PatchSwap: A Regularization Technique for Vision Transformers - Supplementary Material

Sachin Chhabra[1]
schhabr6@asu.edu
Hemanth Venkateswara[2]
hvenkateswara@gsu.edu
Baoxin Li[1]
baoxin.li@asu.edu

[1] Arizona State University,
699 S Mill Ave.,
Tempe, AZ, USA - 85281

[2] Georgia State University,
25 Park Pl NE,
Atlanta, GA, USA - 30303

1 Training Progress

We showcase the training progression of PatchSwap vs Cross-Entropy for the CIFAR-100 and SVHN datasets. The results are displayed in Figure 1. Cross-entropy achieves better test accuracy at the start of the training but PatchSwap results in the best final accuracy for all the scenarios. The training accuracy was around 100 % for all the cases at the end of the training. However, PatchSwap results in much less overfitting.

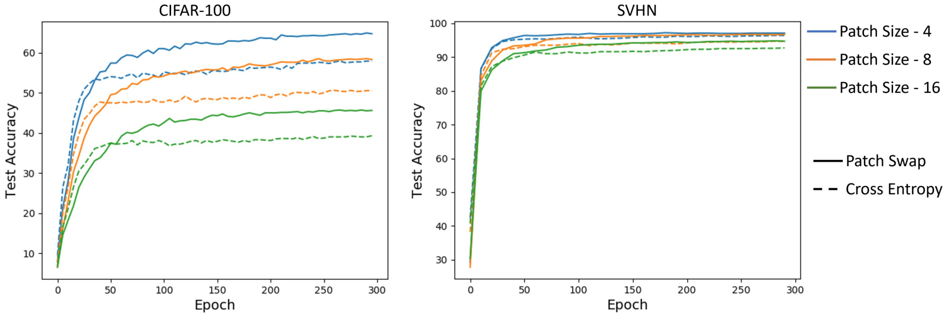

Figure 1: Test accuracy graph for training process on CIFAR-100 (left) and SVHN (right). Different patch sizes used for training the transformers are denoted by different colors. Solid lines represent PatchSwap and dashed lines represent Cross-entropy loss.

Method	CIFAR-10	CIFAR-100
Cross Entropy	95.1	76.0
Label smoothing [4]	95.1	76.4
Cutout [1]	95.9	76.8
Mixup [6]	96.0	77.1
Cutmix [5]	**96.2**	**78.2**
PatchSwap-1	94.1	73.4
PatchSwap-2	95.3	75.8
PatchSwap-4	95.9	77.6
PatchSwap-8	**96.2**	78.0
PatchSwap-16	96.0	77.8

Table 1: Top-1 classification accuracies on CIFAR-10 and CIFAR-100 for ResNet-18. PatchSwap-*k* denotes PatchSwap was performed using *k* patch size. **Bold** denotes the highest performance and underline denotes the second highest.

2 PatchSwap for ConvNets?

PatchSwap is designed for Vision Transformers however, the regularization being applied at the input level makes it compatible with ConvNets as well. However, while using PatchSwap for ConvNets patch size become a hyper-parameter. We apply PatchSwap with different patch sizes - $\{1,2,4,8,16\}$ on the input. We showcase the impact of training a ResNet-18 on CIFAR-10 and CIFAR-100 with different loss functions. Similar to transformer experiments, we run the baselines using a similar setting. The images are resized to 32×32 to be compatible with the network. The network is trained with a batch size of 128 for 300 epochs, using an SGD optimizer with a learning rate of 0.1 and momentum of 0.9. The learning rate is warmed up over the first 10 epochs and then decayed by a factor of 0.1 at 150^{th} and 225^{th} epoch.

The results are in Table 1. PatchSwap-*k* denotes the PatchSwap was performed using *k* patch size. Resnet-18 achieves much higher accuracy than our Vision Transformer mainly due to its translation equivariance and locality inductive biases [2]. To acquire these, the Vision Transformers require a huge amount of data [2]. Resnet-18 has 11 million parameters whereas the transformer has only about one-third of its parameters. PatchSwap does yield good performance for ConvNets as well. However, this requires tuning the patch size hyper-parameter. PatchSwap achieves the best performance with $k = 8$ for CIFAR-10 and CIFAR-100.

