Supplementary Materials: Augmenting Supervised Neural Networks with Unsupervised Objectives for Large-scale Image Classification

Yuting Zhang YUTINGZH@UMICH.EDU

Kibok Lee KIBOK@UMICH.EDU

Honglak Lee HONGLAK@EECS.UMICH.EDU

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA

S1. Parameters for VGGNet-based models

Macro-layer	Learning rate	Loss weighting [1]
	SAE-layerwise	SAE-layer/all
1	3×10^{-9}	1×10^{-4}
2	1×10^{-8}	1×10^{-12}
3	3×10^{-12}	1×10^{-12}
4	1×10^{-12}	1×10^{-12}
5	1×10^{-11}	1×10^{-10}

LR: learning rate; [1] the top-level softmax is weighted by 1.

Table S-1. Layer-wise training parameters for networks augmented from VGGNet

We report the learning parameters for 16-layer VGGNet-based model in Table S-1. We chose the learning rates that lead to the largest decrease in the reconstruction loss in the first 2000 iterations for each layer. The "loss weighting" are balancing factors for reconstruction losses in different layers varied to make them comparable in magnitude. In particular, we computed image reconstruction loss against RGB values normalized to [0,1], which are different in scale from intermediate features. We also did not normalize the reconstruction loss with feature dimensions for any layer.

S2. More experimental results and discussions

S2.1. Learned filters

Compared to the baseline VGGNet, the finetuned SWWAE-all model demonstrated $\sim 35\%$ element-wise relative change of the filter weights on average for all the layers. A small portion of the filters showed stronger contrast after finetuning. Qualitatively, the finetuned filters kept the pretrained visual shapes. In Figure S-1, we visualize the first-layer 3×3 convolution filters.

(a) Pretrained VGGNet (b) Finetuned SWWAE-all

Figure S-1. Visualization of the normalizaed first-layer convolution filters in 16-layer VGGNet-based network. The filters of the SWWAE-all model had nearly the same patterns to those of the pretrained VGGNet, but showed stronger contrast. It is more clear see the difference if displaying the two images alternatively in the same place. (online example: `http://www.ytzhang.net/files/publications/2016-icml-recon-dec/filters/`)

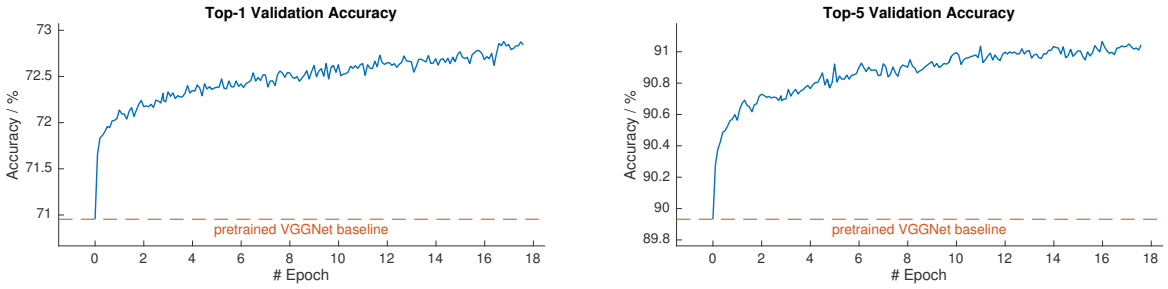


Figure S-2. Training curves for the single-crop validation accuracy of VGGNet-based SWWAE-all models.

S2.2. Training curve

In Figure S-2, we report the training curves of validation accuracy for SWWAE-all, where the pretrained VGGNet classification network and decoder network were taken as the starting point.

S2.3. Selection of different model variants

The performance for different variants of the augmented network are comparable, but we can still choose the best available one. In particular, we provide following discussions.

