Real-time Core-Periphery Guided ViT with Smart Data Layout Selection on Mobile Devices

Zhihao Shu, Xiaowei Yu, Zihao Wu, Wenqi Jia, Yinchen Shi

Miao Yin, Tianming Liu, Dajiang Zhu, Wei Niu

Presented by: Zhihao Shu (Zhihao.Shu@uga.edu)

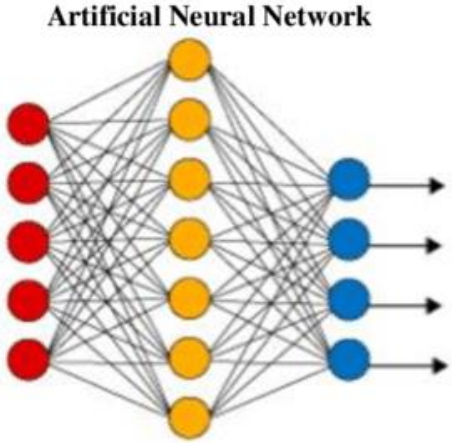
Accepted by NeurIPS'24

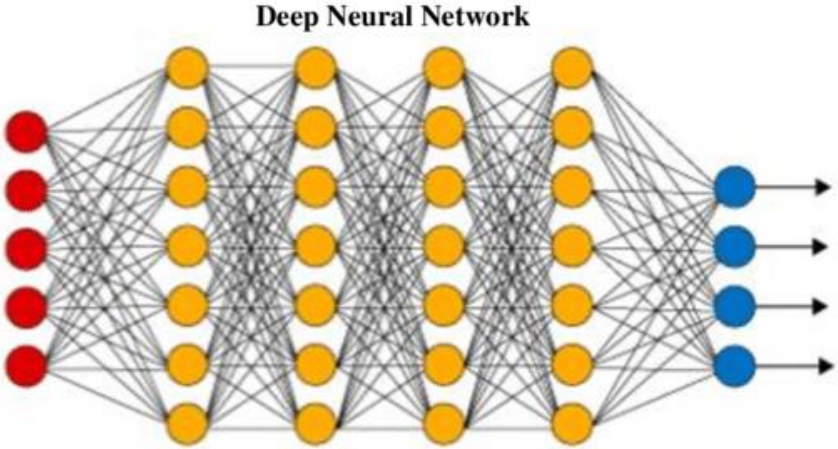
Real-Time DNN Inference On Mobile Device

Object Detection

Credit: www

Virtual Assistants





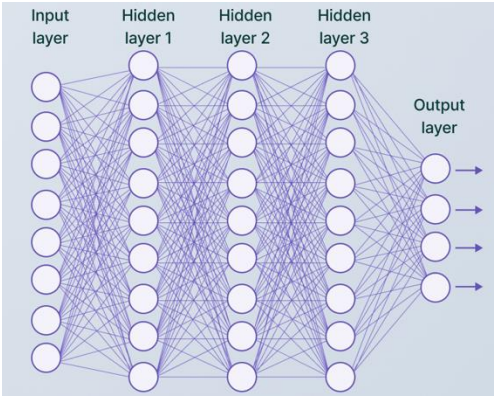
Segmentation

face.co/
Anything-Model

Execution Latency Matters !!!