3 Training Losses

Cross-entropy is the standard loss for training a neural network. However, it results in poor generalization. Nowadays, Mixup and Cutmix are replacing cross-entropy loss as the standard network training procedure. At any training step, either Mixup or Cutmix is chosen with equal probability. This procedure is used for training Vision Transformers as well [3]. In this section, we show the impact of adding PatchSwap to Mixup-Cutmix by applying one of the three at each step. We perform the experiments on CIFAR-100 and Tiny-Imagenet and the results are shown in Table 2. Cross-entropy results in a model with maximum overfitting. This is alleviated to an extend by PatchSwap. Mixup+Cutmix achieve higher accuracy than PatchSwap but using all the three losses results in the highest performance with minimum overfitting.

Dataset	CIFAR-100			Tiny-ImageNet	
Patch Size	4	8	16	8	16
Cross Entropy	57.9	50.6	39.3	41.9	34.4
PatchSwap	64.9	58.5	45.7	49.9	41.8
Mixup+CutMix	66.4	60.3	47.7	51.2	43.8
Mixup+Cutmix+PatchSwap	**67.2**	**60.9**	**48.3**	**52.3**	**44.6**

Table 2: Top-1 classification accuracies on CIFAR-100 and Tiny-ImageNet using different combination of losses for training. **Bold** denotes the highest performance.

4 Additional Results

In this section, we show additional results on CIFAR-10, CIFAR-100, SVHN, and Fashion-MNIST datasets with different augmentation settings than the main paper. The results are in Table 3, 4 and 5. PatchSwap outperforms other approaches on CIFAR-10 and CIFAR-100 datasets. In the case of SVHN and FashionMNIST (no augmentation used), when the patch size is high, the performance of PatchSwap degrades. We believe this is because a high patch size results in a small number of patches. This leads to only a small number of combinations for PatchSwap. Also, with no augmentation, the regularization is limited. As the PatchSwap size reduces, the number of patches increases, leading to an increase in the number of combinations and the performance as well.

Patch Size	4	8	16
Cross Entropy	87.6	83.2	74.8
Label smoothing [4]	87.4	83.6	74.9
Cutout [1]	88.3	84.3	74.4
Mixup [6]	89.1	85.1	75.8
Cutmix [5]	90.3	85.7	**76.8**
PatchSwap	**90.4**	**87.0**	**76.8**

Table 3: Comparison of Top-1 classification accuracies on CIFAR-10 dataset (with RandAugment augmentation) using different patch sizes.

Patch Size	4		8		16	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Cross Entropy	63.7	86.2	57.3	81.5	44.5	70.4
Label smoothing [4]	64.1	84.8	57.7	80.8	45.3	70.4
Cutout [1]	63.8	86.2	56.9	82.0	44.6	71.7
Mixup [6]	66.1	87.8	60.1	83.4	48.5	74.7
Cutmix [5]	67.9	88.7	60.5	84.3	48.2	75.1
PatchSwap	**68.0**	**89.1**	**61.2**	**86.0**	**48.8**	**76.5**

Table 4: Comparison of Top-1 and Top-5 classification accuracies on CIFAR-100 dataset (with RandAugment augmentation) using different patch sizes.

Dataset	FashionMNIST			SVHN		
Patch Size	4	8	16	4	8	16
Cross Entropy	90.1	91.0	90.1	93.8	91.8	88.9
Label smoothing [4]	89.8	91.3	90.5	94.1	92.0	88.8
Cutout [1]	92.8	92.2	91.5	93.5	**94.2**	90.9
Mixup [6]	91.3	92.8	91.7	94.7	92.9	90.8
Cutmix [5]	92.1	**93.0**	**92.0**	96.1	94.0	**92.0**
PatchSwap	**93.2**	92.2	91.4	**96.4**	93.3	89.7

Table 5: Comparison of Top-1 classification accuracies on FashionMNIST and SVHN datasets (with no augmentation) using different patch sizes.

5 PatchSwap Images

We show sample PatchSwap images generated with different PatchSwap sizes in Figure 2. PatchSwap size of 1 is pixel-wise swapping. The λ values are rounded to the closest multiple of $\frac{1}{N}$ where N is the number of patches. For example, when $\lambda = 0.33$ and the PatchSwap size is 32 for an image of size 64×64, the total number of patches is equal to 4. So, the closest value to λ is 0.25.

