STRUCT Group Paper Reading

Image as Set of Points

ICLR 2023

Notable top 5%

Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, Yun Fu

Northeastern University, Boston
University of Illinois, Urbana-Champaign

Presented by Zejia Fan
2023.2.19

Backbone development

- MLP->CNN->Transformer->MLP?
  - 21年

    12月: “图像识别也是Transformer最强(ViT)”

    2月: “Transformer is All you Need”

    3月: “Attention is not All you Need”

    5月: “在MLP上的ViT并(MLPmixer)”

    5月: “Convolution比Transformer强”

    5月: “在MLP上加个门，跨越Transformer (Pay Attention to MLPs)”

CNN

- LeNet-5
  - Convolution & pooling

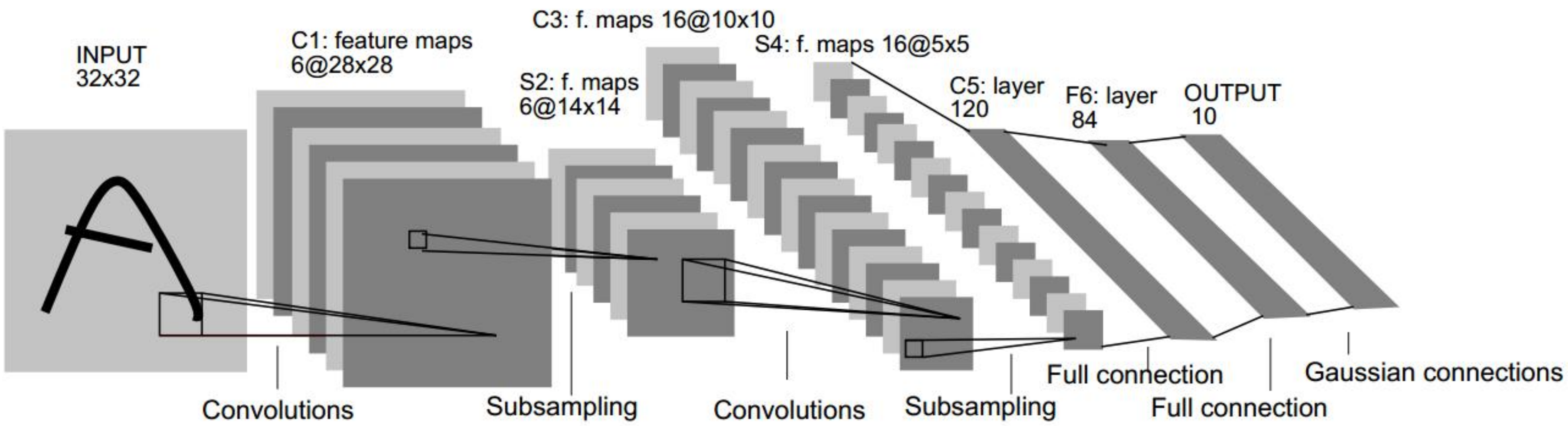
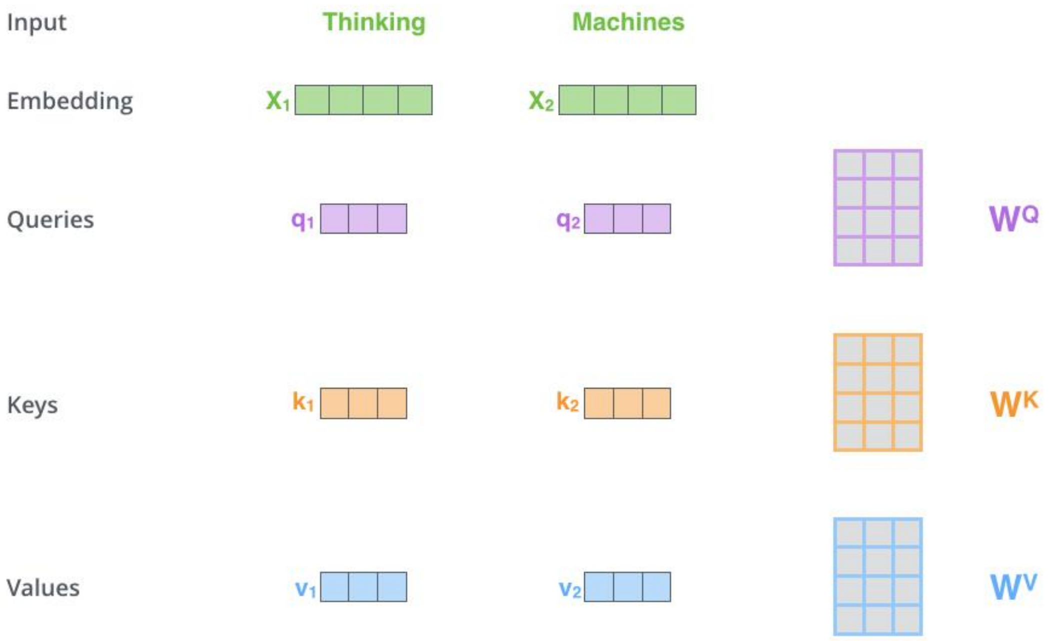
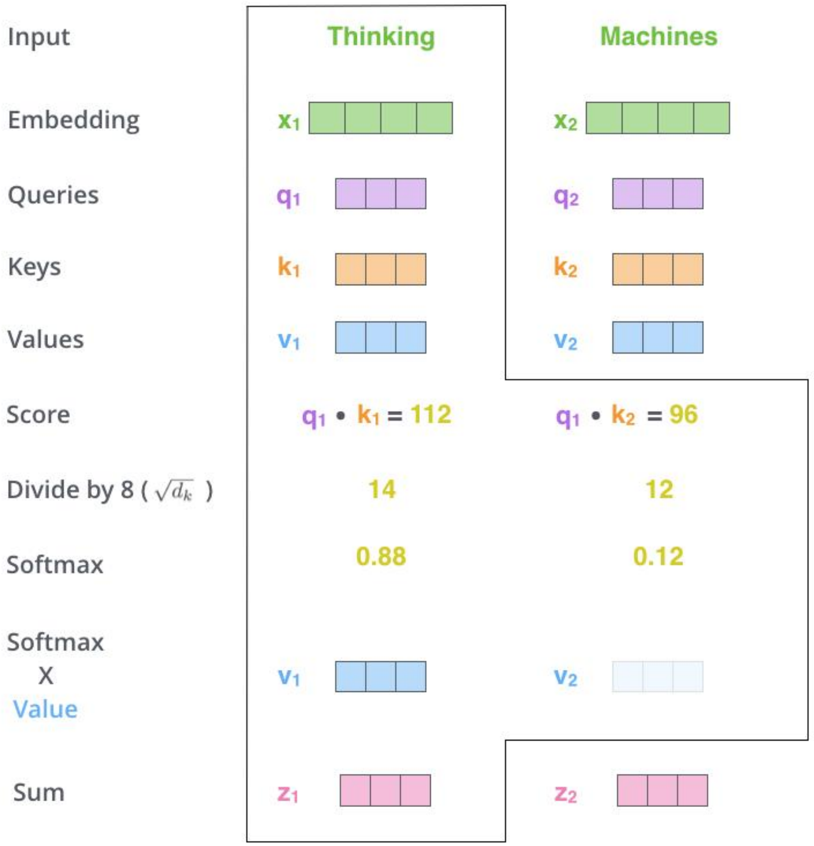


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

LeCun, Yann, et al. "Gradient-based learning applied to document recognition.“ 1998. cite 52538.

Attention





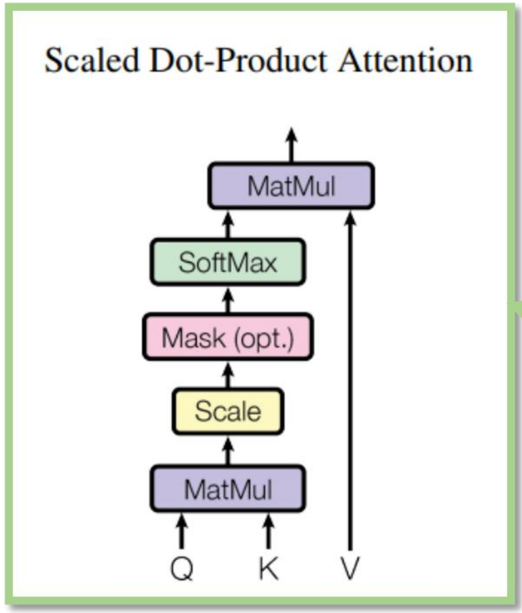
$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V$$

Transformer in NLP

N encoder layers and N decoder layers together form the transformer.

