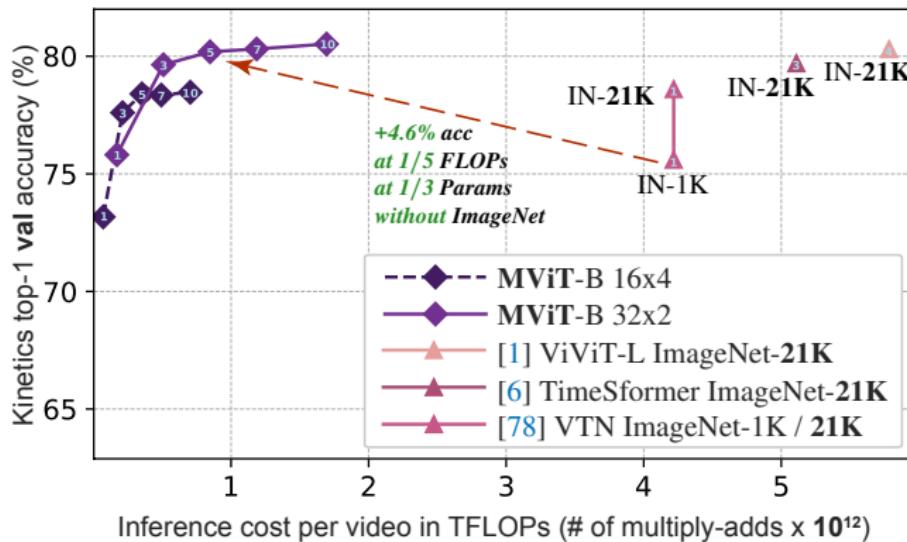
Self-Attention Computational Complexity

Scales quadratically in input sequence length n

- 1 Calculation of $S = \frac{QK^T}{\sqrt{\tilde{d}}}$ takes $\mathcal{O}(n^2\tilde{d})$
- 2 Exponentiation and calculation of row sum of S takes $\mathcal{O}(n^2)$
- 3 Division of each element of S with the corresponding row sum takes $\mathcal{O}(n^2)$
- 4 Post-multiplication with V takes $\mathcal{O}(n^2\tilde{d})$

Motivation

- Decrease computing requirements
- A better sense of “context” at the lower resolutions guiding the processing at higher resolutions

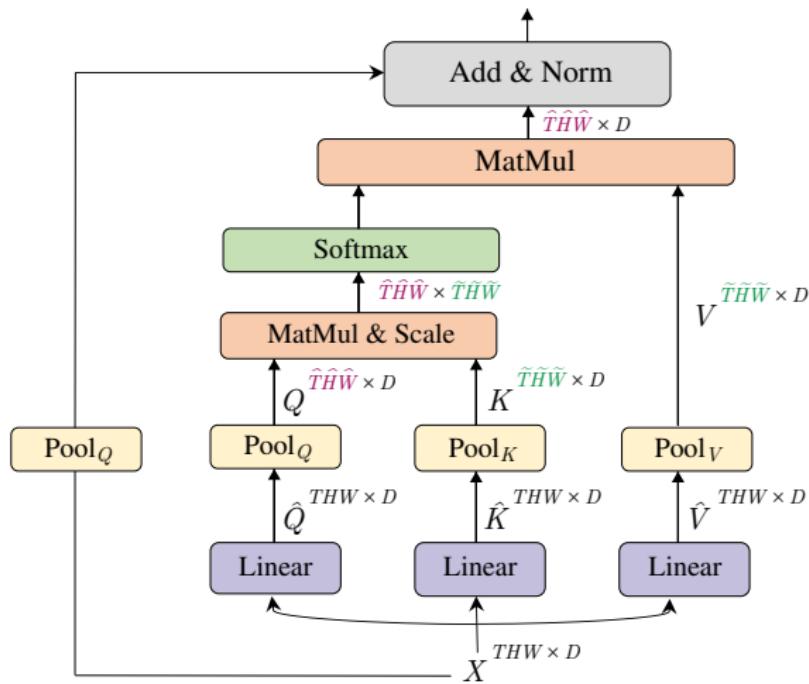


Multiscale: Step by Step

- Scale stages: Transformer blocks that operates on the same scale (identical resolution and channel capacity).

