JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Efficient and Intelligent Computing Lab

Castling-ViT: Compressing Self-Attention via Switching Towards Linear-Angular Attention During Vision Transformer Inference

Haoran You[1*], Yunyang Xiong[2*], Xiaoliang Dai[2], Bichen Wu[2], Peizhao Zhang[2], Haoqi Fan[2], Peter Vajda[2], and Yingyan (Celine) Lin[1]

[1]Georgia Institute of Technology
[2]Meta Research

Session ID: WED-PM-199

Georgia Tech School of Computer Science
College of Computing

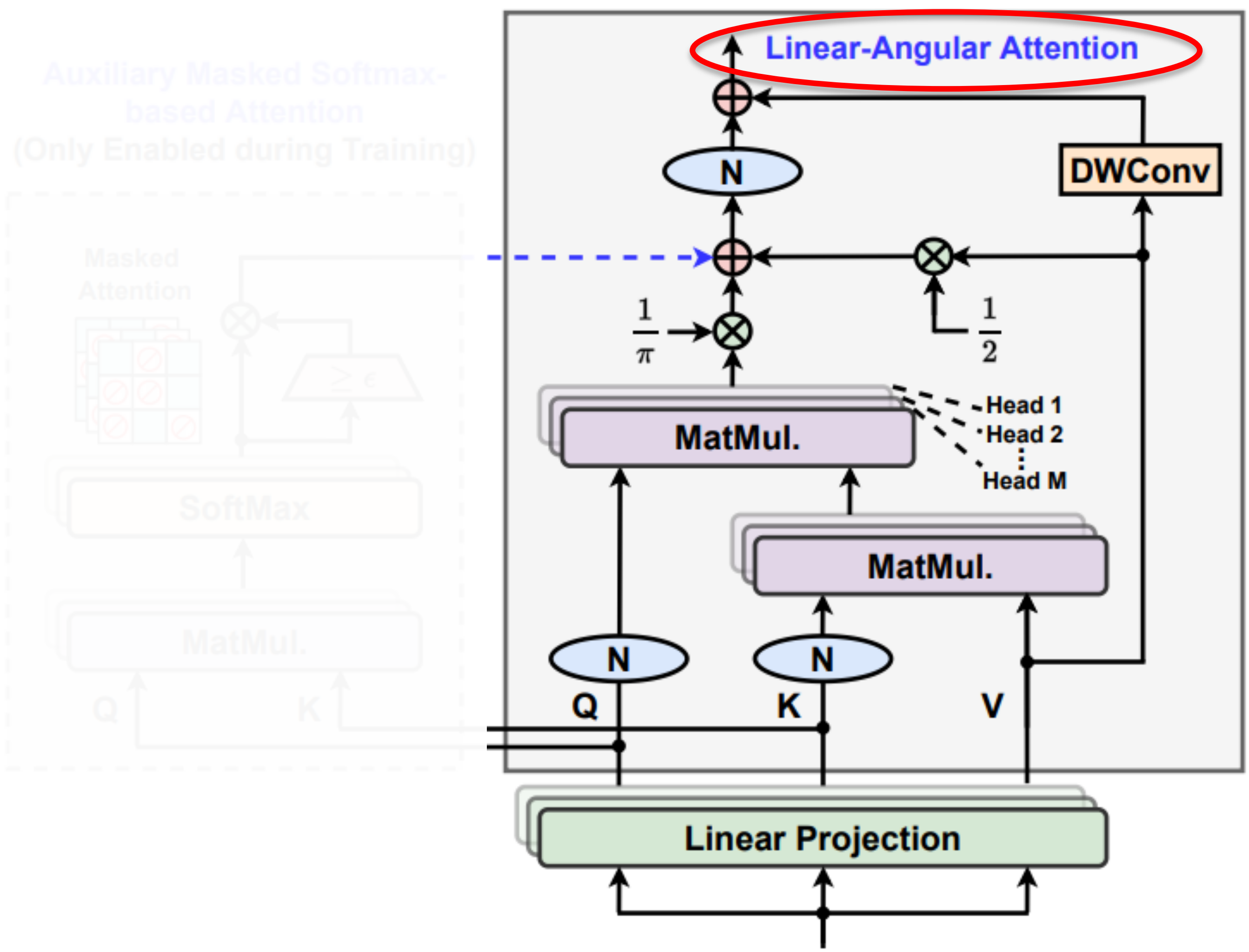
Meta

Quick Review

(1 min)