Several point:

- (Self-/Cross-) attention
- Feed forward
- Residual connection & norm
- Positional encoding



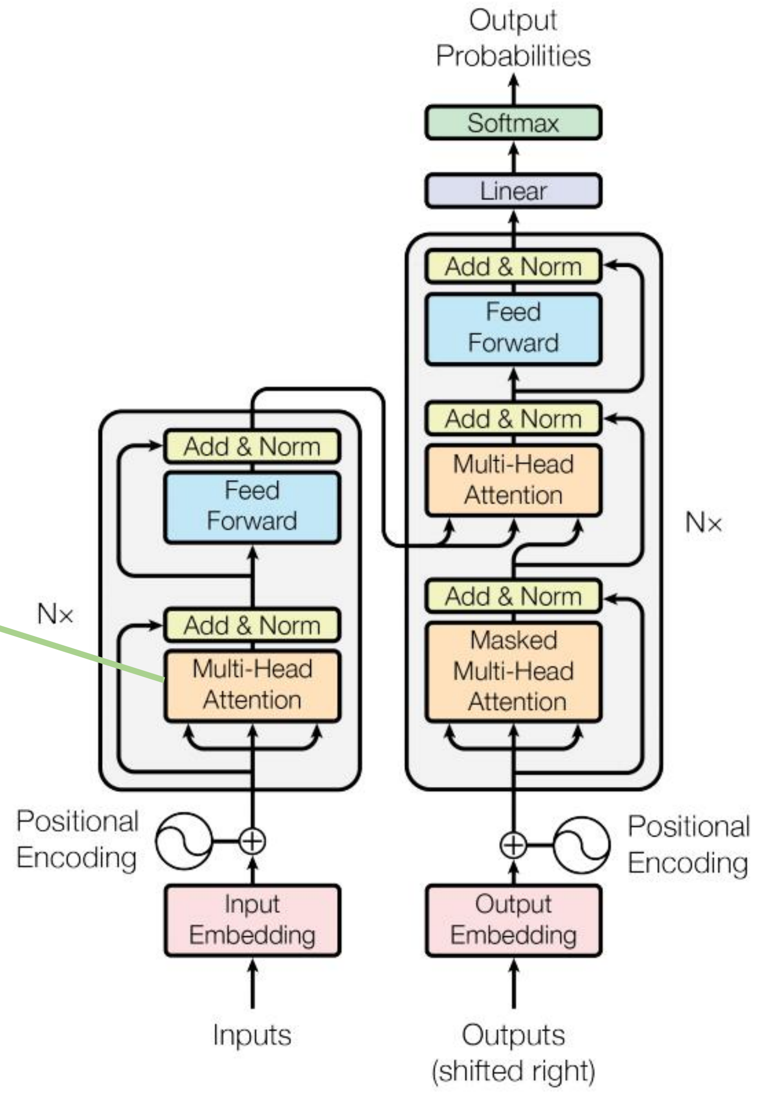
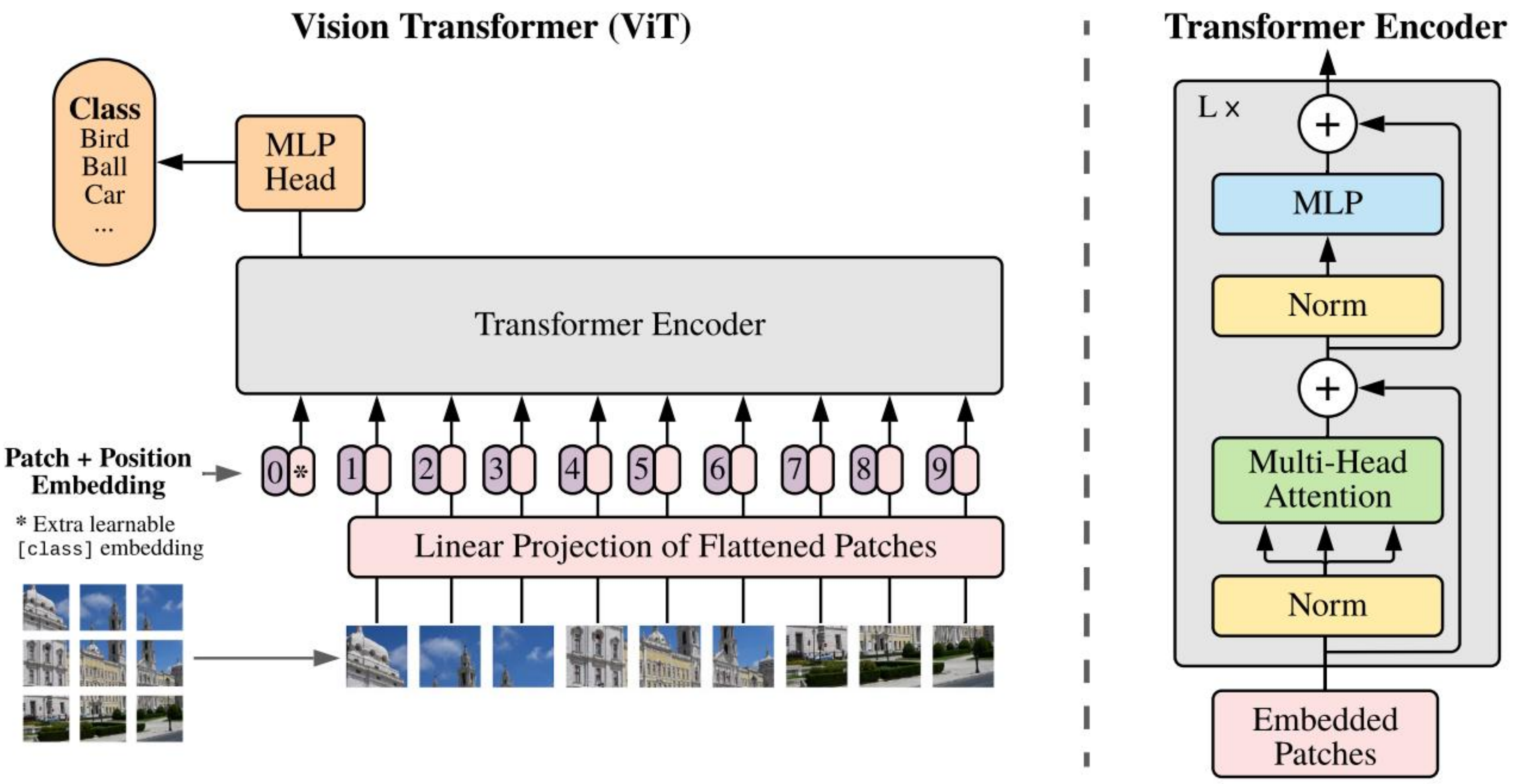


Figure 1: The Transformer - model architecture.

Ashish *et al.* Attention is all you need. *NIPS 2017*.

Vision Transformer



Alexey *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* 2021. cite 12042.

