Paper ID # 996

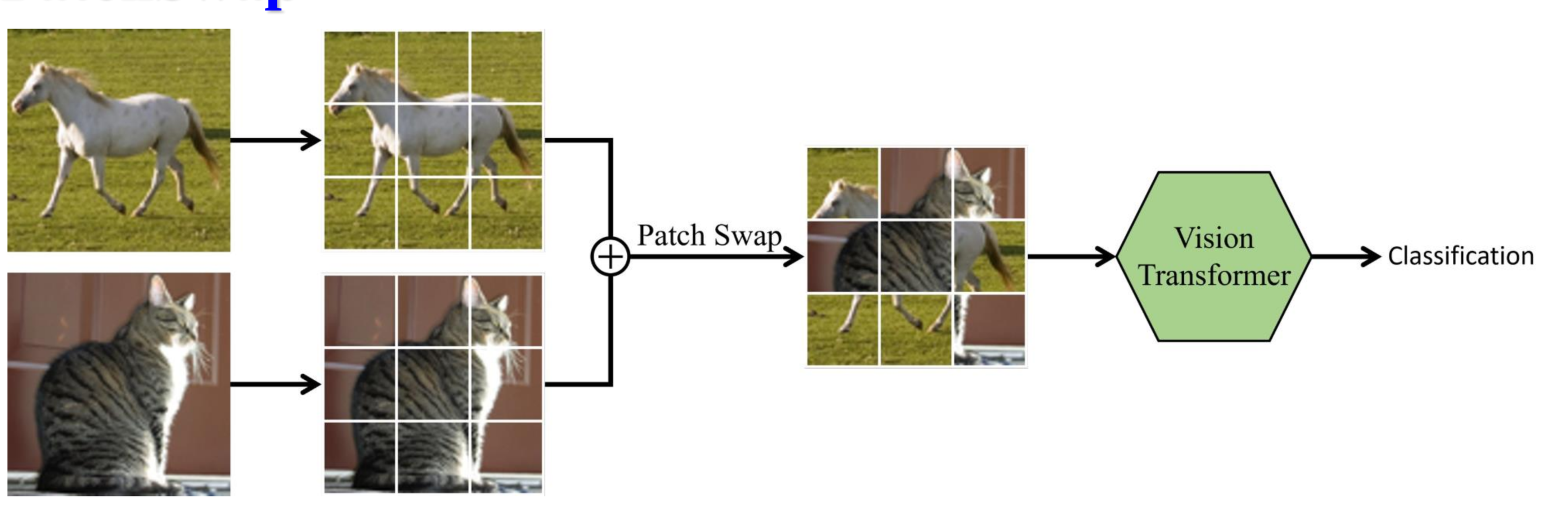
PatchSwap: A Regularization Technique for Vision Transformers

Sachin Chhabra[1], Hemanth Venkateswara[2], Baoxin Li[1]
[1]Arizona State University, [2]Georgia State University

Problem

- Since their introduction Vision Transformers have outperformed ConvNets on various image processing task.
- However, this superior performance is observed only when abundant labeled data is available.
- In the case of limited data, ConvNets beats Vision Transformers.
- For small datasets, Vision Transformers tend to observe a lot more overfitting than ConvNets.
- Existing data augmentation regularization techniques like Mixup and CutMix work with Vision Transformers but were originally designed for ConvNets.

PatchSwap



- Vision Transformers break the image into fixed size patches before passing it through the network.
- PatchSwap swaps the patches using a binary mask between two random images to create a PatchSwap image.
- Network is trained on PatchSwap images instead of the original image.
- The targets for the PatchSwap images are calculated using the ratio of the combined classes.
- It fully utilizes the global receptive field of the vision transformers as the patches can be located anywhere in the image.