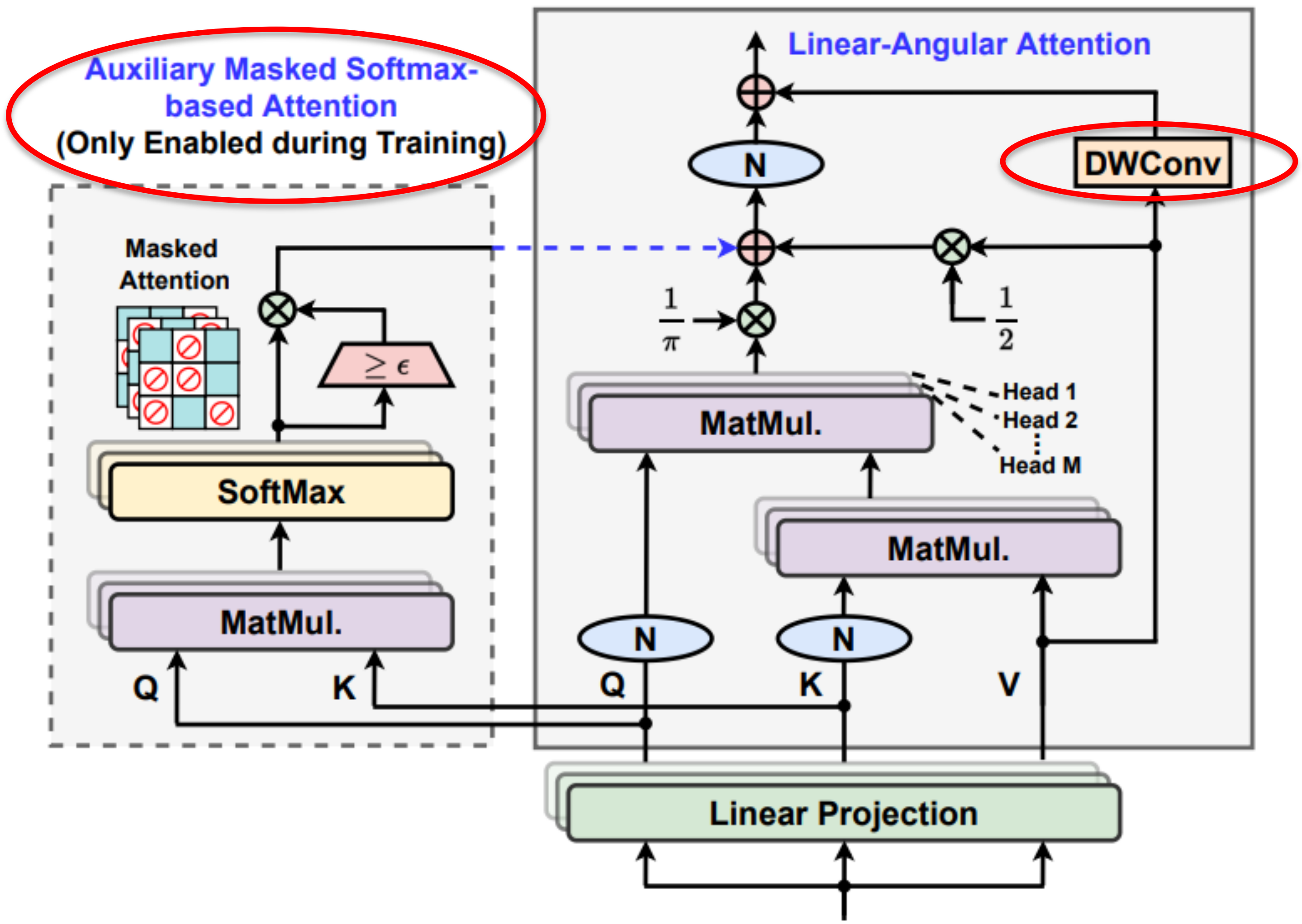
Executive Summary

- Contribution 1: Linear-angular attention
  - Decompose angular kernels into linear terms and high-order residuals