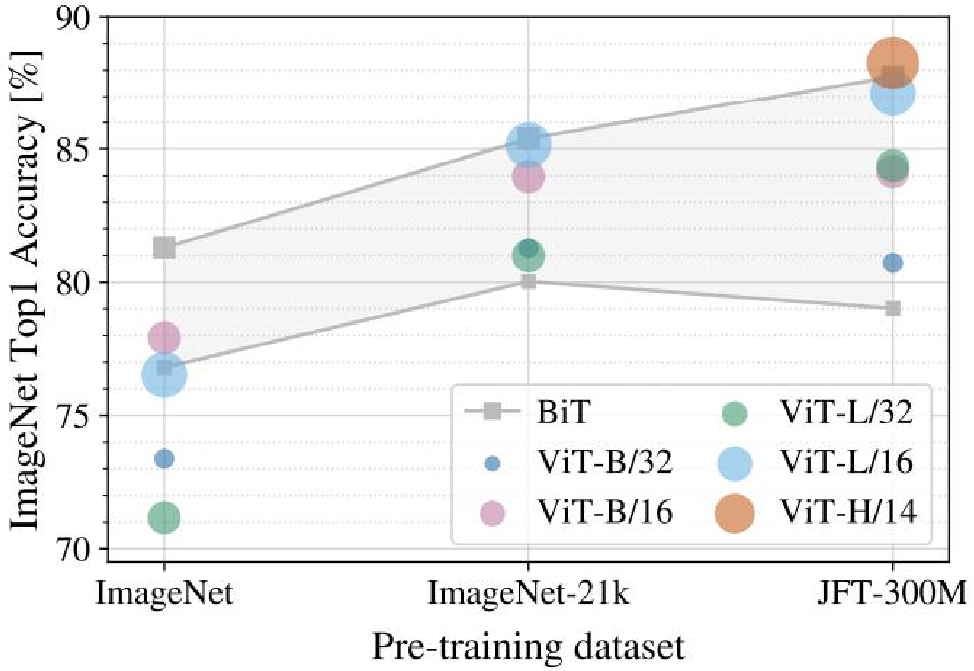
Vision Transformer

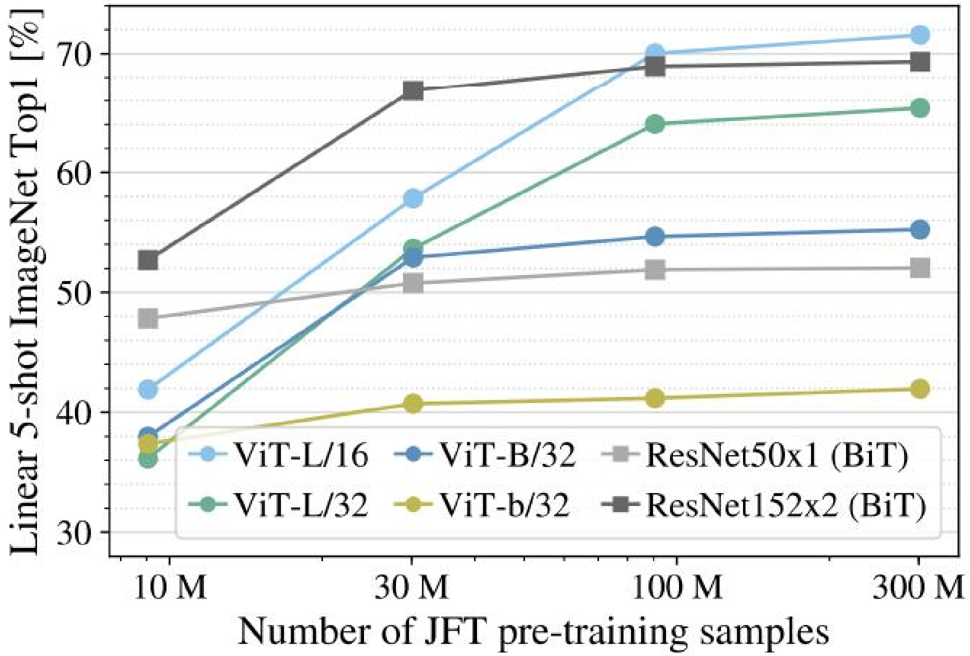
- JFT: 300M images
- ImageNet 21k: 14M images
- ImageNet: 1.3M images

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	–
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	–
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	–
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	–
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Vision Transformer

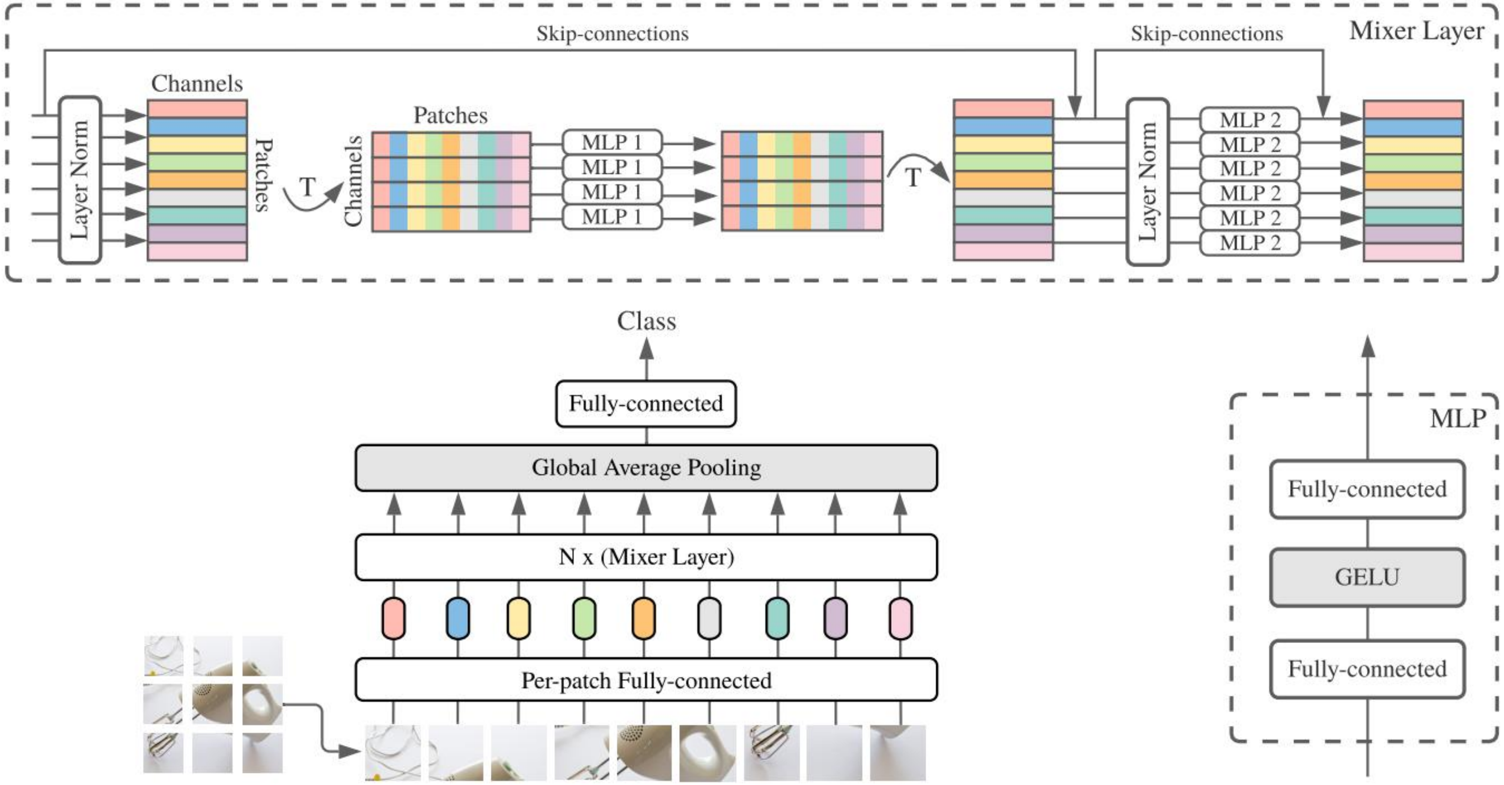
- ViT needs large dataset.





MLP-Mixer

- Remove attention in ViT but keep the structure.



Tolstikhin *et al.* MLP-Mixer: An all-MLP Architecture for Vision. *NIPS* 2021. cite 890.

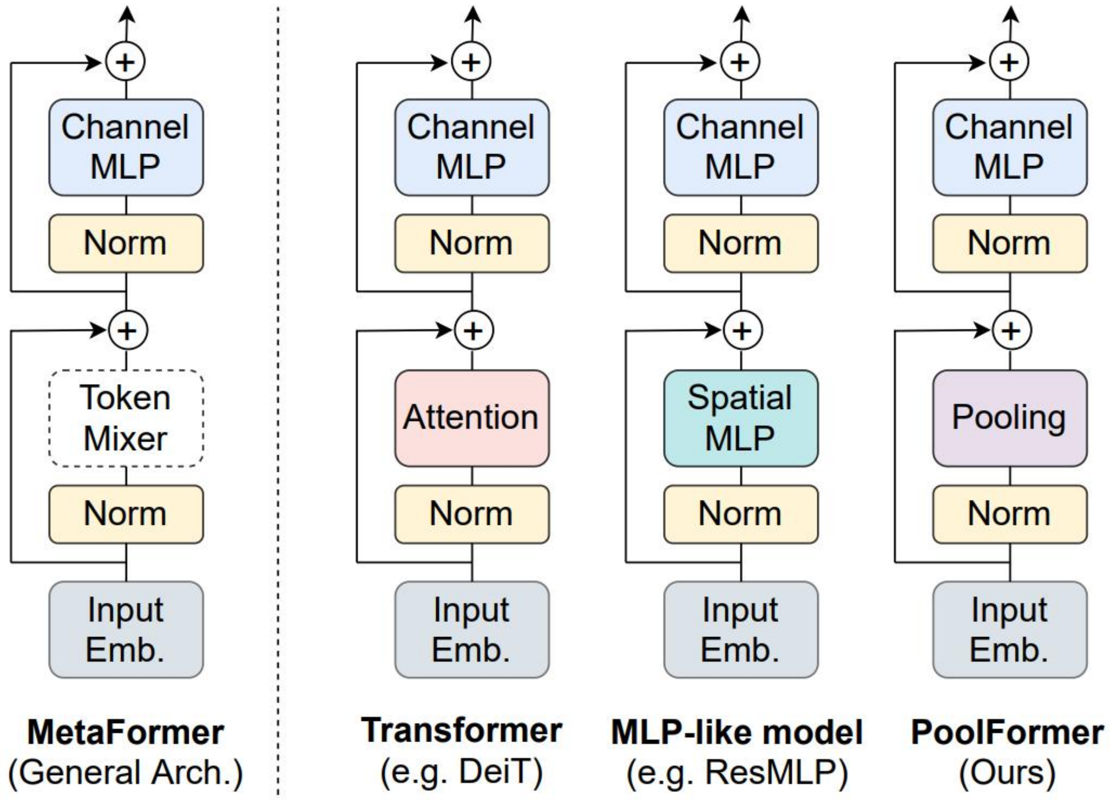
MLP-Mixer

- Sell point: the throughput, perform kind of well on large dataset.