stages	operators	output sizes
data layer	stride $T \times 1 \times 1$	$D \times T \times H \times W$
cube ₁	$c_T \times c_H \times c_W$, D stride $s_T \times 4 \times 4$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale ₂	$\begin{bmatrix} \text{MHPA}(D) \\ \text{MLP}(4D) \end{bmatrix} \times N_2$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale ₃	$\begin{bmatrix} \text{MHPA}(2D) \\ \text{MLP}(8D) \end{bmatrix} \times N_3$	$2D \times \frac{T}{s_T} \times \frac{H}{8} \times \frac{W}{8}$
scale ₄	$\begin{bmatrix} \text{MHPA}(4D) \\ \text{MLP}(16D) \end{bmatrix} \times N_4$	$4D \times \frac{T}{s_T} \times \frac{H}{16} \times \frac{W}{16}$
scale ₅	$\begin{bmatrix} \text{MHPA}(8D) \\ \text{MLP}(32D) \end{bmatrix} \times N_5$	$8D \times \frac{T}{s_T} \times \frac{H}{32} \times \frac{W}{32}$

Reduce Resolution by Pooling



Pooling Operator

$$\mathcal{P}(\cdot; \Theta) \tag{2}$$

where

$$\Theta := (\mathbf{k}, \mathbf{s}, \mathbf{p})$$

poling kernel $\mathbf{k} \in \mathbb{R}^{k_T \times k_H \times k_W}$

stride $\mathbf{s} \in \mathbb{R}^{s_T \times s_H \times s_W}$

padding $\mathbf{p} \in \mathbb{R}^{p_T \times p_H \times p_W}$

then

$$\tilde{\mathbf{L}} = \left\lfloor \frac{\mathbf{L} + 2\mathbf{p} - \mathbf{k}}{\mathbf{s}} \right\rfloor + 1$$

where

$$\mathbf{L} = T \times H \times W, \quad \tilde{\mathbf{L}} = \tilde{T} \times \tilde{H} \times \tilde{W}$$

length reduced by a factor of $s_T s_H s_W$

Computational Complexity

Given sequence length $L = THW$

After pooling $L/f_K, L/f_Q, L/f_V$

$$f_j = s_T^j s_H^j s_W^j, \quad \forall j \in \{Q, K, V\}$$

- 1 Compute key, query, value embeddings

$$\mathcal{O}(THWD^2/h)$$

- 2 Calculate attention matrix and post-multiply with value vectors

$$\mathcal{O}(T^2 H^2 W^2 D / (f_Q f_K h))$$

Overall $\mathcal{O}(THWD/h(D + THW/(f_Q f_K)))$

Multiscale Transformer Networks

- Channel expansion:
e.g., $2D \times \frac{T}{s_T} \times \frac{H}{8} \times \frac{W}{8}$ to $4D \times \frac{T}{s_T} \times \frac{H}{16} \times \frac{W}{16}$
- Query pooling: $\mathcal{P}(Q; \mathbf{k}; \mathbf{p}; \mathbf{s})$
- Key-Value pooling: $\Theta_K \equiv \Theta_V$
- Skip connections
- Multiscale attention block