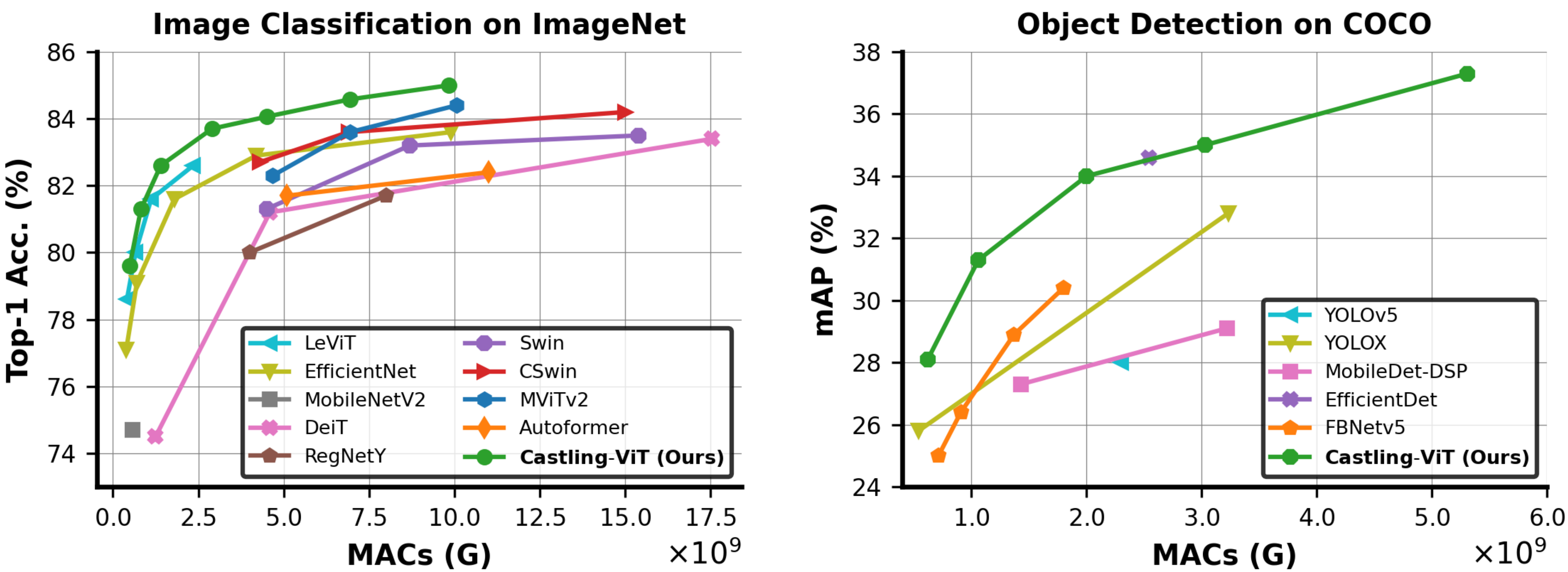


Executive Summary

- **Contribution 1: Linear-angular attention**
  - **Decompose angular kernels into linear terms and high-order residuals**
- **Contribution 2: Approximate high-order residuals**
  - **Two parameterized modules to approximate high-order residuals**



Executive Summary

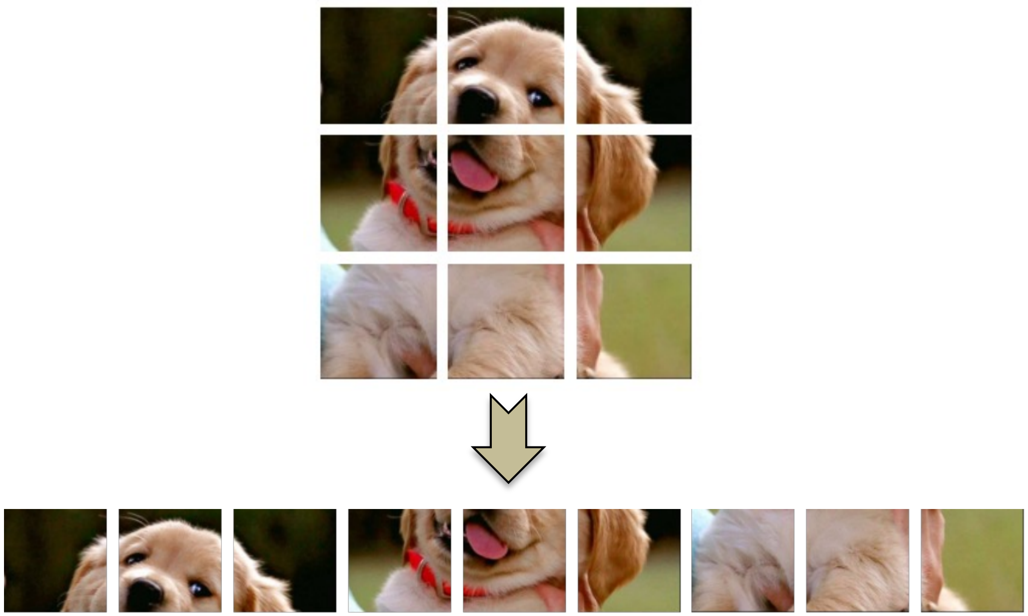


- **Experiment results** (over vanilla softmax ViTs)
  - 1.8% higher accuracy or 40% MACs reduction on classification tasks
  - 1.2 higher mAP on detection tasks

Full Presentation
(7 min)

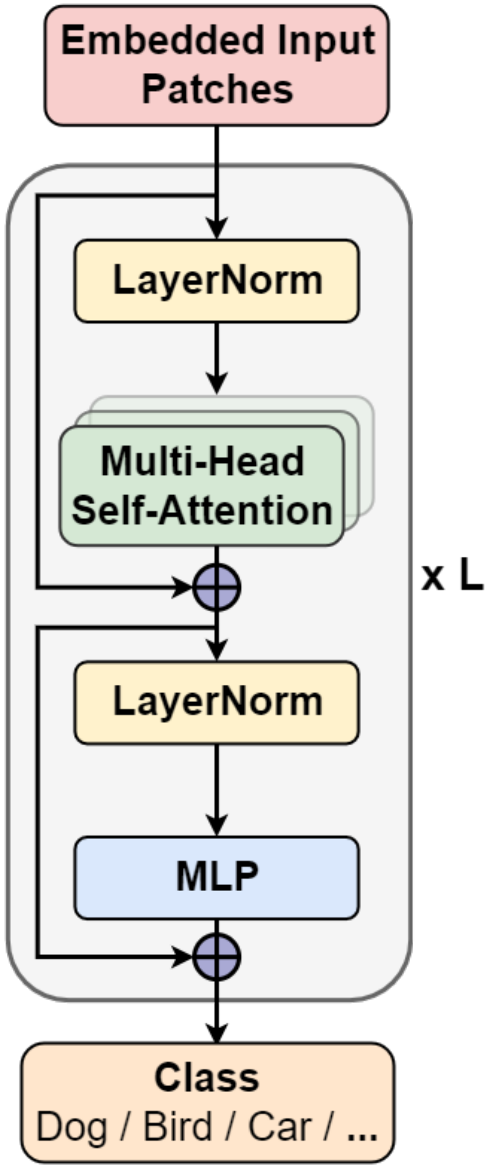
Background of Vision Transformer (ViTs)