	Image size	Pre-Train Epochs	ImNet top-1	ReaL top-1	Avg. 5 top-1	Throughput (img/sec/core)	TPUv3 core-days
Pre-trained on ImageNet (with extra regularization)							
Mixer-B/16	224	300	76.44	82.36	88.33	1384	0.01k$^{(\ddagger)}$
ViT-B/16 (☎)	224	300	79.67	84.97	90.79	861	0.02k$^{(\ddagger)}$
Mixer-L/16	224	300	71.76	77.08	87.25	419	0.04k$^{(\ddagger)}$
ViT-L/16 (☎)	224	300	76.11	80.93	89.66	280	0.05k$^{(\ddagger)}$
Pre-trained on ImageNet-21k (with extra regularization)							
Mixer-B/16	224	300	80.64	85.80	92.50	1384	0.15k$^{(\ddagger)}$
ViT-B/16 (☎)	224	300	84.59	88.93	94.16	861	0.18k$^{(\ddagger)}$
Mixer-L/16	224	300	82.89	87.54	93.63	419	0.41k$^{(\ddagger)}$
ViT-L/16 (☎)	224	300	84.46	88.35	94.49	280	0.55k$^{(\ddagger)}$
Mixer-L/16	448	300	83.91	87.75	93.86	105	0.41k$^{(\ddagger)}$
Pre-trained on JFT-300M							
Mixer-S/32	224	5	68.70	75.83	87.13	11489	0.01k
Mixer-B/32	224	7	75.53	81.94	90.99	4208	0.05k
Mixer-S/16	224	5	73.83	80.60	89.50	3994	0.03k
BiT-R50x1	224	7	73.69	81.92	—	2159	0.08k
Mixer-B/16	224	7	80.00	85.56	92.60	1384	0.08k
Mixer-L/32	224	7	80.67	85.62	93.24	1314	0.12k
BiT-R152x1	224	7	79.12	86.12	—	932	0.14k
BiT-R50x2	224	7	78.92	86.06	—	890	0.14k
BiT-R152x2	224	14	83.34	88.90	—	356	0.58k
Mixer-L/16	224	7	84.05	88.14	94.51	419	0.23k
Mixer-L/16	224	14	84.82	88.48	94.77	419	0.45k
ViT-L/16	224	14	85.63	89.16	95.21	280	0.65k
Mixer-H/14	224	14	86.32	89.14	95.49	194	1.01k
BiT-R200x3	224	14	84.73	89.58	—	141	1.78k
Mixer-L/16	448	14	86.78	89.72	95.13	105	0.45k
ViT-H/14	224	14	86.65	89.56	95.57	87	2.30k
ViT-L/16 [14]	512	14	87.76	90.54	95.63	32	0.65k

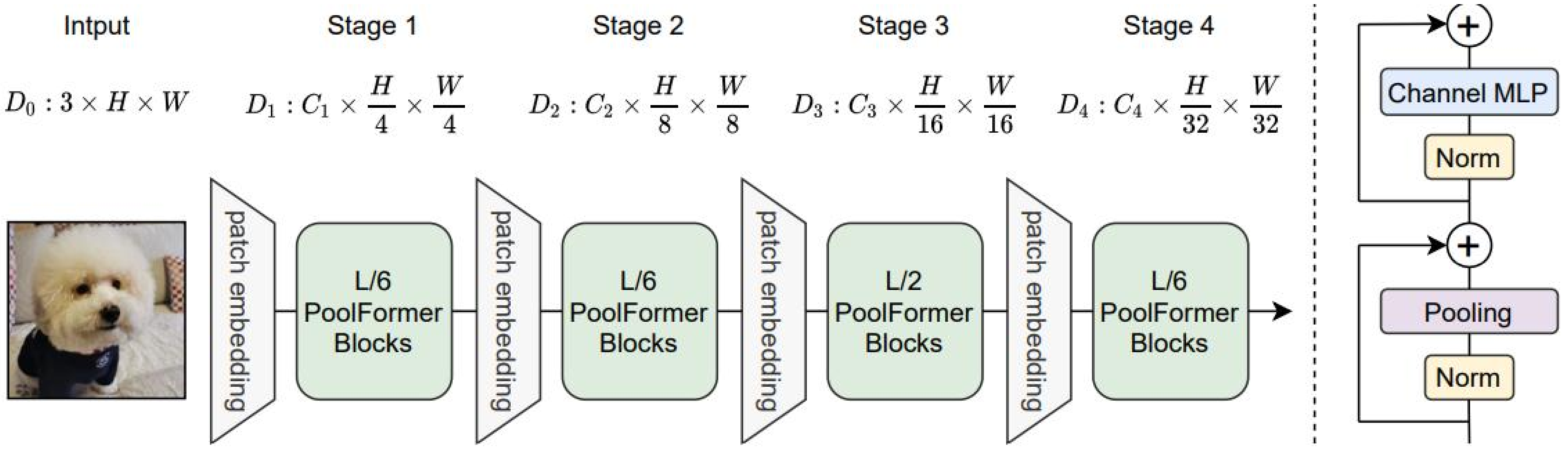
Metaformer

- Attention-based module can be replaced by spatial MLPs
- General architecture of the Transformers is essential



Metaformer

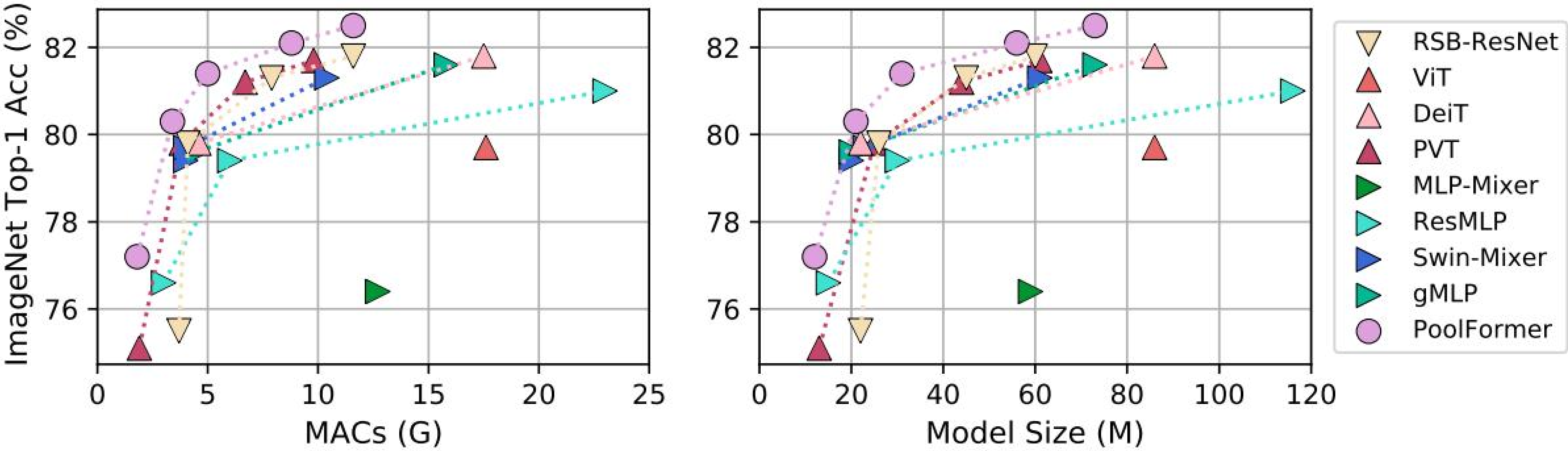
- Replace module with Average Pooling



(a) Overal framework with L PoolFormer blocks (b) PoolFormer block

Metaformer

- Reduce computation notably



COCs

- What is an image and how to extract latent features?
- CNN: organized pixels in rectangular shape, convolutional operation in local region
- ViTs: a sequence of patches, attention mechanism in a global range
- CoCs: a set of unorganized points, simplified clustering algorithm

COCs

- Context cluster
- Each pixel as a 5-dimensional data point with the information of color and position
- Convert image to a set of point clouds, utilize methodologies from point cloud analysis