Experimental Results on Kinetics-400

model	pre-train	top-1	top-5	FLOPs×views	Param
Two-Stream I3D [11]	-	71.6	90.0	$216 \times \text{NA}$	25.0
ip-CSN-152 [96]	-	77.8	92.8	$109 \times 3 \times 10$	32.8
SlowFast $8 \times 8 + \text{NL}$ [30]	-	78.7	93.5	$116 \times 3 \times 10$	59.9
SlowFast $16 \times 8 + \text{NL}$ [30]	-	79.8	93.9	$234 \times 3 \times 10$	59.9
X3D-M [29]	-	76.0	92.3	$6.2 \times 3 \times 10$	3.8
X3D-XL [29]	-	79.1	93.9	$48.4 \times 3 \times 10$	11.0
ViT-B-VTN [78]	ImageNet-1K	75.6	92.4	$4218 \times 1 \times 1$	114.0
ViT-B-VTN [78]	ImageNet- 21K	78.6	93.7	$4218 \times 1 \times 1$	114.0
ViT-B-TimeSformer [6]	ImageNet- 21K	80.7	94.7	$2380 \times 3 \times 1$	121.4
ViT-L-ViT [1]	ImageNet- 21K	81.3	94.7	$3992 \times 3 \times 4$	310.8
ViT-B (our baseline)	ImageNet- 21K	79.3	93.9	$180 \times 1 \times 5$	87.2
ViT-B (our baseline)	-	68.5	86.9	$180 \times 1 \times 5$	87.2
MViT-S	-	76.0	92.1	$32.9 \times 1 \times 5$	26.1
MViT-B , 16×4	-	78.4	93.5	$70.5 \times 1 \times 5$	36.6
MViT-B , 32×3	-	80.2	94.4	$170 \times 1 \times 5$	36.6
MViT-B , 64×3	-	81.2	95.1	$455 \times 3 \times 3$	36.6

Good Enough?

MViTv2: Improved Multiscale Vision Transformers for Classification and Detection²

- 1 Decomposed relative positional embeddings
- 2 Residual pooling connection

Decomposed Relative Positional Embedding

$$\text{Attn}(Q, K, V) = \text{Softmax} \left((QK^T + E^{(\text{rel})}) / \sqrt{d} \right) V \quad (3)$$

where

$$E_{ij}^{(\text{rel})} = Q_i \cdot R_{p^{(i)}, p^{(j)}}, \quad R_{p^{(i)}, p^{(j)}} \in \mathbb{R}^d$$

decompose along axes

$$R_{p^{(i)}, p^{(j)}} = R_{h(i), h(j)}^{\text{h}} + R_{w(i), w(j)}^{\text{w}} + R_{t(i), t(j)}^{\text{t}} \quad (4)$$

Wait...

$$e_{ij} = \frac{Q_i K_j + E_{ij}^{(\text{rel})}}{\sqrt{d}}, \quad E_{ij}^{(\text{rel})} = Q_i \cdot R_{p(i), p(j)} \quad (5)$$

vs.

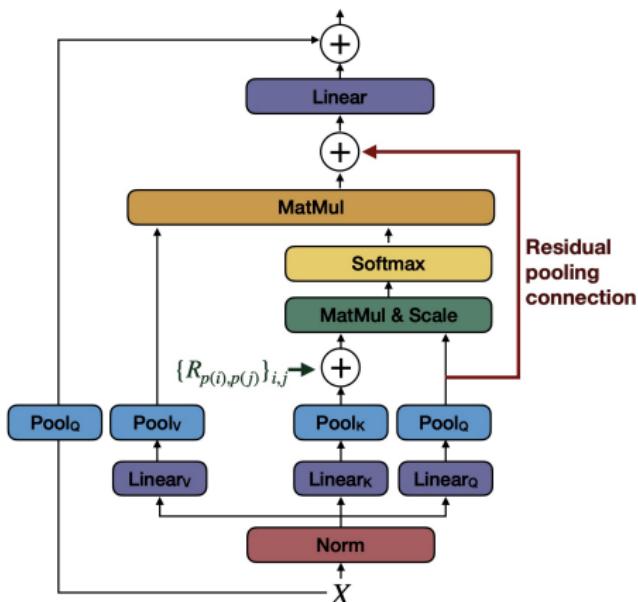
$$e_{ij} = \frac{Q_i(K_j + R_{p(i), p(j)})}{\sqrt{d}} \quad (6)$$

why?