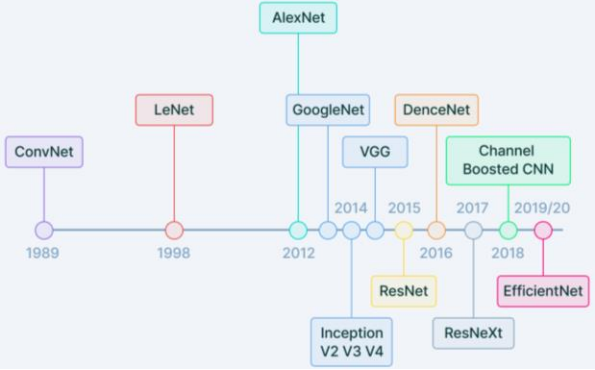
Evolution of DNN Architectures

Simple FFNs



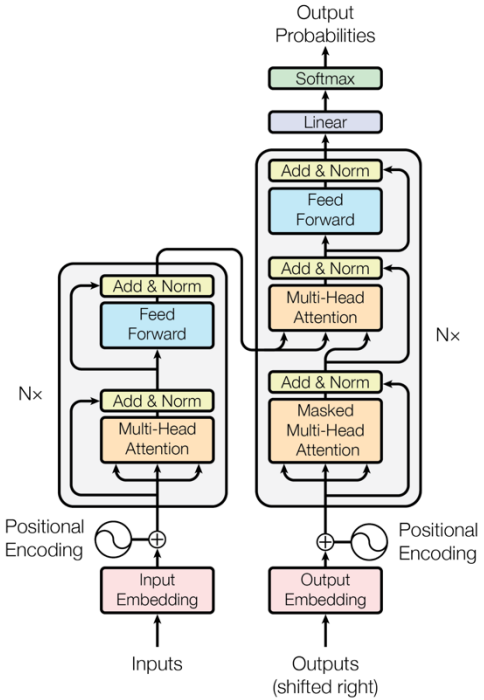
Credit: V7 Labs

Evolving Convolution Networks



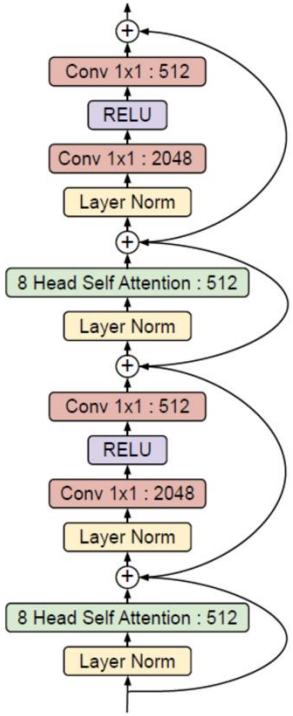
Credit: V7 Labs

Transformers



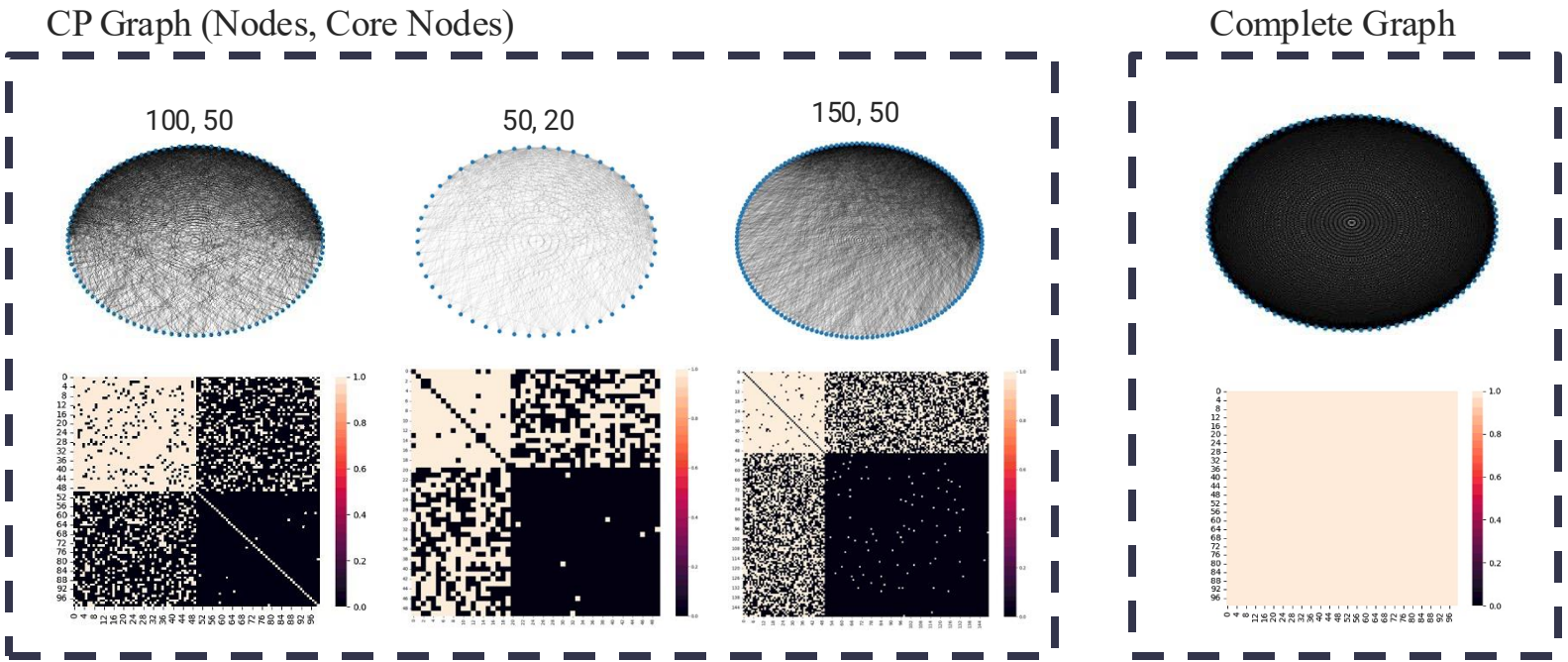
Credit: Attention is all your need

Transformer variants



Credit: devopedia.org

Core-Periphery Graph Generation



Core-Periphery Graph Generator, Verified by the CP Network Detection Algorithm [1]

[1] S. P. Borgatti and M. G. Everett. Models of core/periphery structures. Social Networks, 21, 375–395, 2000

Poor Performance on Mobile GPU

Models	Latency (ms)	Latency breakdown	
		Data Transformation (%)	Computation (%)