- ViTs achieve **SOTA performance** on various vision tasks
  - **Input**: 2D image → input tokens/patches



Input Tokens

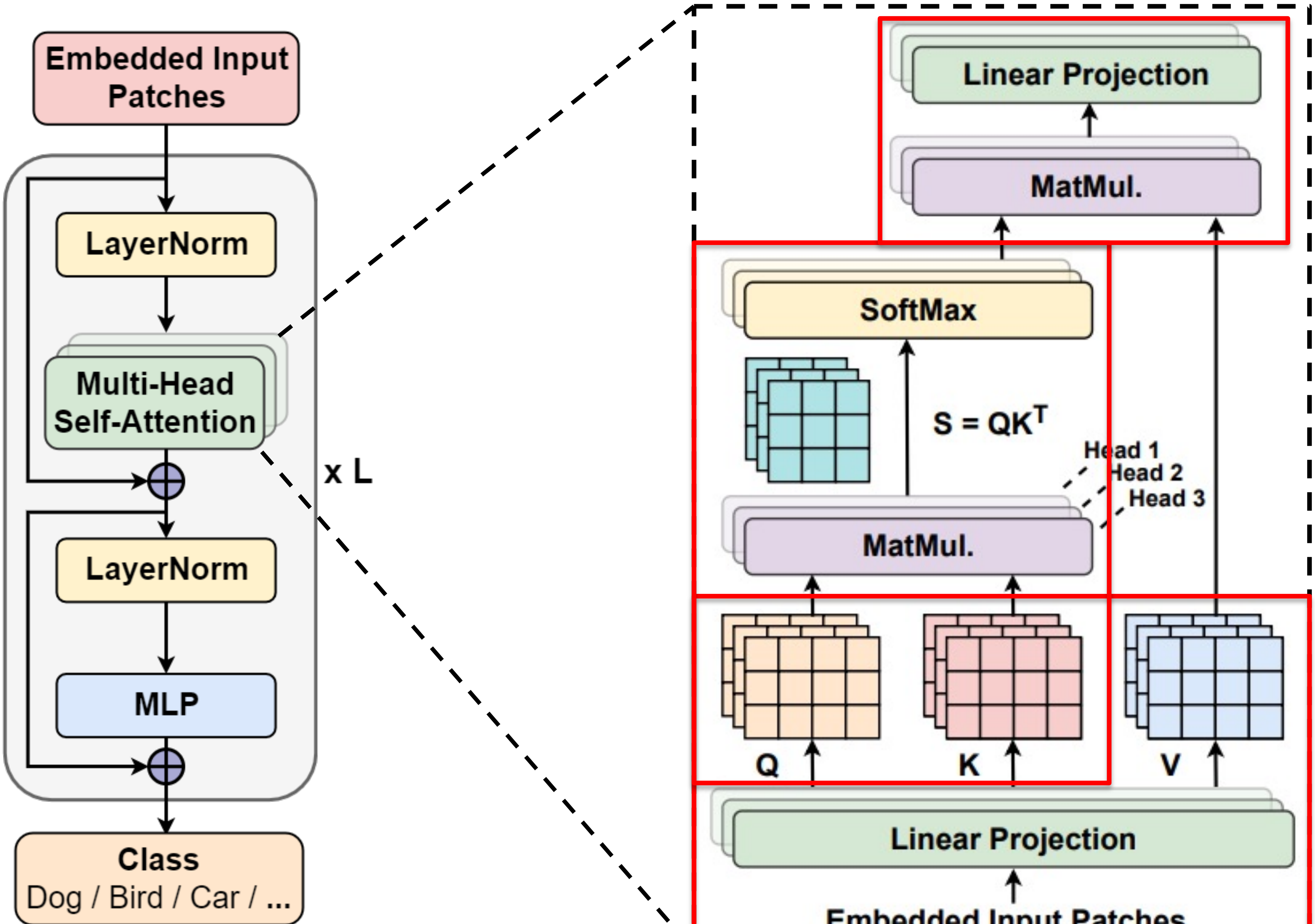
  - **Core Model**: Self-Attention and MLP



ViT Models

Background of Vision Transformer (ViTs)

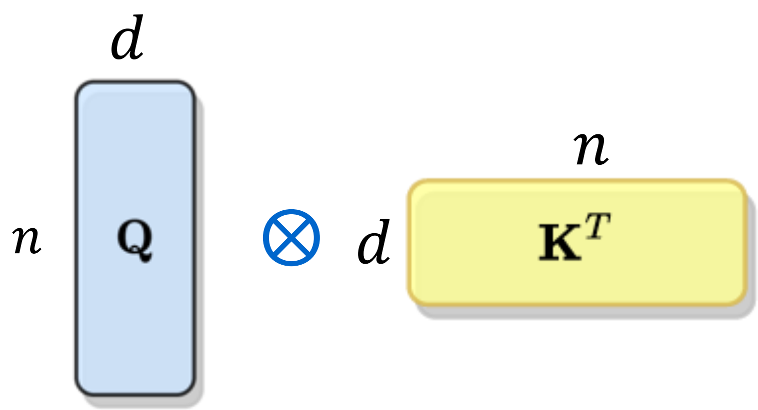
- **Illustrate the core self-attention module**