- Since the computational costs were similar for training and the same for testing, we can use the best available architecture depending on tasks. For example, when using decoding pathways for spatially corresponded tasks like reconstruction (as in our paper) and segmentation, we can use the SWWAE. For more general objectives like predicting next frames, where pooling switches are non-transferrable, we can still use ordinary SAEs to get competitive performance.
- S(WW)AE-first has less hyper-parameters than S(WW)AE-all, and can be trained first for quick parameter search. It can be switched to *-all for better performance.

S2.4. Ladder networks

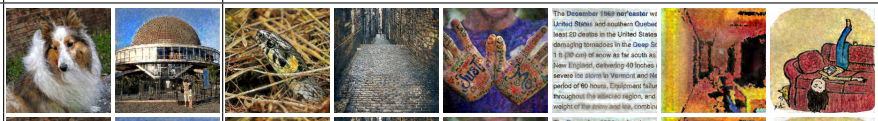
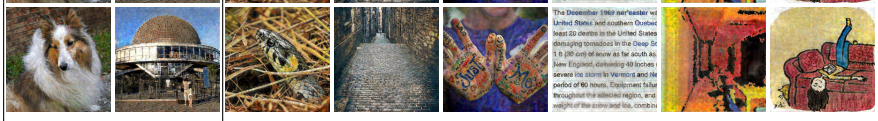
We tried training a ladder network following the same procedures of pretraining auxiliary pathways and finetuning the whole network as for our models, which is also similar to Rasmus et al. (2015)'s strategy. We used the augmented multi-layer perceptron (AMLP) combinator, which Pezeshki et al. (2016) proposed as the best combinator function. Different

from the previous work conducted on the variants of MNIST dataset, the pretrained VGGNet does not have batch normalization (BN) layers, which pushed us to remove the BN layers from the ladder network. However, BN turned out to be critical for proper noise injection, and the non-BN ladder network did not perform well. It might suggest that our models are easier to pair with a standard convolutional network and train on large-scale datasets.

S2.5. Image reconstruction

In Figure S-3, we visualize the images reconstructed by the pretrained decoder of SWWAE-first and the final models for SWWAE-first/all, and reported the L2 reconstruction loss on the validation set. Finetuning the entire networks also resulted in better reconstruction quality, which is consistent with our assumption that enhancing the ability of preserving input information can lead to better features for image classification. Since the shape details had already been well recovered by the pretrained decoder, the finetuned SWWAE-first/all mainly improved the accuracy of colors. Note that the decoder learning is more difficult for SWWAE-all than SWWAE-first, which explains its slightly higher reconstruction loss and better regularization ability.

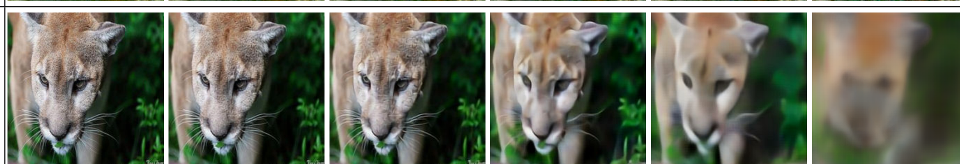
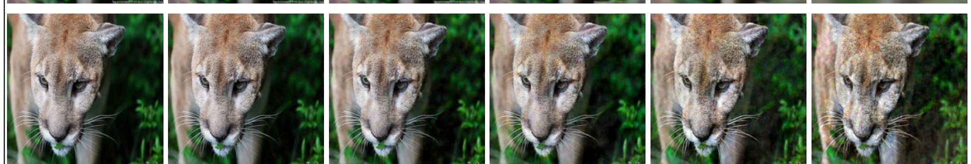
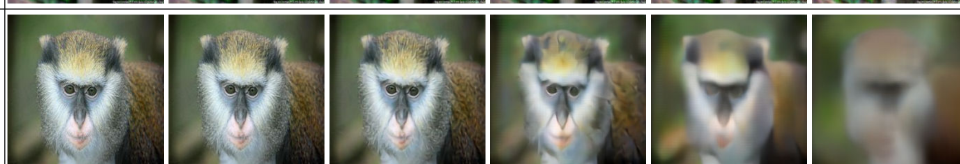
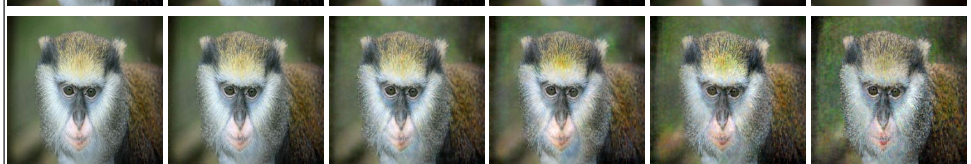
In Figure S-4 and S-5, we showed more examples for reconstructing input images from pretrained neural network features for AlexNet and VGGNet.