Results

Dataset	CIFAR-10			FashionMNIST			SVHN		
Patch Size	4	8	16	4	8	16	4	8	16
Cross Entropy	83.3	78.3	69.8	92.1	92.8	91.2	96.4	94.7	92.7
Label smoothing [24]	83.0	79.0	69.6	92.0	92.9	91.5	96.5	94.8	92.8
Cutout [8]	84.0	79.2	70.1	94.2	93.5	91.4	96.8	96.2	94.5
Mixup [32]	87.4	82.3	74.3	93.0	93.4	92.2	97.0	95.7	94.2
Cutmix [31]	88.0	82.7	73.8	94.0	93.8	92.5	96.9	96.2	**94.8**
PatchSwap	**88.3**	**84.7**	**74.9**	**94.4**	**93.9**	**92.6**	**97.2**	**96.8**	**94.8**

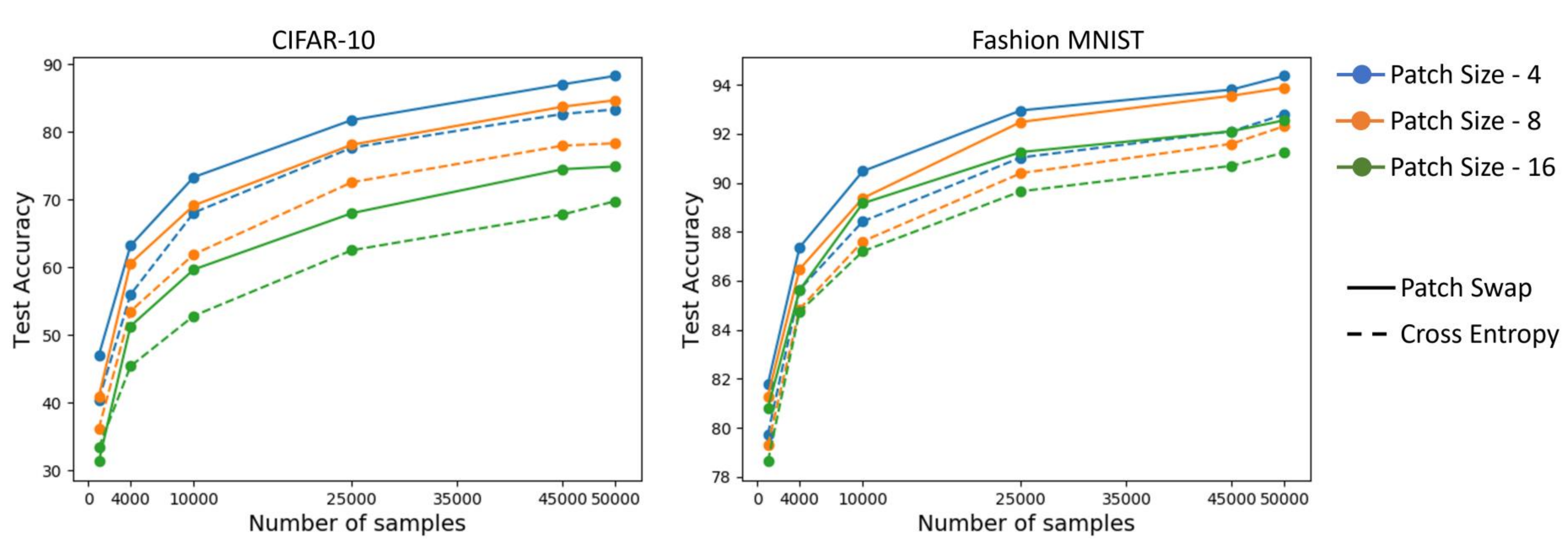
Patch Size	4		8		16	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Cross Entropy	57.9	81.5	50.6	76.2	39.3	65.0
Label smoothing [24]	58.3	77.0	51.5	71.7	39.8	62.3
Cutout [8]	57.0	81.1	50.2	76.1	39.1	64.7
Mixup [32]	63.5	85.0	56.8	80.0	45.3	70.6
Cutmix [31]	63.7	85.2	57.0	80.4	44.2	69.5
PatchSwap	**64.9**	**86.4**	**58.5**	**82.5**	**45.7**	**71.6**

