Visual Transformer with Differentiable Channel Selection: An Information Bottleneck Inspired Approach

Yancheng Wang[1], Ping Li[2], Yingzhen Yang[1].
[1]School of Computing and Augmented Intelligence, Arizona State University.
[2]VecML Inc.

Introduction

- Visual transformers have demonstrated remarkable performance compared to state-of-the-art CNNs across a wide range of computer vision tasks. However, the achievements of visual transformers are accompanied by heavy computational costs, making their deployment impractical under resource-limited scenarios.

- Our goal of this work is to prune the attention output channels of visual transformers while maintaining and even improving the prediction accuracy of the original transformers.

- Such goal is achieved by encouraging the compressed transformers by channel pruning to better adhere to the Information Bottleneck (IB) principle. This is inspired by the fact that extensive empirical and theoretical works have evidenced that models respecting the IB principle enjoy compelling generalization.

Contributions

- We present a novel and compact transformer block termed Transformer with Differentiable Channel Selection, or DCS-Transformer. Using our proposed channel selection in both the computation for attention weights and the features of the MLP, DCS-Transformer blocks automatically select channels in queries and keys to compute more informative attention weights inspired by the Information Bottleneck (IB) principle.

- DCS-Transformer blocks can be used to replace all the transformer blocks in many popular visual transformers, rendering compact visual transformers with comparable or even better performance. The effectiveness of DCS-Transformer is evidenced by replacing all the transformer blocks with DCS-Transformer blocks into popular visual transformers which are already compact, including MobileViT, EfficientViT, ViT-S/16, and Swin-T, for image classification, object detection and instance segmentation tasks.

Contributions (Cont'd)

- In order to improve the generalization capability of the compressed transformers after channel pruning, we propose to reduce the IB loss of our DCS-Transformer model. A model with a smaller IB loss indicates that the model better adhere to the IB principle. To this end, we derive the first separable variational upper bound for the IB loss. Such separable upper bound for IB can be directly incorporated into existing training loss of deep neural networks even beyond transformers, and optimized in an end-to-end manner by standard SGD. Experimental results demonstrate that the IB loss of the visual transformer can be reduced by optimizing the composite loss formed by our variational upper bound for the IB loss and the regular cross-entropy loss, and the transformer network trained with such variational upper bound exhibits stronger generalization.

Formulation

- **The definition of IB loss.** Let X be the input training features, $\tilde{X}$ be the learned features by the network, and Y be the ground truth training labels for a classification task. Then the IB loss is $I(\tilde{X}) - I(\tilde{X}, Y)$, where $I(\cdot, \cdot)$ denotes mutual information.

- There are two types of channel selection in our DCS-Transformer, which are (1) channel selection for attention weights that renders more informative attention weights or affinity between tokens; (2) channel selection for attention outputs which prunes the channels of the MLP features so as to reduce the FLOPs of the transformer block.