Bottlenecks of ViTs

- **Self-attention is one of the runtime bottleneck** [1,2,3]
- **Why?**

$$O_{Attn} = softmax(\boldsymbol{Q} \cdot \boldsymbol{K}^T) \cdot \boldsymbol{V}$$



[1] H. You et al, HPCA 2023
[2] J. Dass et al, HPCA 2023
[3] H. Fan, et al, MICRO 2022

Attention's quadratic complexity:

Time: $\boldsymbol{O}(\boldsymbol{n}^2 d)$, Memory: $\boldsymbol{O}(\boldsymbol{n}^2)$

n : number of tokens (e.g., 196~16K)

d : hidden dimension per head (e.g., 64)

Previous Efficient Linear Attention

- **Decompose the similarity measurement function, e.g., softmax or EXP(·), into separate kernel embeddings**

$$softmax(\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V} \approx \boxed{\{\phi(\boldsymbol{Q})\phi(\boldsymbol{K})^T\}}\,\boldsymbol{V} = \phi(\boldsymbol{Q})\boxed{\{\phi(\boldsymbol{K})^T\boldsymbol{V}\}}$$

$\mathbf{O}(\boldsymbol{n}^2 d)$ — Separate Kernel embeddings → $\mathbf{O}(\boldsymbol{n}d^2)$

Quadratic → **Linear**

Motivating Research Question

- **Previous SOTA Linear or Local Attention**

$$softmax(QK^T)V \approx \{\phi(Q)\phi(K)^T\}V = \phi(Q)\{\phi(K)^TV\}$$

$O(n^2d)$ Linear or Local Attention $O(nd^2)$

Kernel embeddings

Quadratic Linear

- **But all these linear or local approximations suffer from large accuracy drops; Our research question is:**

Can ViTs learn both global and local context while being more efficient during inference?

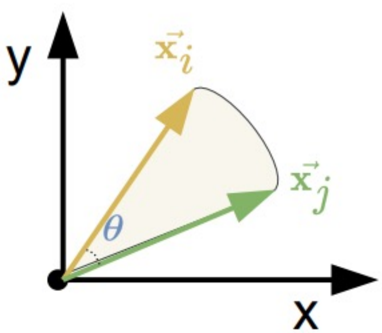
Proposed Castling-ViT

Castling-ViT: Linear-Angular Attention

- **Angular kernel from a spectral perspective**
  - **Spectral angle between two vectors:**

$$\theta(x_i, x_j) = \arccos\left(\frac{\langle x_i, x_j \rangle}{\|x_i\| \cdot \|x_j\|}\right)$$

Example of 2D inputs:



Angle in a 2D input space

Castling-ViT: Linear-Angular Attention

- **Angular kernel from a spectral perspective**
  - Spectral angle between two vectors:

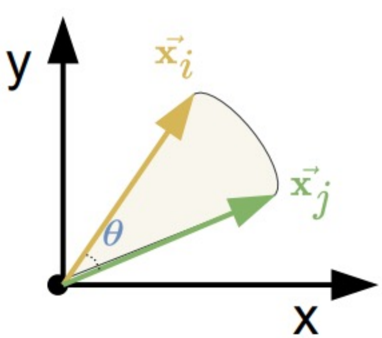
$$\theta(\mathrm{x}_i, \mathrm{x}_j) = \arccos\left(\frac{\langle \mathrm{x}_i, \mathrm{x}_j \rangle}{\|\mathrm{x}_i\| \cdot \|\mathrm{x}_j\|}\right)$$

  - **We design the angular kernel as a similarity measurement function**

$$Sim(\boldsymbol{Q}_i, \boldsymbol{K}_j) = 1 - \frac{1}{\pi} \cdot \theta(\boldsymbol{Q}_i, \boldsymbol{K}_j) = \boxed{\phi(\boldsymbol{Q}_i) \cdot \phi(\boldsymbol{K}_j)}$$

Decomposable

Example of 2D inputs:



Angle in a 2D input space

If $\boldsymbol{Q}_i$ and $\boldsymbol{K}_j$ are aligned, then $\theta \to 0$ and $Sim(\boldsymbol{Q}_i, \boldsymbol{K}_j) \to 1$

If $\boldsymbol{Q}_i$ and $\boldsymbol{K}_j$ are opposite, then $\theta \to \pi$ and $Sim(\boldsymbol{Q}_i, \boldsymbol{K}_j) \to 0$

Similarity Measurement

Castling-ViT: Linear-Angular Attention

- **Angular kernel from a spectral perspective**
  - Spectral angle between two vectors:

$$\theta(\mathrm{x}_i, \mathrm{x}_j) = \arccos\left(\frac{\langle \mathrm{x}_i, \mathrm{x}_j \rangle}{\|\mathrm{x}_i\| \cdot \|\mathrm{x}_j\|}\right)$$