Swin	342	68.8	**31.2**
Cswin	703	64.5	**35.5**
ViT	421	76.3	**23.7**

Operator fusion can leads to less intermediate results, but cannot eliminate data transformation.

The data transformation overhead exceeds even the computational cost.

ECP-ViT Contributions

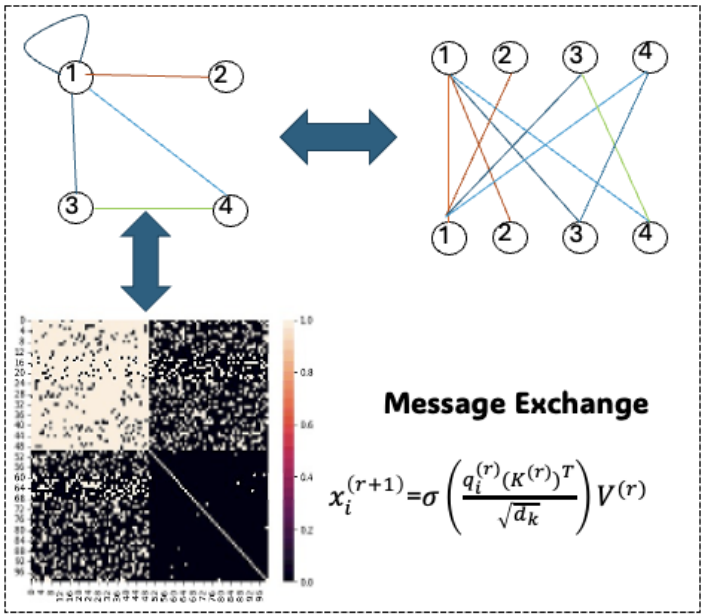
1. Incorporates Core-periphery to guide self-attention in ViTs
2. Designs a mechanism for eliminating layout transformation and selecting optimal layout
3. Builds ECP-ViT, a framework that combines algorithm and system design to achieve real-time performance

Compared to state-of-the-art frameworks -- TVM, MNN

Speedup

25x ***7.2x***

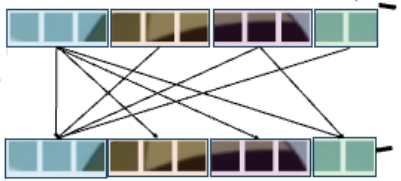
Core-Periphery Transformer (Contribution)