- 1 Compute all (original) e_{ij} in a single matrix multiplication
- 2 Avoid broadcasting relative position representations

Residual Pooling Connection

$$Z := \text{Attn}(Q, K, V) + Q \quad (7)$$



Pooling Attention vs. Window Attention

Local aggregation then **global** self-attention computation vs. computing self-attention **locally** within non-overlapping windows

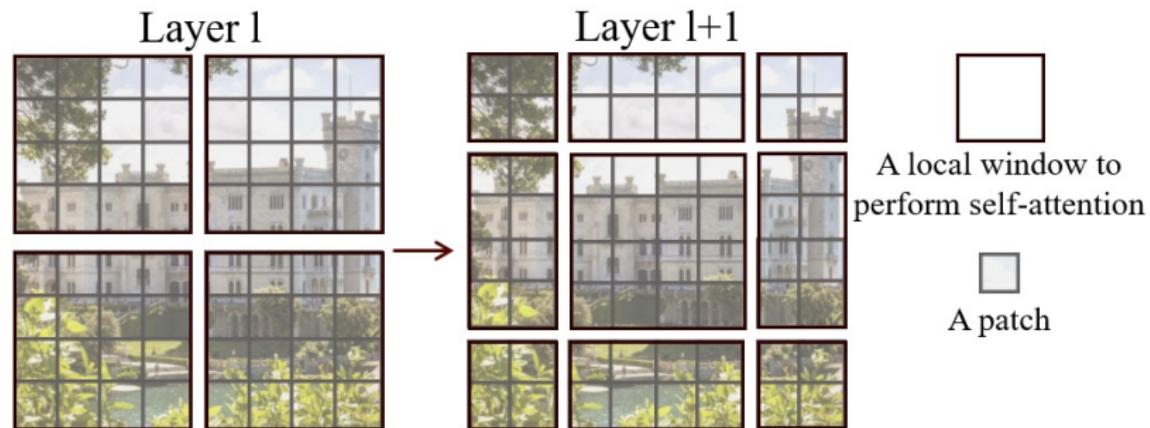


Figure: An illustration of the *shifted window* approach in Swin

Experimental Results on Kinetics-400

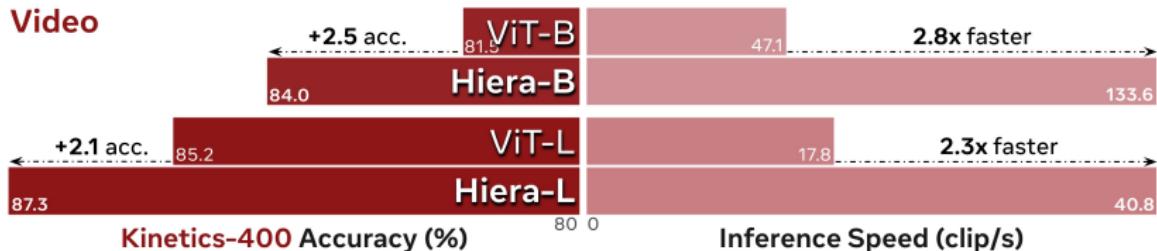
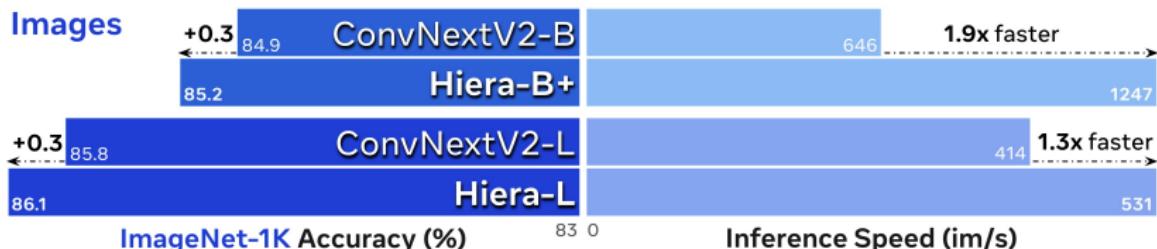
model	pre-train	top-1	top-5	FLOPs×views	Param
SlowFast 16×8 +NL [23]	-	79.8	93.9	$234 \times 3 \times 10$	59.9
X3D-XL [22]	-	79.1	93.9	$48.4 \times 3 \times 10$	11.0
MoViNet-A6 [45]	-	81.5	95.3	$386 \times 1 \times 1$	31.4
MViTv1, 16×4 [21]	-	78.4	93.5	$70.3 \times 1 \times 5$	36.6
MViTv1, 32×3 [21]	-	80.2	94.4	$170 \times 1 \times 5$	36.6
MViTv2-S , 16×4	-	81.0	94.6	$64 \times 1 \times 5$	34.5
MViTv2-B , 32×3	-	82.9	95.7	$225 \times 1 \times 5$	51.2
ViT-B-VTN [59]	IN-21K	78.6	93.7	$4218 \times 1 \times 1$	114.0
ViT-B-TimeSformer [3]		80.7	94.7	$2380 \times 3 \times 1$	121.4
ViT-L-ViViT [1]		81.3	94.7	$3992 \times 3 \times 4$	310.8
Swin-L \uparrow 384^2 [56]		84.9	96.7	$2107 \times 5 \times 10$	200.0
MViTv2-L\uparrow 312^2, 40×3		86.1	97.0	$2828 \times 3 \times 5$	217.6