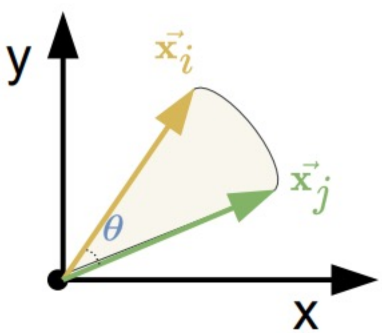
  - **We design the angular kernel as a similarity measurement function**

$$Sim(\boldsymbol{Q}_i, \boldsymbol{K}_j) = 1 - \frac{1}{\pi} \cdot \theta(\boldsymbol{Q}_i, \boldsymbol{K}_j) = \boxed{\phi(\boldsymbol{Q}_i) \cdot \phi(\boldsymbol{K}_j)}$$

Decomposable

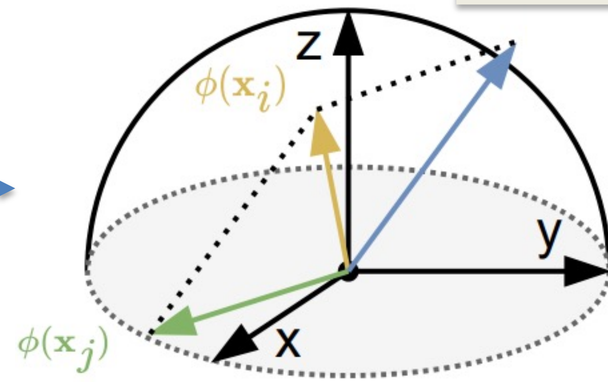
$$\left\|\phi(\mathrm{x}_i) - \phi(\mathrm{x}_j)\right\|^2 = \frac{2}{\pi} \cdot \theta\left(\mathrm{x}_i, \mathrm{x}_j\right)$$

Example of 2D inputs and 3D features:



mapped onto
the sphere of radius 1

$$\|\phi(\mathrm{x}_i)\|^2 = Sim(\mathrm{x}_i, \mathrm{x}_i) = 1$$



Angle in a 2D input space

Distance in a 3D feature space

Castling-ViT: Linear-Angular Attention

- **Angular kernel from a spectral perspective**
  - **Expansion of the angular kernel**

$$Sim(\boldsymbol{Q}_i, \boldsymbol{K}_j) = \underbrace{\frac{1}{2} + \frac{1}{\pi} \cdot (\boldsymbol{Q}_i \cdot \boldsymbol{K}_j)}_{\text{Linear-angular terms}} + \underbrace{\frac{1}{\pi} \cdot \sum_{k=1}^{\infty} \frac{(2k)!}{2^{2k}(k!)^2(2k+1)} (\boldsymbol{Q}_i \cdot \boldsymbol{K}_j)^{2k+1}}_{\text{High-order residuals when } \boldsymbol{k} \geq 1 \text{ (non-linear)}}$$

Keep | **Approximate**

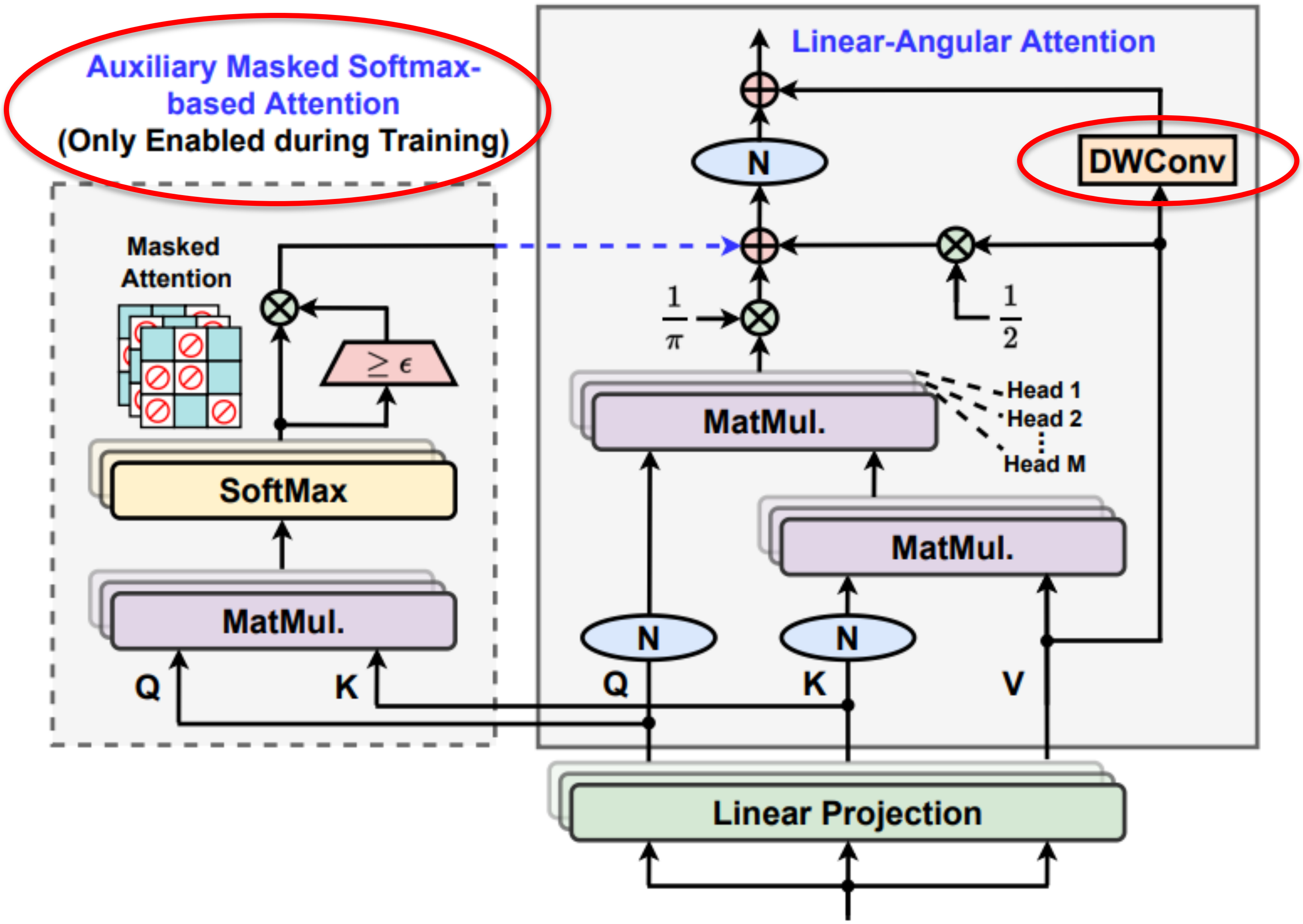
Referred to as: **Linear-Angular Attention**