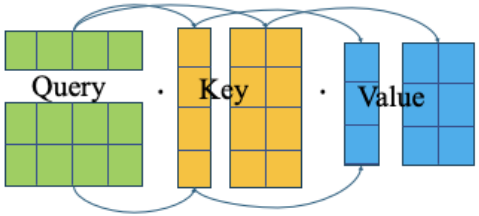
(a) Core-Periphery Graph Generation

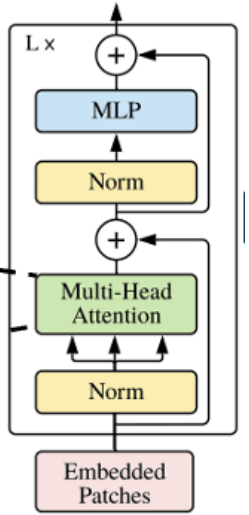
(b1) Self-Attention Re-schedule

$$CPAttention = softmax\left(CP\left(\frac{QK^T}{\sqrt{d_k}}\right)V\right)$$

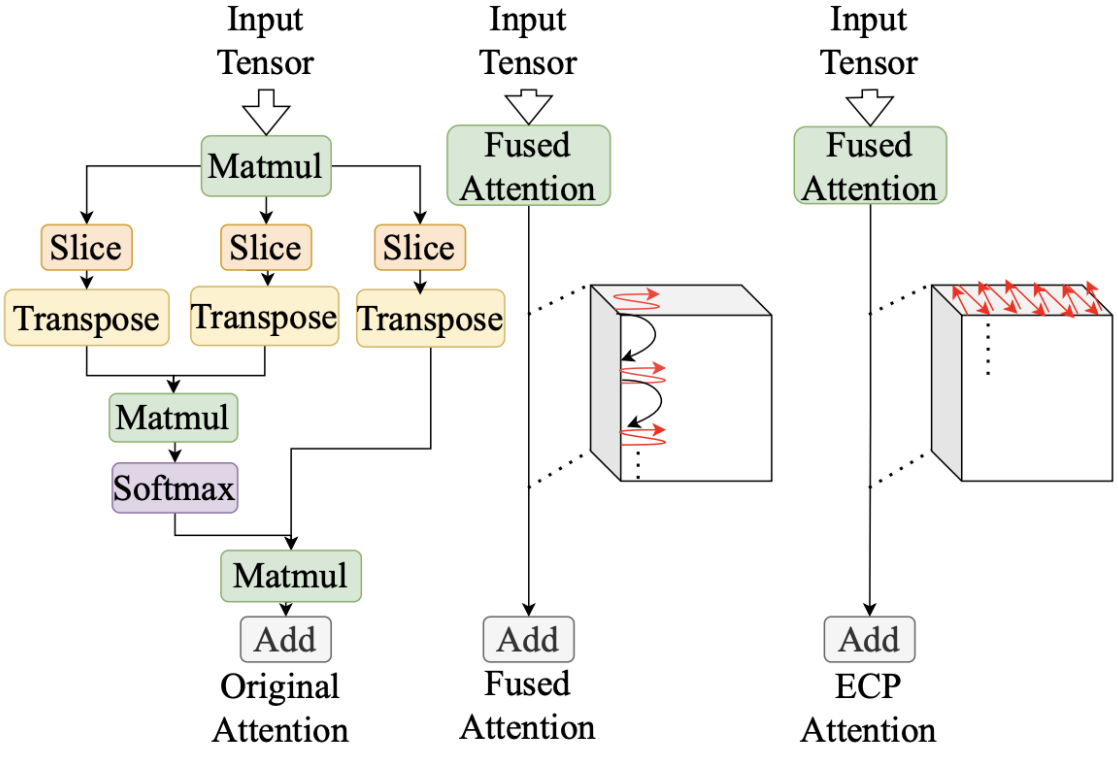


(b2) Core-Periphery Guided Matrix Multiplication





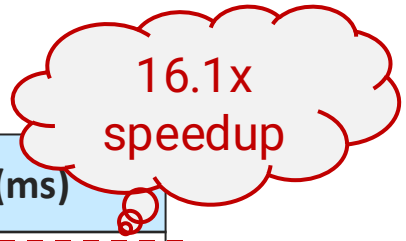
Smart Data Layout Selection *(Contribution)*



Results – Accuracy Comparison on ImageNet-21K

Models	Top-1 (%)	# Params.	# MACs
ViT-Base/16	83.9	86.6M	17.6G
PVT-Large	83.8	82.0M	11.84G
TNT-Base	84.1	66.0M	14.16G
DeiT-Base/16	84.2	86.6M	17.76G
ECP-ViT-Base	**84.6**	86.5M	16.96G

Results – Latency Comparison on OnePlus 11 (Smartphone)



Models	TVM (ms)	MNN (ms)	Ours (ms)
ECP-ViT-Tiny	380.1	110.1	**15.2**
ECP-ViT-Small	837.6	157.9	**31.1**
ECP-ViT-Base	2033	563.2	**122.8**

ECP-ViT enables real-time performance (< 33ms) on mobile devices for Vision Transformer

Thanks for listening!

Email: Zhihao.Shu@uga.edu