CIFAR-100

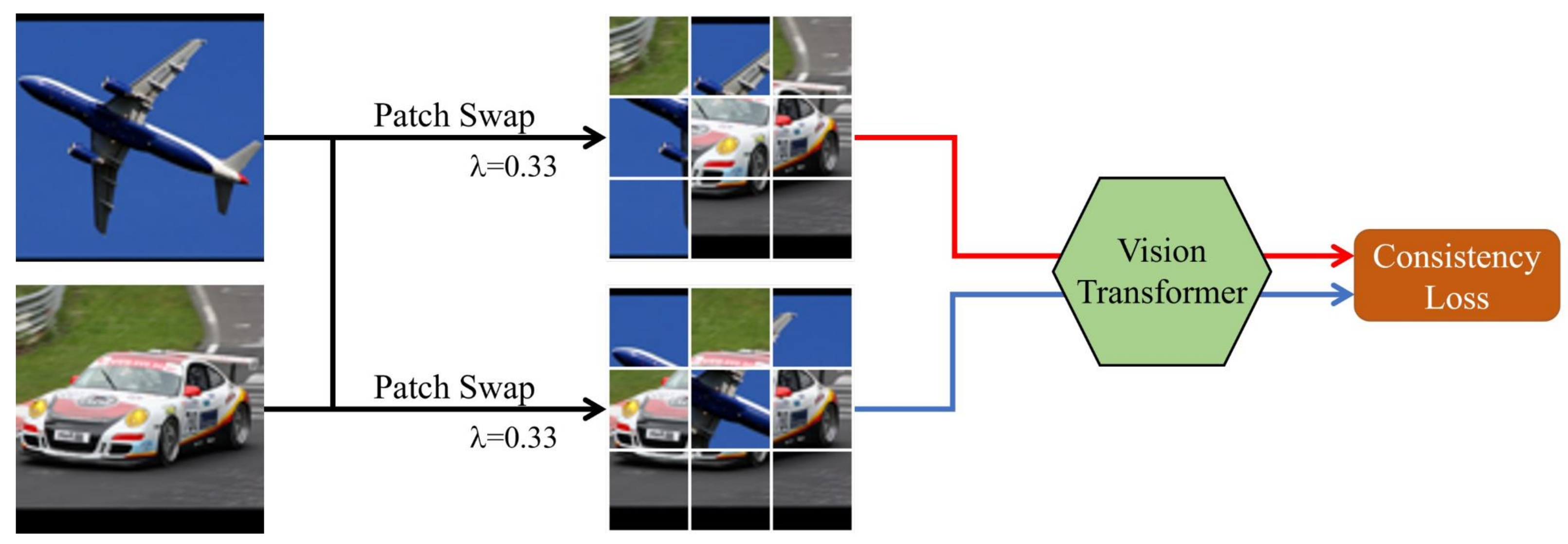
Augmentation	Standard				RandAugment [7]			
Patch Size	8		16		8		16	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Cross Entropy	41.9	65.2	34.4	57.7	46.2	70.4	39.1	63.4
Label smoothing [24]	42.8	63.0	34.6	56.5	47.0	69.7	39.3	62.7
Cutout [8]	42.8	66.6	33.8	58.1	47.5	71.5	40.2	65.1
Mixup [32]	46.6	69.0	38.5	62.4	49.9	73.5	43.2	67.5
Cutmix [31]	48.4	71.6	39.5	63.5	48.4	74.9	44.0	68.0
PatchSwap	**49.9**	**73.4**	**41.8**	**66.3**	**52.8**	**77.0**	**45.6**	**70.8**