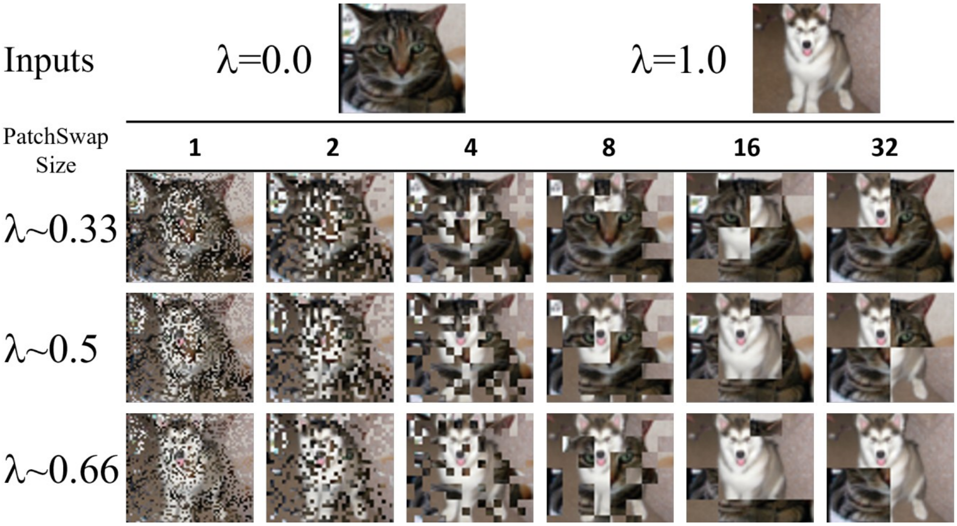

Figure 2: Sample PatchSwap Images with different PatchSwap sizes. The first row shows the original images. The second row displays the PatchSwap size. The next 3 rows display sample PatchSwap images for varying λ. Best viewed in color.

6 More Attention Maps

In this section, we show additional attention maps for different types of settings for TinyImageNet. We show more class-specific attention maps similar to the ones shown in the main paper for PatchSwap images are displayed in Figure 3. The attention maps for the original (non-mixed) images in in Figure 4. We also show the attention maps for PatchSwap images belonging to fine-grained classes (for example, Egyptian cat v/s Persian cat) in 5.