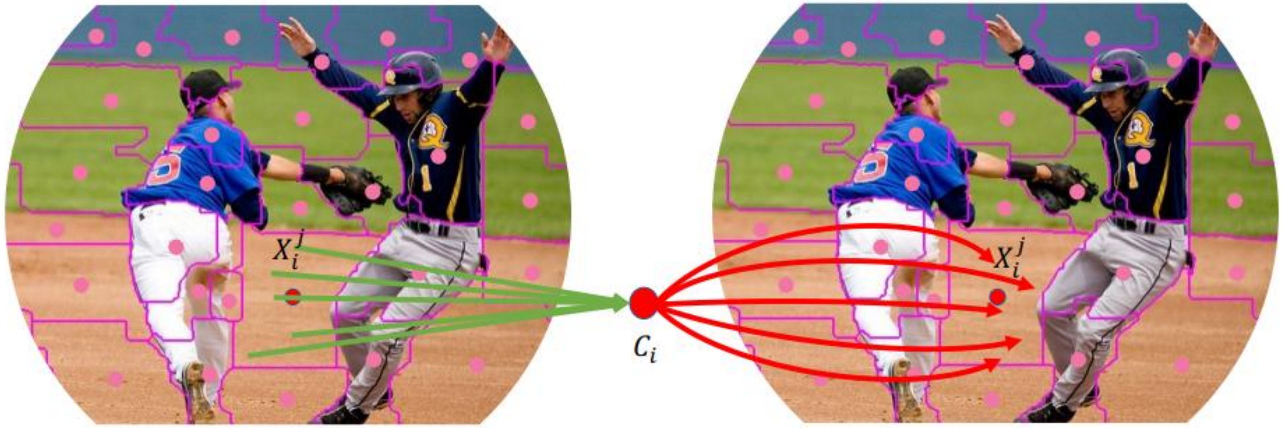
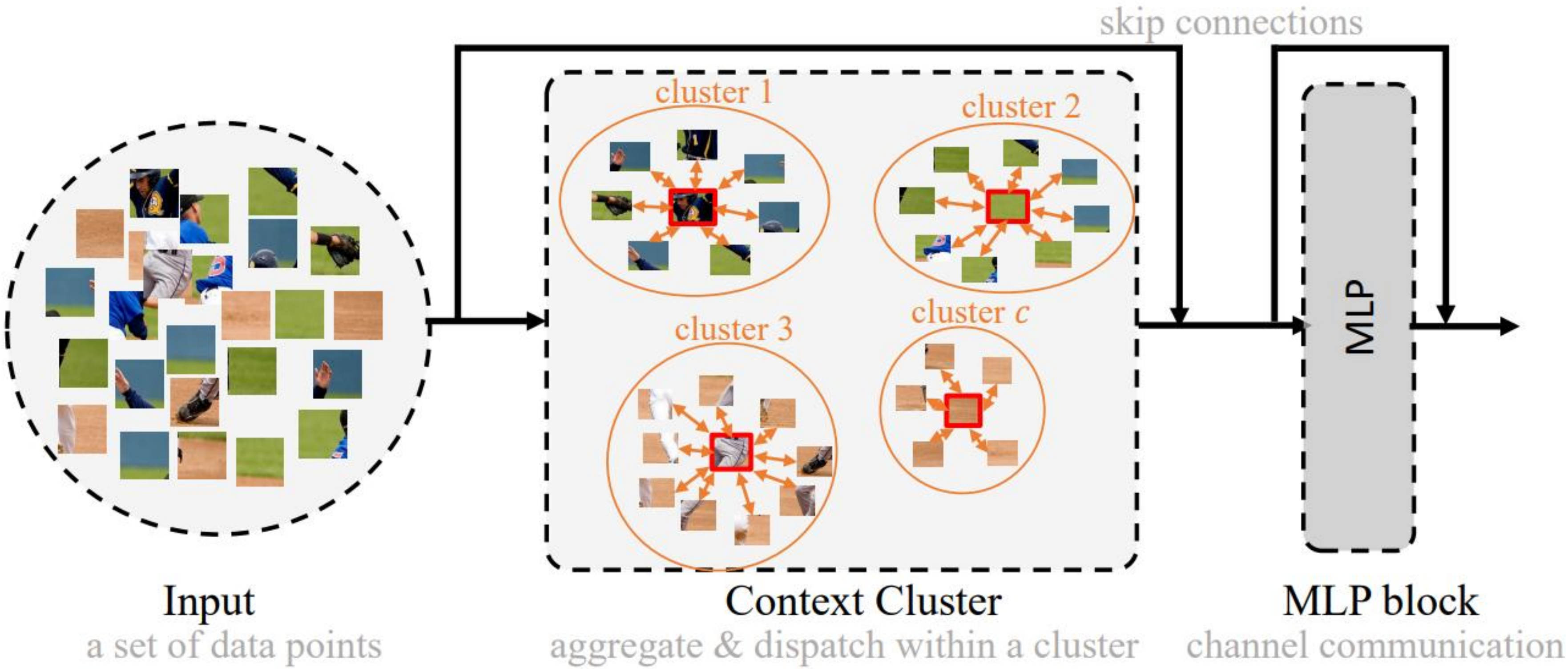


Figure 1: A context cluster in our network trained for image classification. We view *an image as a set of points* and sample c centers for points clustering. Point features are aggregated and then dispatched within a cluster. For cluster center C_i, we first aggregated all points $\{x_i^0, x_i^1, \cdots, x_i^n\}$ in ith cluster, then the aggregated result is distributed to all points in the clusters dynamically. See § 3 for details.

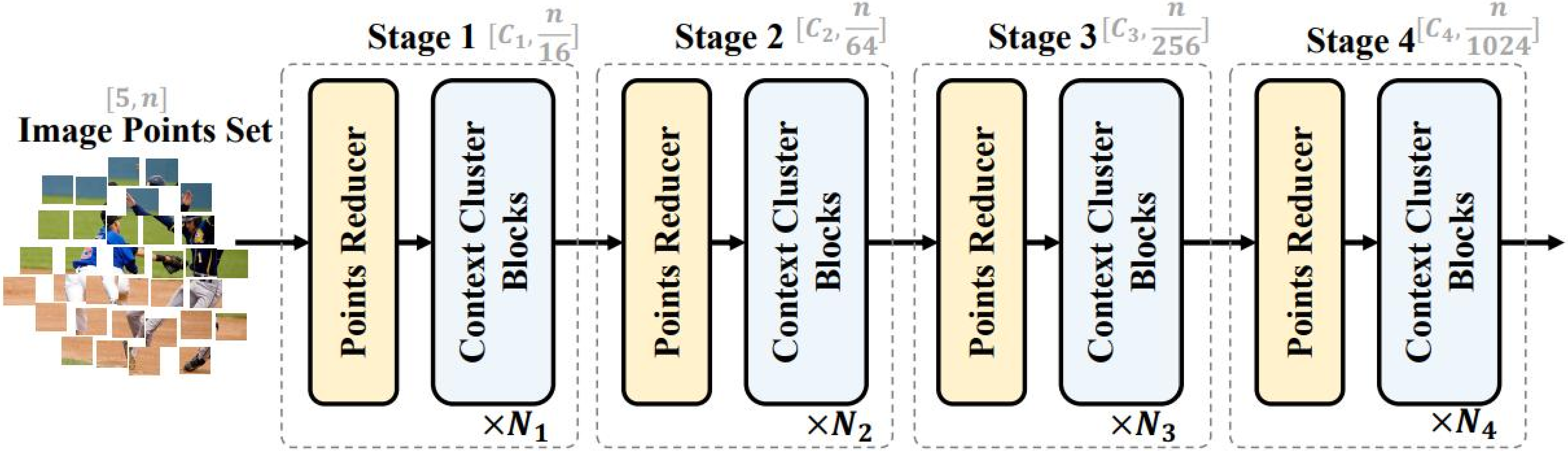
COCs

- Group a set of unorganized data points
- Communicate the points within clusters.
- Applied MLP block



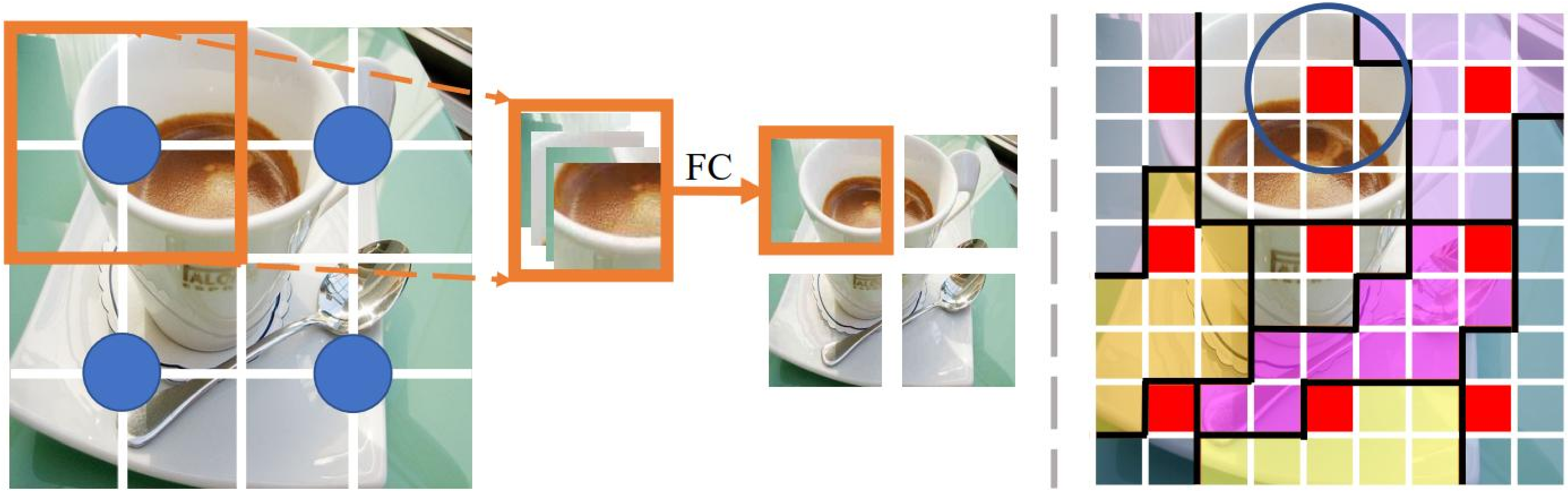
COCs

- Context Cluster architecture with four stages, extract deep feature
- Points reducer as down-sampler