Formulation (Cont'd): Two Types of Channel Selection

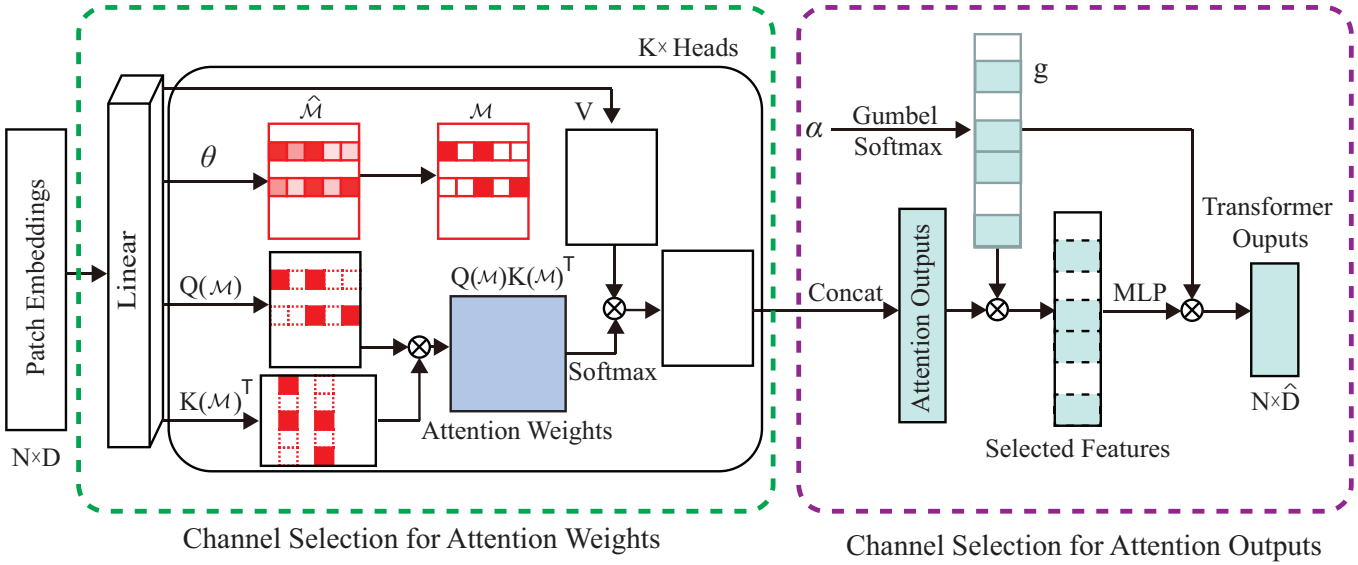

Figure 1: DCS-Transformer

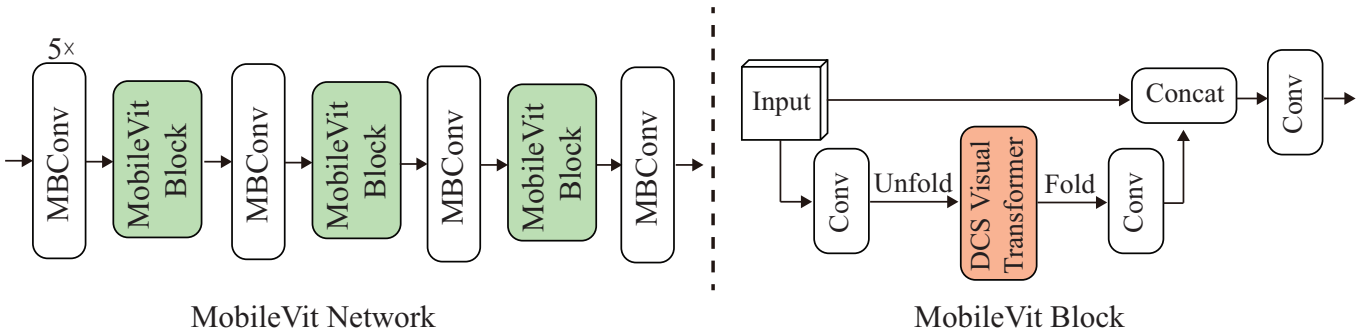

Figure 2: Architecture of DCS-MobileViT

Formulation (Cont'd): Separable Variational Upper Bound for the IB Loss

- The separable variational upper bound for the IB loss ($\mathsf{IB}(\mathcal{W})$) of a DCS-Transformer network with weights $\mathcal{W}$ is presented in the following theorem.

Theorem

$$\mathsf{IB}(\mathcal{W}) \leq \mathsf{IBB}(\mathcal{W}),$$

where

$$\mathsf{IBB}(\mathcal{W}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{A}\sum_{b=1}^{B}\phi(\tilde{X}_i, a)\phi(X_i, b)\log\phi(X_i, b)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{A}\sum_{y=1}^{C}\phi(\tilde{X}_i, a)\mathbb{1}_{\{y_i=y\}}\log Q(\tilde{X}\in a|Y=y).$$

- $\mathsf{IBB}(\mathcal{W})$ is the separable variational upper bound for the IB loss which can be incorporated into the existing training loss and optimized by the standard SGD.

Experiments

- We first evaluate the performance of DCS-Transformers on the ImageNet-1k dataset for image classification, and show that both models render better performance than state-of-the-art networks with more compact models.

- We then show that using DCS-MobileViT and DCS-EfficientViT as the feature extraction backbones achieve better mAP with lower FLOPs than the competing baselines for semantic segmentation and object detection. Please refer to more details in our paper.

Experiments (Cont'd): Image Classification on ImageNet-1k

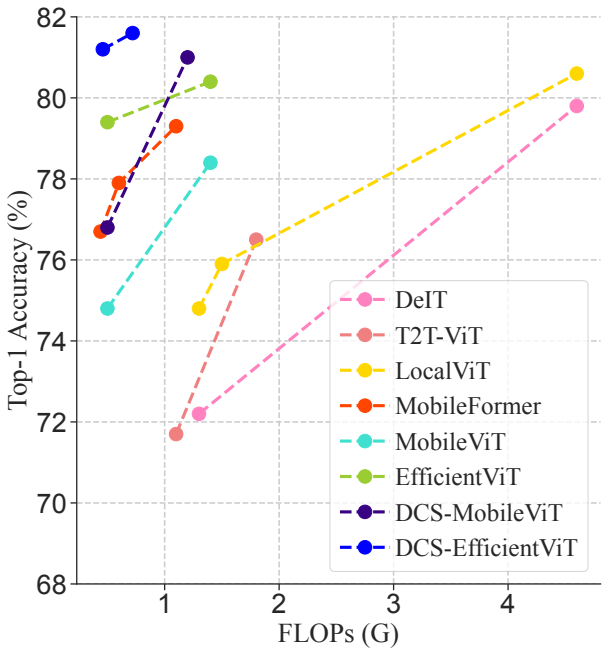

Figure 3: Top-1 accuracy vs FLOPs (G) on ImageNet-1k validation set.