Are They Necessary?

Hiera: A Hierarchical Vision Transformer without
Bells-and-Whistles³

- Observation:
Vision-specific modules make models slower
 - e.g., attention pooling, relative positional embedding
- Question:
Why should we slow down our architecture to add the spatial biases?
- Hypothesis:
Use MAE pretraining to teach ViTs spatial reasoning

Test Hypothesis on MViTv2



Hiera Setup

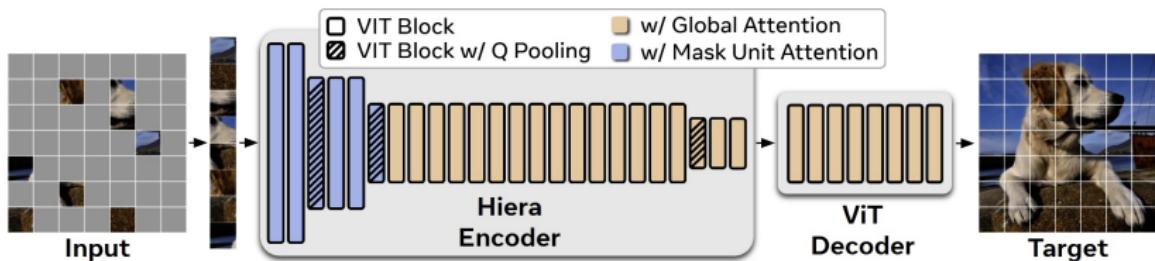
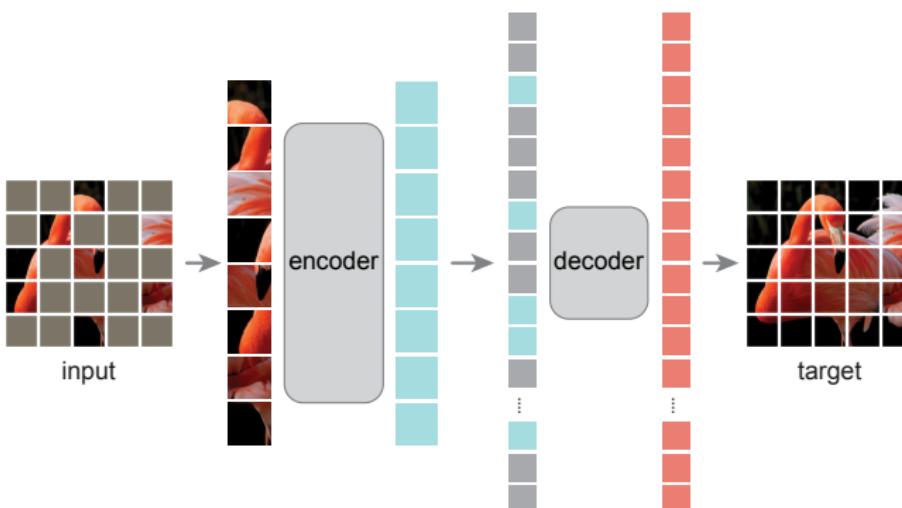


Figure: Local attention within “mask units” for the first two stages and global attention for the rest

MAE Recap

- encoder: e.g., ViT, applied to **visible, unmasked** patches
- decoder: lightweight, **independent** of the encoder design
- reconstruction target: **pixel** values for masked tokens

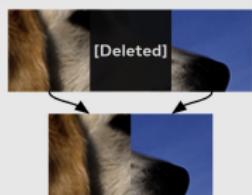


MAE for Hierarchical Models

(a) Use Mask Units instead of tokens.