COCs

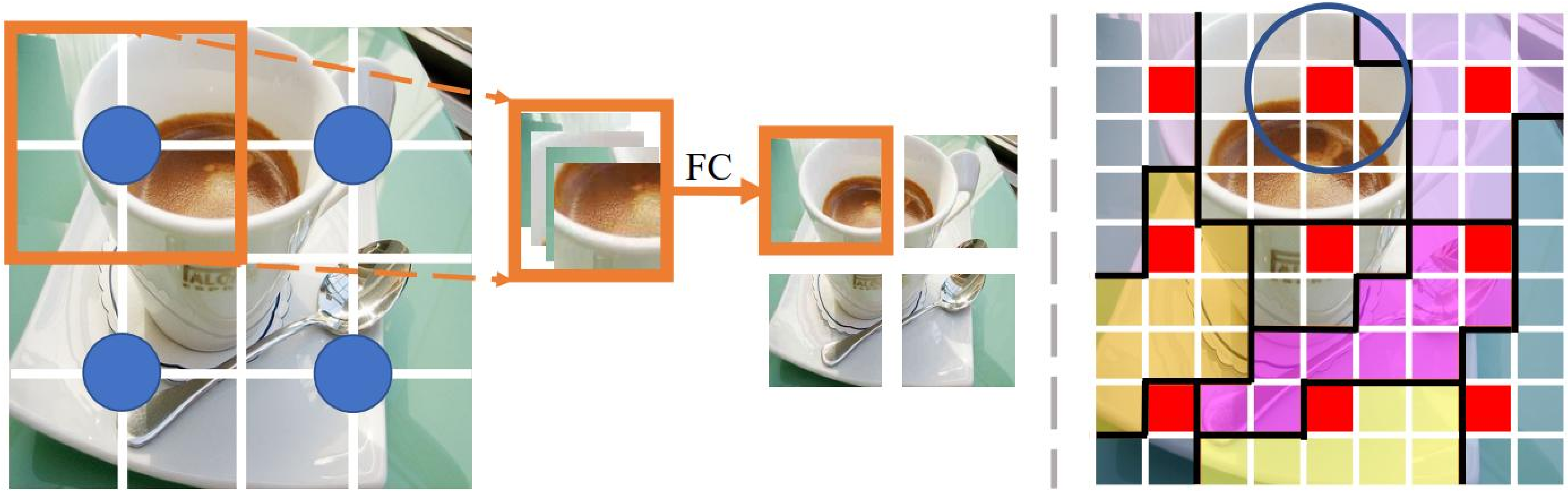
- Anchors in points reducer block and centers for context cluster block
- The center feature value is achieved by averaging its k neighbors as blue circle.



(a) Illustration of anchors for points reduction. (b) Demo of centers in CoC.

COCs

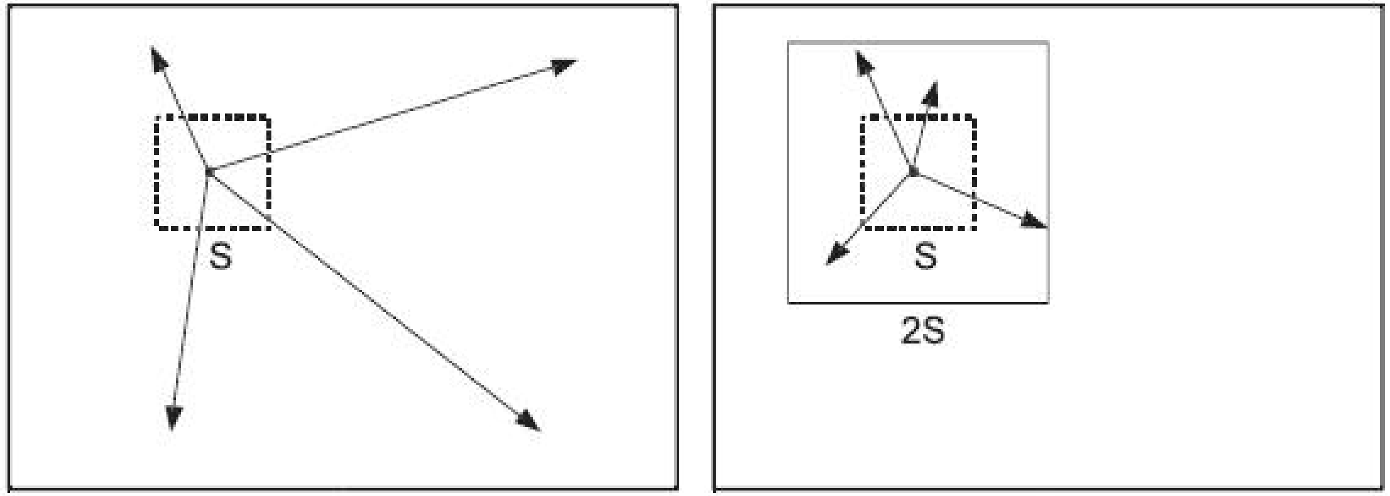
- Fixed center for cluster, but feature updated, aggregate in cluster and assign back.
- Calculation complexity consideration.



(a) Illustration of anchors for points reduction.　　(b) Demo of centers in CoC.

SLIC

- K-means clustering method with local searching region
- Linear complexity



(a) standard k-means searches the entire image

(b) SLIC searches a limited region

Experiment

- Imagenet-1k classification,
- Comparable, even better some case

	Method	Param.	GFLOPs	Top-1	Throughputs (images/s)
MLP	♣ ResMLP-12 (Touvron et al., 2021a)	15.0	3.0	76.6	511.4
	♣ ResMLP-24 (Touvron et al., 2021a)	30.0	6.0	79.4	509.7
	♣ ResMLP-36 (Touvron et al., 2021a)	45.0	8.9	79.7	452.9
	♣ MLP-Mixer-B/16 (Tolstikhin et al., 2021)	59.0	12.7	76.4	400.8
	♣ MLP-Mixer-L/16 (Tolstikhin et al., 2021)	207.0	44.8	71.8	125.2
	♣ gMLP-Ti (Liu et al., 2021a)	6.0	1.4	72.3	511.6
	♣ gMLP-S (Liu et al., 2021a)	20.0	4.5	79.6	509.4
Attention	♦ ViT-B/16 (Dosovitskiy et al., 2020)	86.0	55.5	77.9	292.0
	♦ ViT-L/16 (Dosovitskiy et al., 2020)	307	190.7	76.5	92.8
	♦ PVT-Tiny (Wang et al., 2021)	13.2	1.9	75.1	-
	♦ PVT-Small (Wang et al., 2021)	24.5	3.8	79.8	-
	♦ T2T-ViT-7 (Yuan et al., 2021a)	4.3	1.1	71.7	-
	♦ DeiT-Tiny/16 (Touvron et al., 2021b)	5.7	1.3	72.2	523.8
	♦ DeiT-Small/16 (Touvron et al., 2021b)	22.1	4.6	79.8	521.3
Convolution	♠ ResNet18 (He et al., 2016)	12	1.8	69.8	584.9
	♠ ResNet50 (He et al., 2016)	26	4.1	79.8	524.8
	♠ ConvMixer-512/16 (Trockman et al., 2022)	5.4	-	73.8	-
	♠ ConvMixer-1024/12 (Trockman et al., 2022)	14.6	-	77.8	-
	♠ ConvMixer-768/32 (Trockman et al., 2022)	21.1	-	80.16	142.9
Cluster	♥ Context-Cluster-Ti (ours)	5.3	1.0	71.8	518.4
	♥ Context-Cluster-Ti‡ (ours)	5.3	1.0	71.7	510.8
	♥ Context-Cluster-Small (ours)	14.0	2.6	77.5	513.0
	♥ Context-Cluster-Medium (ours)	27.9	5.5	81.0	325.2

Experiment

- Imagenet-1k classification

	Method	Param.	GFLOPs	Top-1	Throughputs (images/s)
MLP	♣ ResMLP-12 (Touvron et al., 2021a)	15.0	3.0	76.6	511.4
	♣ ResMLP-24 (Touvron et al., 2021a)	30.0	6.0	79.4	509.7
	♣ ResMLP-36 (Touvron et al., 2021a)	45.0	8.9	79.7	452.9
	♣ MLP-Mixer-B/16 (Tolstikhin et al., 2021)	59.0	12.7	76.4	400.8
	♣ MLP-Mixer-L/16 (Tolstikhin et al., 2021)	207.0	44.8	71.8	125.2
	♣ gMLP-Ti (Liu et al., 2021a)	6.0	1.4	72.3	511.6
	♣ gMLP-S (Liu et al., 2021a)	20.0	4.5	79.6	509.4
Attention	♦ ViT-B/16 (Dosovitskiy et al., 2020)	86.0	55.5	77.9	292.0
	♦ ViT-L/16 (Dosovitskiy et al., 2020)	307	190.7	76.5	92.8
	♦ PVT-Tiny (Wang et al., 2021)	13.2	1.9	75.1	-
	♦ PVT-Small (Wang et al., 2021)	24.5	3.8	79.8	-
	♦ T2T-ViT-7 (Yuan et al., 2021a)	4.3	1.1	71.7	-
	♦ DeiT-Tiny/16 (Touvron et al., 2021b)	5.7	1.3	72.2	523.8
	♦ DeiT-Small/16 (Touvron et al., 2021b)	22.1	4.6	79.8	521.3
Convolution	♠ ResNet18 (He et al., 2016)	12	1.8	69.8	584.9
	♠ ResNet50 (He et al., 2016)	26	4.1	79.8	524.8
	♠ ConvMixer-512/16 (Trockman et al., 2022)	5.4	-	73.8	-
	♠ ConvMixer-1024/12 (Trockman et al., 2022)	14.6	-	77.8	-
	♠ ConvMixer-768/32 (Trockman et al., 2022)	21.1	-	80.16	142.9
Cluster	♥ Context-Cluster-Ti (ours)	5.3	1.0	71.8	518.4
	♥ Context-Cluster-Ti‡ (ours)	5.3	1.0	71.7	510.8
	♥ Context-Cluster-Small (ours)	14.0	2.6	77.5	513.0
	♥ Context-Cluster-Medium (ours)	27.9	5.5	81.0	325.2