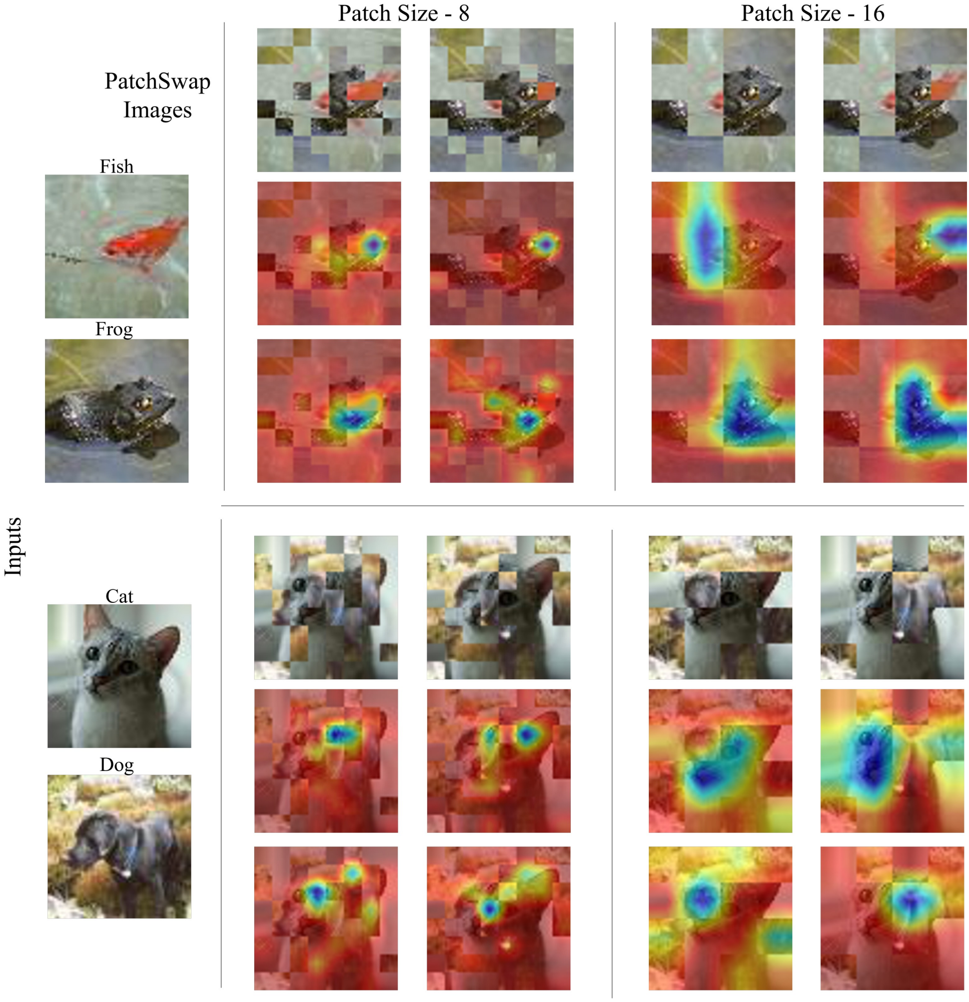

Figure 3: Class-specific Attention Maps for PatchSwap images.

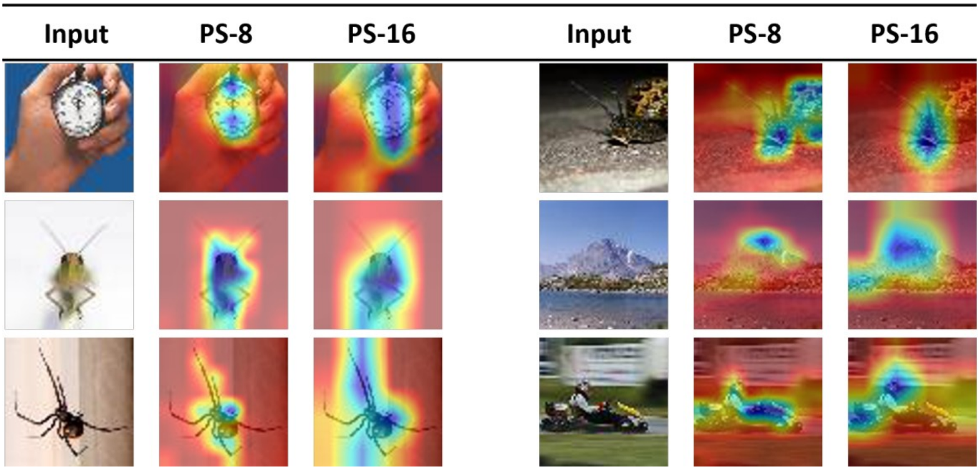

Figure 4: Attention maps for TinyImagenet images. PS-*k* denotes the attention maps using a Vision Transformer trained with patch size of *k*.

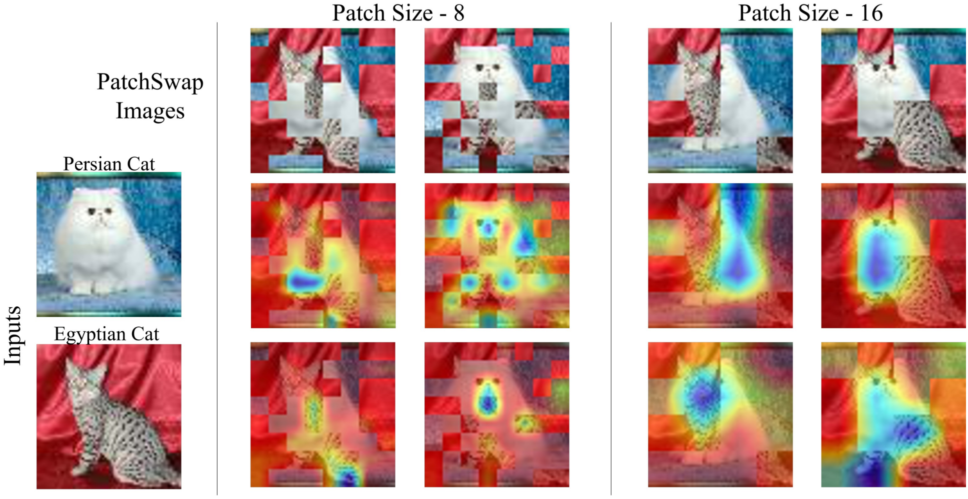

Figure 5: Class-specific attention maps for PatchSwap images for fine-grained classes (Persian cat and Egyptian cat).

References

[1] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. URL http://arxiv.org/abs/1708.04552.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[5] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.