Bi-Directional Self-Attention for Vision Transformers

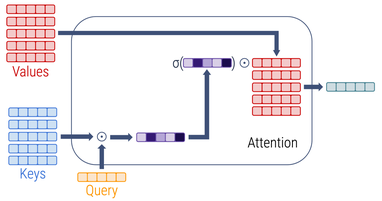
George Stoica, Taylor Hearn, Bhavika Devnani, Judy Hoffman
{gstoica3, thearn6, bdevnani3, judy}@gatech.edu

VTTA @ Neurips

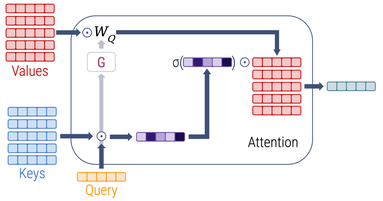
1 Background

Process a set of **key**-**value** vector pairs (context) according to a **query** vector (source)

(Self) Attention



Generalized (Self) Attention



Combine **values** (transformed by the **query**) using the compatibility of **keys** to the **query**

2 Problem

*(Generalized) Attention limits the information flow between **query** and **key**-**value** pairs*

- **Attention Can:** Transform context differently based on the source (e.g., emphasize foreground while filtering out background)
- **Attention Cannot:** Transform source information of interest differently based on context (e.g., emphasize differences between foreground elements)

*The **query** cannot be transformed according to the **key**-**value** set*

3 Contributions

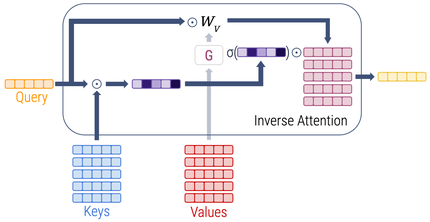
- **Inverse (Self) Attention (ISA):** Inverts attention to transform a **query** according to the **key**-**value** set
- **Bi-Directional (Self) Attention (BiSA):** Enables the **query** and **key**-**value** set to simultaneously transform one another by coupling Attention and Inverse Attention — *replaces attention mechanism*

Code: *https://github.com/gstoica27/BiSA*

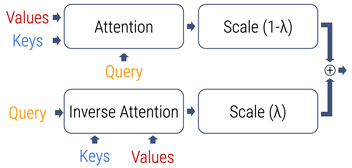
4 Approach

Process a **query** vector (source) according to a set of **key**-**value** vector pairs (context)

Inverse (Self) Attention



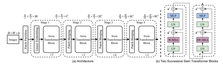
Bi-Directional (Self) Attention



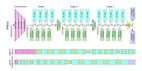
5 Experimental Setup

Models

Swin-Transformer



LeViT



Tasks

CIFAR100 ImageNet1K ADE20K

Image classification & Semantic Segmentation

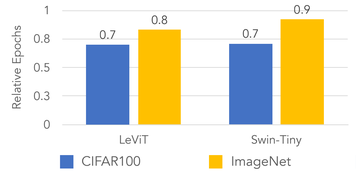
6 Findings

Dataset	Model	Variant	Norm	λ	#Layers Replaced	#Params	#Flops	Metrics	
								Top-1	Top-5
CIFAR100	Swin-Tiny	Baseline	-	-	0	27.6M	4.5G	80.09	94.62
		BiSA	-	0.5	2	27.7M	5.3G	**81.68**	**95.53**
		BiSA	x	0.5	2	27.7M	5.3G	80.75	95.14
		BiSA	x	Learned	2	27.7M	5.3G	80.34	94.91
	LeViT-128s	Baseline	-	-	0	7.8M	306M	73.58	92.98
		BiSA	-	0.5	2	8.0M	395M	**77.63**	**94.79**
		BiSA	x	0.5	2	8.0M	395M	75.29	93.76
		BiSA	x	Learned	2	8.0M	395M	75.05	93.42
ImageNet1K	Swin-Tiny	Baseline	-	-	0	28.3M	4.5G	81.19	95.52
		BiSA	-	0.5	2	28.4M	5.3G	**81.45**	**95.66**
	LeViT-128s	Baseline	-	-	0	7.8M	306M	76.45	92.97
		BiSA	-	0.5	2	8.0M	395M	**77.52**	**93.50**

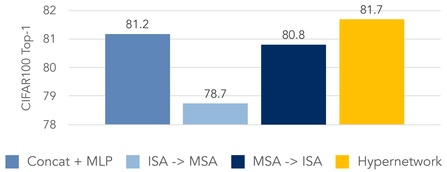
ADE20K: Swin-Tiny + Upernet

Type	Variant	Norm	λ	#Layers Replaced	#params	Metrics		
						mIoU	mAcc	aAcc
All	Baseline	-	-	0	60M	44.51	55.61	81.09
	BiSA	-	0.5	2	60.1M	**45.11 (+1.3)**	**57.15 (+2.8)**	**81.21 (+0.1)**
Boundary	Baseline	-	-	0	60M	24.37	35.75	53.86
	BiSA	-	0.5	2	60.1M	**24.80 (+1.7)**	**37.11 (+3.8)**	**53.87 (+0.0)**
Boundary vs. All						+0.4	+1.0	-0.1

Relative Epoch Time



BiSA Mechanism Ablations



BiSA emphasizes image boundary regions