Castling-ViT: High-Order Residuals Approximation

- We leverage parameterized DNN modules to approximate it
  - A learnable depthwise convolution (DWConv)
    - To capture a strong inductive bias in neighboring tokens
  - An auxiliary masked softmax-based attention
    - To capture global similarity for nonadjacent tokens

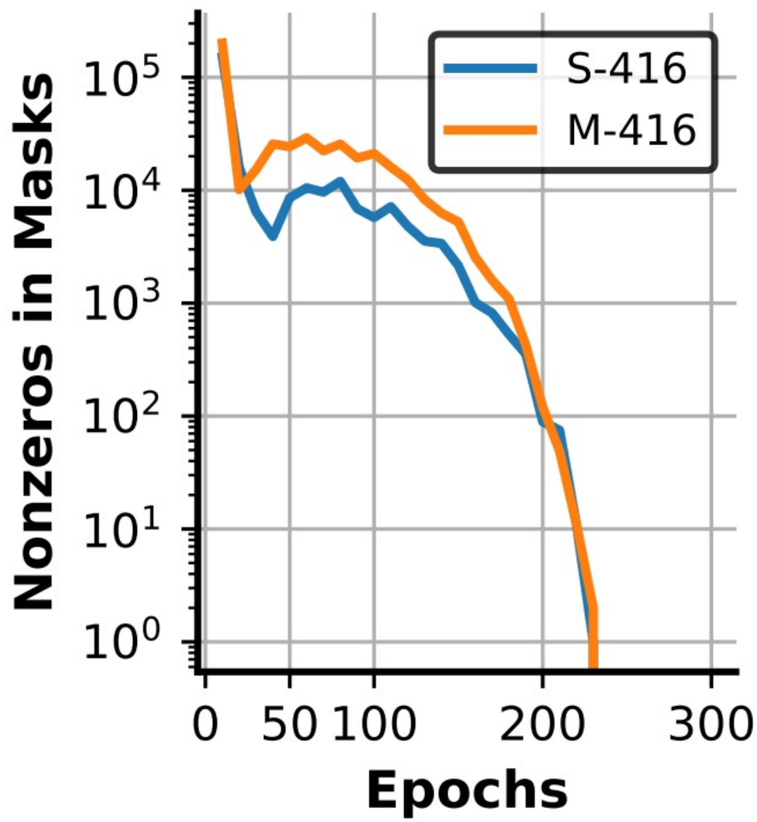
Castling-ViT: High-Order Residuals Approximation

- We leverage parameterized DNN modules to approximate it
  - A learnable depthwise convolution (DWConv)
  - An auxiliary masked softmax-based attention



Castling-ViT: High-Order Residuals Approximation

- ***Question***: **Softmax-based attention is costly?**
  - **It can be dropped during inference! → "Castling"**



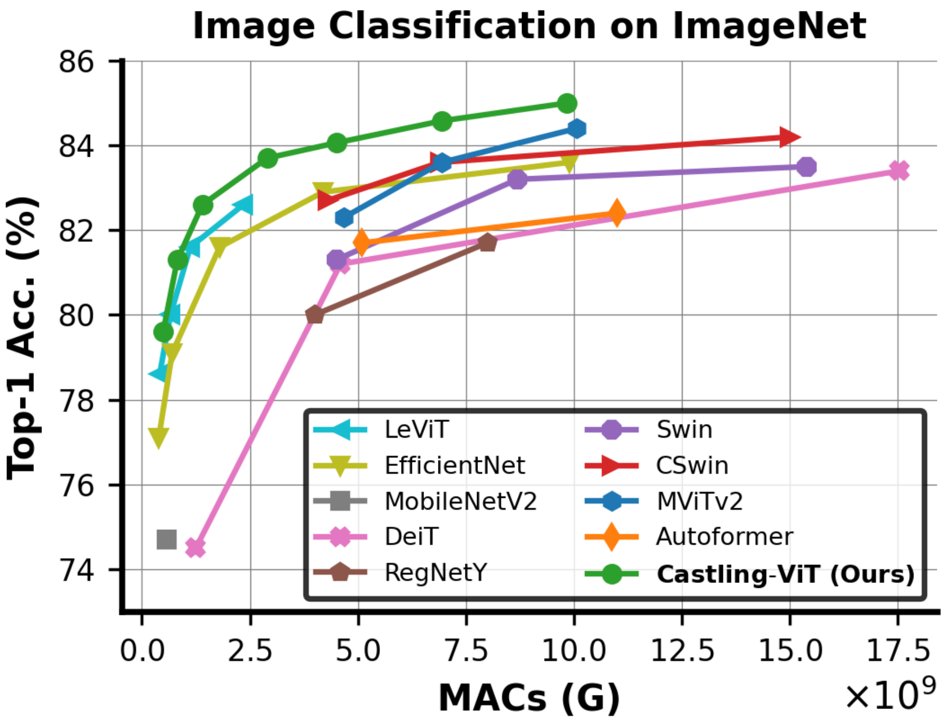
Visualizing the trajectories of nonzeros
in auxiliary masks during training detection models on COCO.