Experiment

- Imagenet-1k classification

	Method	Param.	GFLOPs	Top-1	Throughputs (images/s)
MLP	♣ ResMLP-12 (Touvron et al., 2021a)	15.0	3.0	76.6	511.4
	♣ ResMLP-24 (Touvron et al., 2021a)	30.0	6.0	79.4	509.7
	♣ ResMLP-36 (Touvron et al., 2021a)	45.0	8.9	79.7	452.9
	♣ MLP-Mixer-B/16 (Tolstikhin et al., 2021)	59.0	12.7	76.4	400.8
	♣ MLP-Mixer-L/16 (Tolstikhin et al., 2021)	207.0	44.8	71.8	125.2
	♣ gMLP-Ti (Liu et al., 2021a)	6.0	1.4	72.3	511.6
	♣ gMLP-S (Liu et al., 2021a)	20.0	4.5	79.6	509.4
Attention	♦ ViT-B/16 (Dosovitskiy et al., 2020)	86.0	55.5	77.9	292.0
	♦ ViT-L/16 (Dosovitskiy et al., 2020)	307	190.7	76.5	92.8
	♦ PVT-Tiny (Wang et al., 2021)	13.2	1.9	75.1	-
	♦ PVT-Small (Wang et al., 2021)	24.5	3.8	79.8	-
	♦ T2T-ViT-7 (Yuan et al., 2021a)	4.3	1.1	71.7	-
	♦ DeiT-Tiny/16 (Touvron et al., 2021b)	5.7	1.3	72.2	523.8
	♦ DeiT-Small/16 (Touvron et al., 2021b)	22.1	4.6	79.8	521.3
Convolution	♠ ResNet18 (He et al., 2016)	12	1.8	69.8	584.9
	♠ ResNet50 (He et al., 2016)	26	4.1	79.8	524.8
	♠ ConvMixer-512/16 (Trockman et al., 2022)	5.4	-	73.8	-
	♠ ConvMixer-1024/12 (Trockman et al., 2022)	14.6	-	77.8	-
	♠ ConvMixer-768/32 (Trockman et al., 2022)	21.1	-	80.16	142.9
Cluster	♥ Context-Cluster-Ti (ours)	5.3	1.0	71.8	518.4
	♥ Context-Cluster-Ti‡ (ours)	5.3	1.0	71.7	510.8
	♥ Context-Cluster-Small (ours)	14.0	2.6	77.5	513.0
	♥ Context-Cluster-Medium (ours)	27.9	5.5	81.0	325.2

Experiment

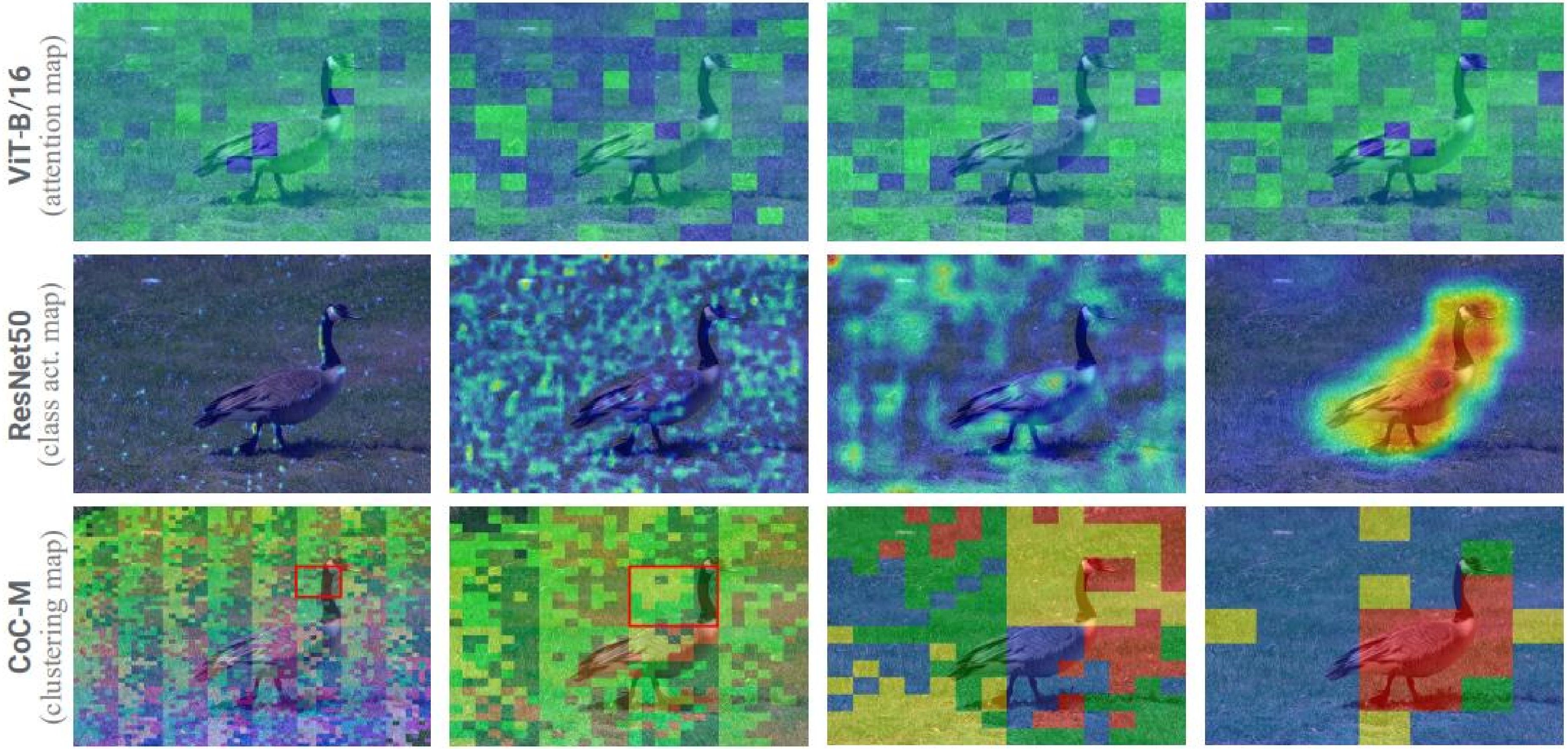
- Imagenet-1k classification

下一页有模糊的鸟！

	Method	Param.	GFLOPs	Top-1	Throughputs (images/s)
MLP	♣ ResMLP-12 (Touvron et al., 2021a)	15.0	3.0	76.6	511.4
	♣ ResMLP-24 (Touvron et al., 2021a)	30.0	6.0	79.4	509.7
	♣ ResMLP-36 (Touvron et al., 2021a)	45.0	8.9	79.7	452.9
	♣ MLP-Mixer-B/16 (Tolstikhin et al., 2021)	59.0	12.7	76.4	400.8
	♣ MLP-Mixer-L/16 (Tolstikhin et al., 2021)	207.0	44.8	71.8	125.2
	♣ gMLP-Ti (Liu et al., 2021a)	6.0	1.4	72.3	511.6
	♣ gMLP-S (Liu et al., 2021a)	20.0	4.5	79.6	509.4
Attention	♦ ViT-B/16 (Dosovitskiy et al., 2020)	86.0	55.5	77.9	292.0
	♦ ViT-L/16 (Dosovitskiy et al., 2020)	307	190.7	76.5	92.8
	♦ PVT-Tiny (Wang et al., 2021)	13.2	1.9	75.1	-
	♦ PVT-Small (Wang et al., 2021)	24.5	3.8	79.8	-
	♦ T2T-ViT-7 (Yuan et al., 2021a)	4.3	1.1	71.7	-
	♦ DeiT-Tiny/16 (Touvron et al., 2021b)	5.7	1.3	72.2	523.8
	♦ DeiT-Small/16 (Touvron et al., 2021b)	22.1	4.6	79.8	521.3
Convolution	♠ ResNet18 (He et al., 2016)	12	1.8	69.8	584.9
	♠ ResNet50 (He et al., 2016)	26	4.1	79.8	524.8
	♠ ConvMixer-512/16 (Trockman et al., 2022)	5.4	-	73.8	-
	♠ ConvMixer-1024/12 (Trockman et al., 2022)	14.6	-	77.8	-
	♠ ConvMixer-768/32 (Trockman et al., 2022)	21.1	-	80.16	142.9
Cluster	♥ Context-Cluster-Ti (ours)	5.3	1.0	71.8	518.4
	♥ Context-Cluster-Ti‡ (ours)	5.3	1.0	71.7	510.8
	♥ Context-Cluster-Small (ours)	14.0	2.6	77.5	513.0
	♥ Context-Cluster-Medium (ours)	27.9	5.5	81.0	325.2