Tiny-ImageNet

Results with Limited Labels



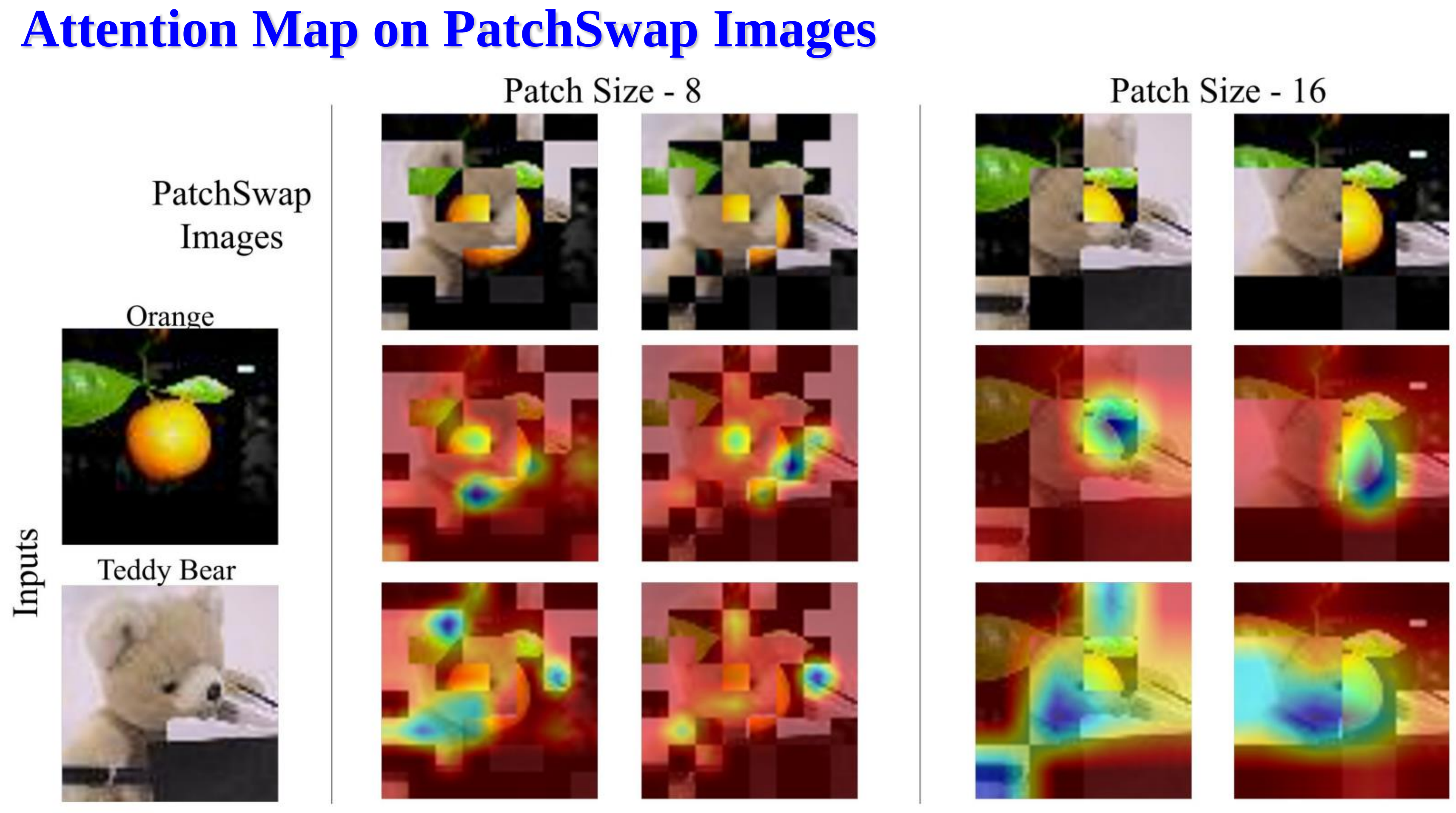
Unsupervised PatchSwap



- An extension of PatchSwap when limited labels are available.
- The number of patches swapped decides the swapping ratio.
- The same swapping ratio can be achieved in multiple ways for PatchSwap images.
- In the above example, both the PatchSwap images have $\lambda = 0.33$ yet they are different.
- A consistency loss is used between the outputs of the two PatchSwap images to train the network.
- Below are the results on using 4000 labeled samples and the rest as unlabeled on CIFAR-10 and SVHN.

Dataset	CIFAR-10			SVHN		
Patch Size	4	8	16	4	8	16
Cross Entropy	56.0	53.5	45.4	87.9	86.7	76.0
Pseudo Label [19]	58.1	54.0	46.3	91.2	88.9	78.0
MeanTeacher [26]	62.6	56.5	48.2	96.2	95.1	90.1
PatchSwap (Labeled only)	63.2	60.6	51.3	89.7	87.2	81.0
PatchSwap (Full)	**67.6**	**62.9**	**54.2**	**96.4**	**96.7**	**90.9**

Attention Map on PatchSwap Images



- We generate class-wise attention maps for various PatchSwap images.
- The network can identify the corresponding and relevant patches for each class in the PatchSwap Images.

Summary

- We presented the PatchSwap, a regularization technique for Vision Transformers.
- PatchSwap swaps patches between two images to create a PatchSwap image.
- Through extensive experiments on multiple datasets and settings, we showcased that PatchSwap results in superior performance.
- PatchSwap can also be extended to an unsupervised setting and results in a superior performance than vanilla consistency training.

Acknowledgment

The work was supported in part by a grant from ONR. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ONR.