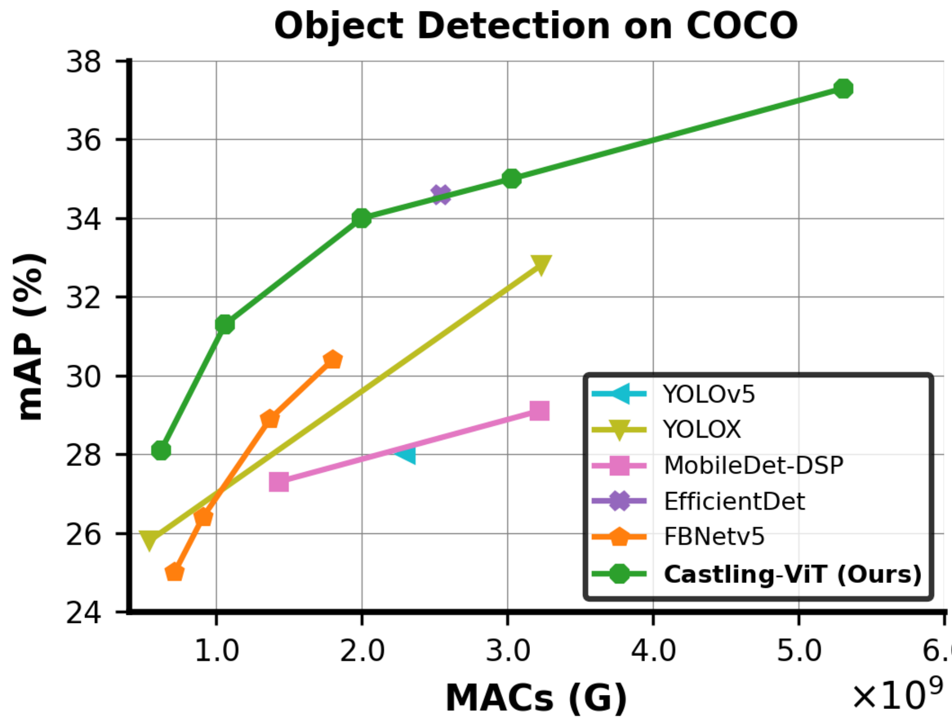
Evaluation Setup

- Evaluation Setup
  - Three Tasks
    - Classification (CLS) / Detection (DET) / Segmentation (SEG)
  - Datasets
    - ImageNet / COCO / ADE20K
  - Models
    - LeViT, MViTv2, DeiT / PicoDet / Mask2former

- Benchmark Baselines
  - CLS
    - LeViT, MViTv2, DeiT, Swin, CSwin, PVT, etc
  - DET
    - FBNetV5, YOLOX, YOLOv5, MobileDet, EfficientDet
  - SEG
    - Mask2former w/ ViT backbone

Evaluation: Castling-ViT over SOTA Baselines





- **Castling-ViT over SOTA baselines**
  - 1.8% higher accuracy or 40% MACs reduction on classification tasks
  - 2.2 higher mAP on detection tasks

Evaluation: Castling-ViT over SOTA Baselines

Mask2former w/	MAE Pretrain	Params (M)	MACs (G)	mIoU	mAcc	pAcc
ViT-Base	N	118	229 (182)	34.54	46.36	75.84
Castling-ViT-Base	N	118	**195 (147)**	**34.67**	**46.47**	**76.20**
ViT-Base	Y	118	229 (182)	47.92	61.00	83.02
Castling-ViT-Base	Y	118	**195 (147)**	**48.44**	**61.82**	**83.29**

- **Castling-ViT over SOTA baselines**
  - 0.52% higher mIoU and 15% MACs reduction on segmentation tasks

Summary

In this work, we

- **Propose a framework called Castling-ViT, which trains both quadratic and linear attention while switching to having only linear attention at inference**
- **Develop a new linear-angular attention leveraging angular kernels to close the accuracy gap with softmax attention**
- **Use two parameterized modules to approximate the high-order residuals to compensate the accuracy drop**

Acknowledge:

NSF RTML programs and the CoCoSys, one of the seven centers in JUMP 2.0 sponsored by DARPA

Project page:

SCAN ME