Experiment

- Visualization of activation map, class activation map, and clustering map for ViT-B/16, ResNet50, and our CoC-M



Experiment

- 3D Point Cloud Classification on ScanObjectNN
- PointMLP as baseline

Method	mAcc(%)	OA(%)
♠ SpiderCNN (Xu et al., 2018)	69.8	73.7
♠ DGCNN (Wang et al., 2019)	73.6	78.1
♠ PointCNN (Li et al., 2018)	75.1	78.5
♠ GBNet (Qiu et al., 2021)	77.8	80.5
♦ PointBert (Yu et al., 2022d)	-	83.1
♦ Point-MAE (Pang et al., 2022)	-	85.2
♦ Point-TnT (Berg et al., 2022)	81.0	83.5
♣ PointNet (Qi et al., 2017a)	63.4	68.2
♣ PointNet++ (Qi et al., 2017b)	75.4	77.9
♣ BGA-PN++ (Uy et al., 2019)	77.5	80.2
♣ PointMLP (Ma et al., 2022)	83.9	85.4
♣ PointMLP-elite (Ma et al., 2022)	81.8	83.8
♥ PointMLP-CoC (ours)	**84.4** $_{\uparrow 0.5}$	**86.2** $_{\uparrow 0.8}$

Experiment

- Detection and segmentation

Table 5: Semantic segmentation performance of different backbones with Semantic FPN on the ADE20K validation set.

Backbone	Params	mIoU(%)
♠ ResNet18	15.5M	32.9
♦ PVT-Tiny	17.0M	35.7
♥ CoC-Small/4	17.7M	**36.6**
♥ CoC-Small/25	17.7M	**36.4**
♥ CoC-Small/49	17.7M	**36.3**

Table 4: COCO object detection and instance segmentation results using Mask-RCNN (1×).

Family	Backbone	Params	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Conv.	♠ ResNet-18	31.2M	34.0	54.0	36.7	31.2	51.0	32.7
Attention	♦ PVT-Tiny	32.9M	36.7	59.2	39.3	35.1	56.7	37.3
Cluster	♥ CoC-Small/4	33.6M	35.9	58.3	38.3	33.8	55.3	35.8
	♥ CoC-Small/25	33.6M	**37.5**	**60.1**	**40.0**	**35.4**	**57.1**	**37.9**
	♥ CoC-Small/49	33.6M	37.2	59.8	39.7	34.9	56.7	37.0

Experiment

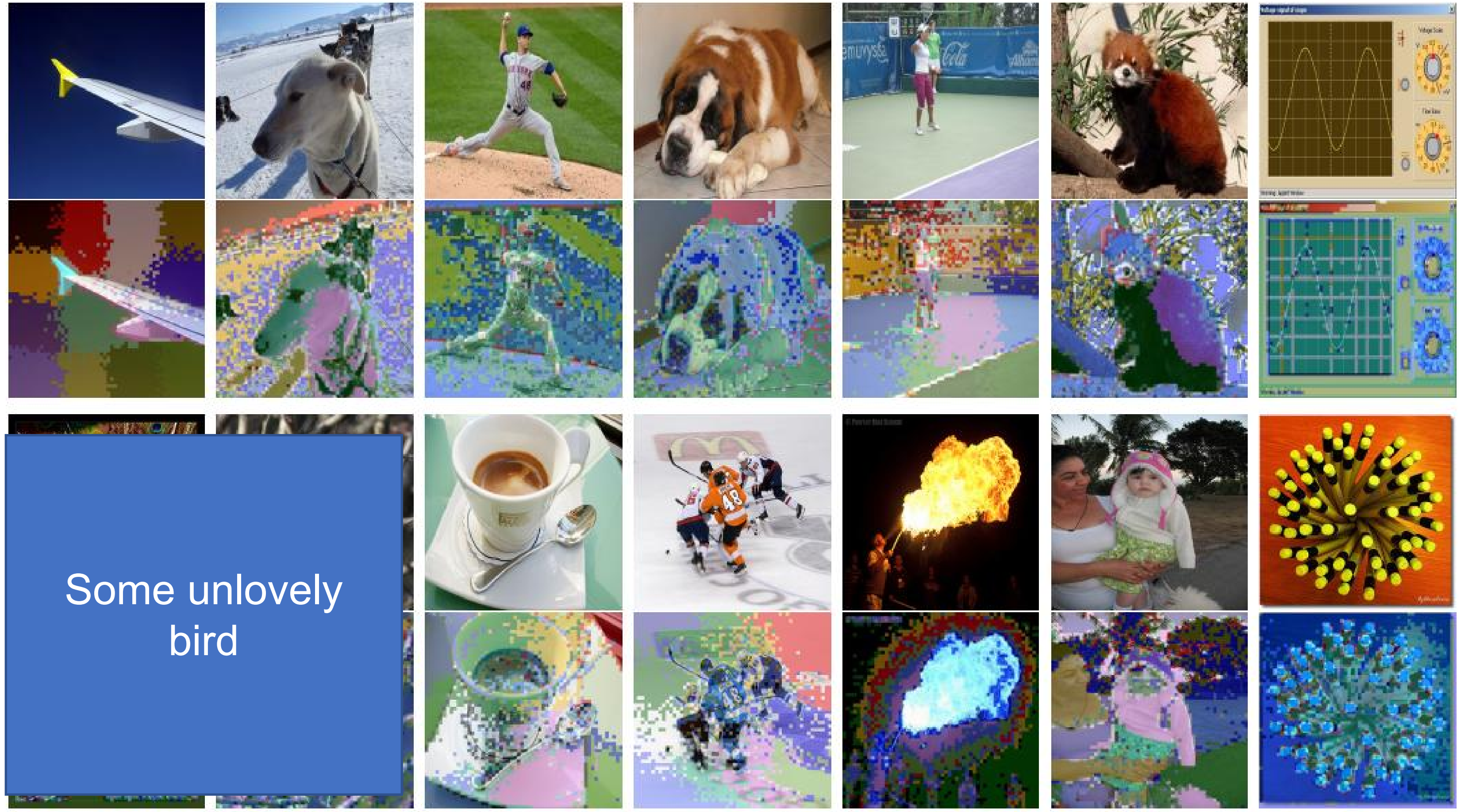
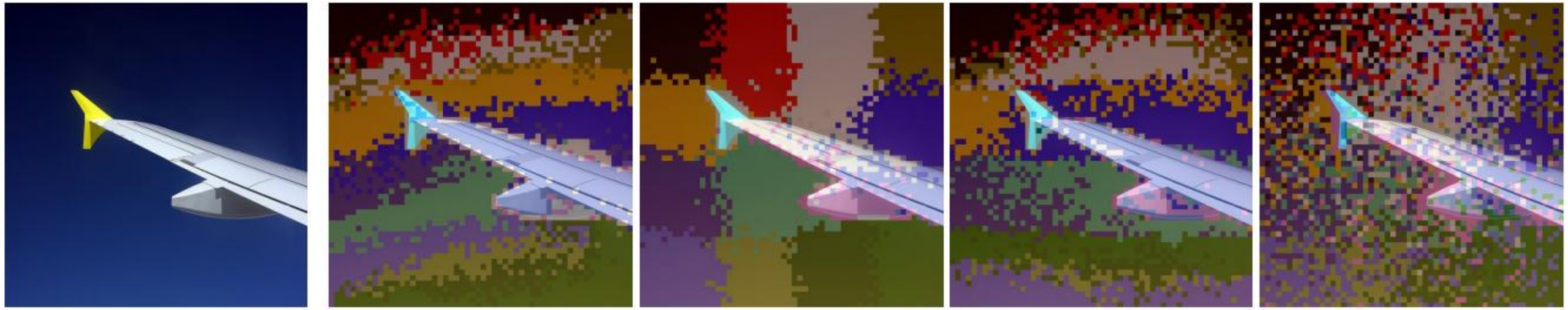


Figure 8: The clustering results of the last context cluster block in the first CoC-Tiny stage (without region partition). Without region partition, Our Context Cluster astonishingly displays "superpixel"-like clustering results, even in the early stage. we pick the most intriguing one out of the four heads.

Experiment

- Multi heads



Input Clustering results in 4 heads (16 clusters)

Figure 9: A sample of all groups' clustering results.

Application

(a) Discrete pixels (b) Masked image (c) Irregular image (d) RGB-D image

Figure 10: Four examples of image formats. Remember that there are no pixels in the white area.

Conclusion

- Propose a backbone with context cluster and metaformer structure
- Show promising performance
- Better interpretability for feature extraction and may support irregular input format

Thanks for your listening!