(b) Problem: MAE *deletes* mask units.

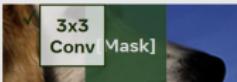


This **breaks the 2D grid**, causing errors for hierarchical models (e.g., w/ convs).



Potential Solutions

(c) MaskFeat: Fill with [mask].



Not sparse: *VERY* slow training.

(d) Baseline: Separate units & pad.



Sparse, but padding has overhead.

(e) Hiera: Just set *kernel size = stride*.

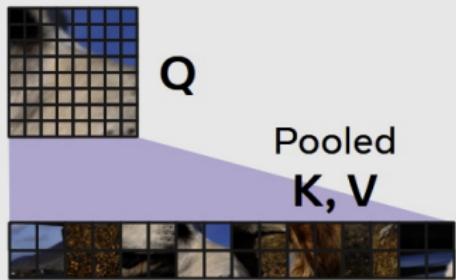


Sparse, no overhead, simple.

Figure: Random mask units rather than tokens

Mask Unit Attention

(a) MViTv2: Pooling Attn



(b) Hiera: Mask Unit Attn

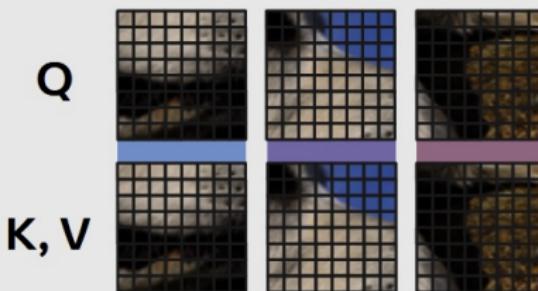


Figure: Mask Unit Attention performs local attention within mask units

Simplifying MViTv2

Setting	Image		Video	
	acc.	im/s	acc.	clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	<u>85.6</u>	253.3	<u>85.3</u>	20.7
b. replace convs with maxpools *	84.4	99.9 [†]	84.1	10.4 [†]
c. delete stride=1 maxpools *	85.4	309.2	84.3	26.2
d. set kernel size equal to stride	85.7	369.8	85.5	29.4
e. delete q attention residuals	<u>85.6</u>	374.3	85.5	29.8
f. replace kv pooling with MU attn	<u>85.6</u>	531.4	85.5	40.8

Experimental Results on Kinetics-400

backbone	pretrain	acc.	FLOPs (G)	Param
ViT-B	MAE	81.5	$180 \times 3 \times 5$	87M
Hiera-B	MAE	<u>84.0</u>	102 $\times 3 \times 5$	51M
Hiera-B+	MAE	85.0	<u>133</u> $\times 3 \times 5$	<u>69M</u>
MViTv2-L	-	80.5	377 $\times 1 \times 10$	<u>218M</u>
MViTv2-L	MaskFeat	84.3	377 $\times 1 \times 10$	<u>218M</u>
ViT-L	MAE	<u>85.2</u>	$597 \times 3 \times 5$	305M
Hiera-L	MAE	87.3	<u>413</u> $\times 3 \times 5$	213M
ViT-H	MAE	86.6	$1192 \times 3 \times 5$	633M
Hiera-H	MAE	87.8	1159 $\times 3 \times 5$	672M

Future Work

- Perception (multimodal) to action learning for robotics
- Functional perspective: proactive conversational agent (demo)
- When and how to perform actions is key given good perception representation
- Can be formulated as a canonical robotic control problem

References

- [1] **ICCV 2021** - Multiscale Vision Transformers
- [2] **CVPR 2022** - MViTv2: Improved Multiscale Vision Transformers for Classification and Detection
- [3] **CVPR 2022** - Masked Autoencoders Are Scalable Vision Learners
- [4] **ICML 2023** - Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles

Thank you very much!
Q&A