Model	L2 Loss	ImageNet	Non-ImageNet [1]
Ground truth	-		
SWWAE-first (Pretrained, fixing encoder)	513.4		
SWWAE-first (Finetuned with encoder)	462.2		
SWWAE-all (Finetuned with encoder)	493.0		

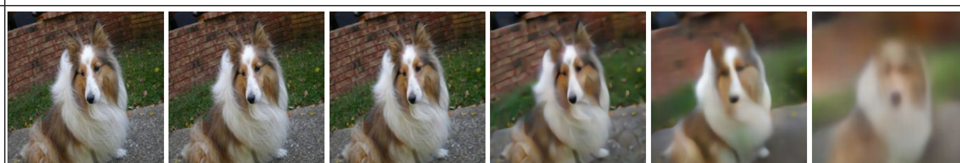
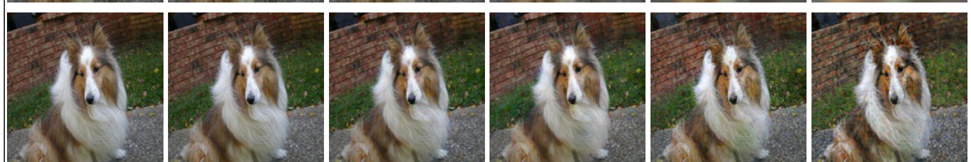
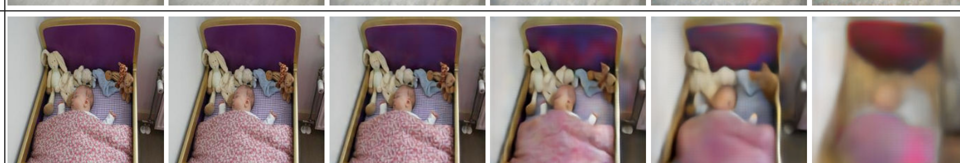
[1] The first three images are from morguefile.com; the fourth is a screenshot of Wikipedia; the fifth is a depth image from NYU dataset; the last is used with permission from Debbie Ridpath Ohi at Inkygirl.com

Figure S-3. Image reconstruction from pool5 features to images. The reconstruction loss is computed on the ILSVRC2012 validation set and measured with L2-distance with the ground truth (RGB values are in $[0, 1]$). The first 2 example images are from the ILSVRC2012 validation set (excluding the 100 categories). The rest are not in ImageNet.

Figure S-4. AlexNet reconstruction on ImageNet ILSVRC2012 validation set. (Best viewed when zoomed in on a screen.)

Layer	image	pool1	pool2	pool3	pool4	pool5
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						

(continued on next page)

Layer	image	pool1	pool2	pool3	pool4	pool5
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						
SAE-first (**fixed** unpooling switches)						
SWWAE-first (**known** unpooling switches)						

(continued on next page)