Model	# Params	FLOPs	Top-1
T2T	4.3 M	1.1 G	71.7
DeIT	5.7 M	1.2 G	72.2
CrossViT	6.9 M	1.6 G	73.4
MobileViT-XS	2.3 M	0.7 G	74.8
DCS-MobileViT-XS (Ours)	2.0 M	0.5 G	**76.8**
DeIT	10 M	2.2 G	76.6
T2T	6.9 M	1.8 G	76.5
PiT	10.6 M	1.4 G	78.1
Mobile-Former	9.4 M	0.2 G	76.7
EViT	12.4 M	0.5 G	77.1
TinyViT	5.4 M	1.3 G	79.1
DeIT	22 M	4.6 G	79.8
ToMe	22 M	2.7 G	79.4
EfficientFormer	12.3 M	1.3 G	79.2
MobileViT-S	5.6 M	1.4 G	78.4
VTC-LFC	5.0 M	1.3 G	78.0
SPViT	4.9 M	1.2 G	77.8
ToMe	5.6 M	1.2 G	77.3
DCS-MobileViT-S (Ours)	4.8 M	1.2 G	**81.0**
EfficientViT-B1 [r224]	9.1 M	0.52 G	79.4
EfficientViT-B1 [r288]	9.1 M	0.86 G	80.4
EViT	8.8 M	0.29 G	74.3
VTC-LFC	8.7 M	0.76 G	79.3
SPViT	8.3 M	0.71 G	79.0
ToMe	9.1 M	0.47 G	78.8
DCS-EfficientViT-B1 [r224] (Ours)	8.2 M	0.46 G	**80.8**
DCS-EfficientViT-B1 [r288] (Ours)	8.2 M	0.72 G	**81.6**
ViT-S/16	22.1 M	4.3 G	81.2
DCS-ViT-S/16 (Ours)	20.2 M	3.9 G	**82.0**
Swin-T	29.0 M	4.5 G	81.3
DCS-Swin-T (Ours)	26.1 M	4.0 G	**82.0**

Table 1: Comparisons with baseline methods on ImageNet-1k validation set.

Experiments (Cont'd): Instance Segmentation and Object Detection

Methods	mAP^{box}	AP^{b}_{50}	AP^{b}_{75}	mAP^{m}	AP^{m}_{50}	AP^{m}_{75}
EViT	32.8	54.4	34.5	31.0	51.2	32.2
EfficientViT-B1	33.5	55.4	34.8	31.9	52.3	32.7
DCS-EfficientViT-B1	**34.8**	**56.3**	**35.3**	**33.2**	**53.1**	**33.3**

Table 2: Instance Segmentation Results on COCO.

Feature backbone	# Params.	FLOPs	mAP
MobileNetv3	4.9 M	1.4 G	22.0
MobileNetv2	4.3 M	1.6 G	22.1
MobileNetv1	5.1 M	2.6 G	22.2
MixNet	4.5 M	2.2 G	22.3
MNASNet	4.9 M	1.7 G	23.0
YoloV5-N (640×640)	1.9 M	4.5 G	28.0
Vidt	7.0 M	6.7 G	28.7
MobileViT-XS	2.7 M	1.7 G	24.8
DCS-MobileViT-XS(Ours)	2.4 M	1.5 G	**25.8**
MobileViT-S	5.7 M	2.4 G	27.7
DCS-MobileViT-S(Ours)	4.7 M	2.1 G	**28.7**
EfficientViT	9.9 M	1.5 G	28.4
DCS-EfficientViT(Ours)	9.0 M	1.4 G	**29.0**

Table 3: Object detection performance with SSDLite.

Key Takeaways

- While channel pruning is an effective method for compressing transformers, it usually results in models with worse prediction accuracy.
- We can maintain a compelling prediction accuracy of a compressed transformers by reducing its IB loss, a information-theoretical measure which benefits generalization capability.
- The IB loss can be reduced by optimizing a novel separable variational upper bound for the IB loss (the IBB), and such IBB can be used to enhance the performance of deep learning models beyond transformers.

Thank you!