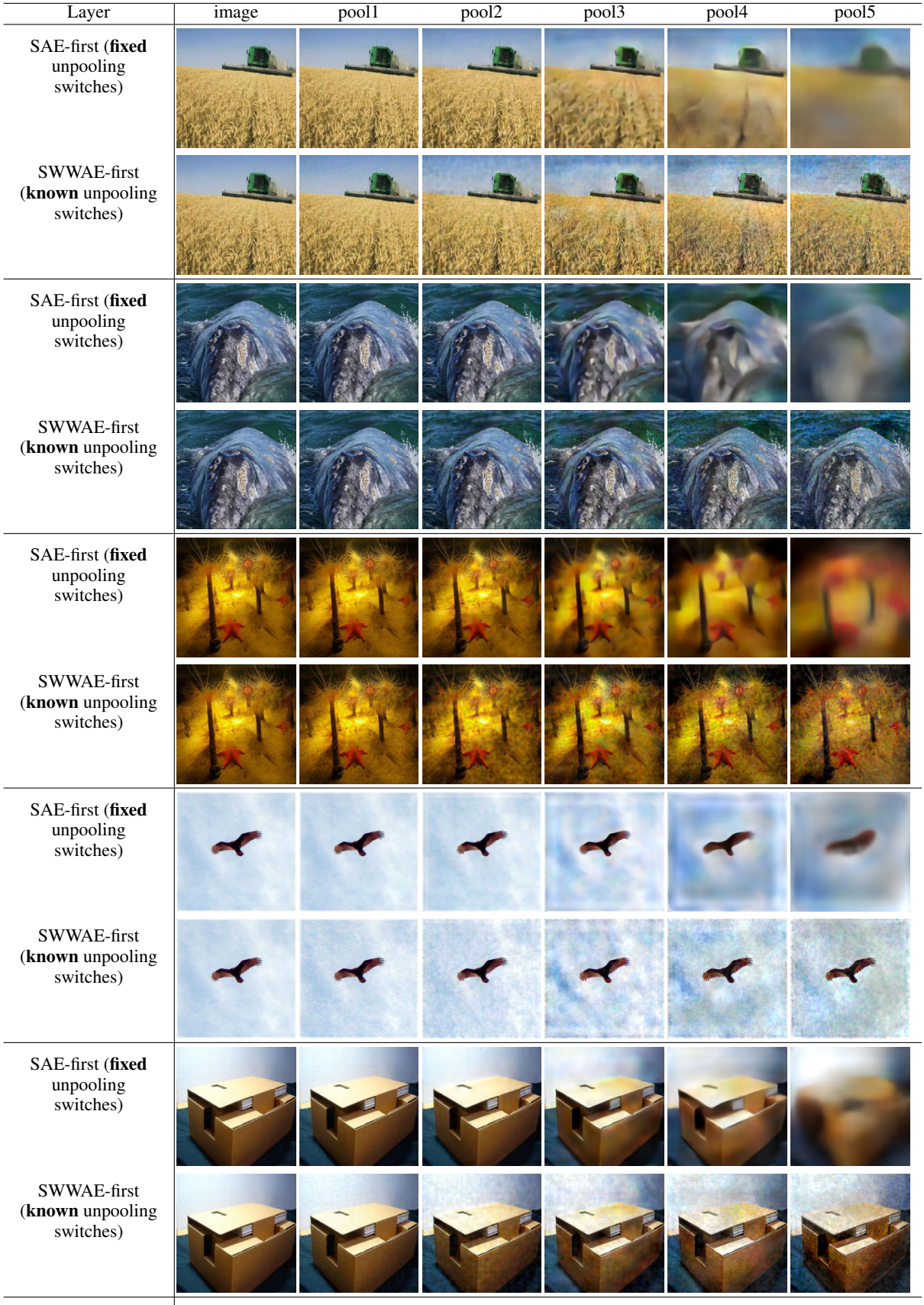


Figure S-5. VGGNet reconstruction on ImageNet ILSVRC2012 validation set. (Best viewed when zoomed in on a screen.)

References

Dosovitskiy, A. and Brox, T. Inverting visual representations with convolutional networks. *CVPR*, 2016.

Pezeshki, M., Fan, L., Brakel, P., Courville, A., and Bengio, Y. Deconstructing the ladder network architecture. *arXiv:1506.02351*, 2016.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. Semi-supervised learning with ladder network. In *NIPS*, 2015.