Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion

Galina Boeva

MIPT, 2024

March 26, 2024

Motivation

Main idea

Often having a huge predictive model, we would like to transfer knowledge from it to a more lightweight version that would be easy to build, for example, on a phone, but then the question arises of a training sample, we would like to transfer knowledge without data transfer.

Background

Knowledge distillation [1] - Transfer of knowledge from one model to another was first introduced by Breiman and Shang when they learned a single decision tree to approximate the outputs of multiple decision trees. [2] - synthesize inputs based on pre-stored auxiliary layer-wise statistics of the teacher network. **Image synthesis** An alternative area of work without GAN in the field of security focuses on the synthesis of images from a single CNN. [3] propose an attack using model inversion to obtain class images from the network by gradient descent over the input data. [4] has enabled the "dreaming" of new object features onto natural images given a single pretrained CNN.

Idea of method DeepInversion

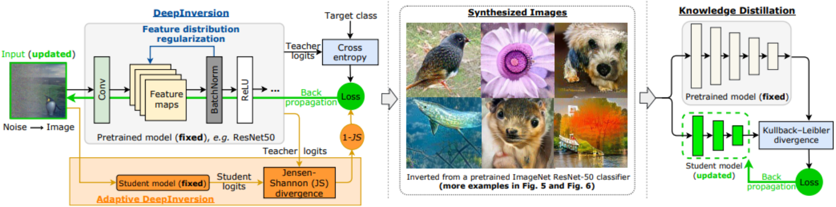

Figure: We introduce DeepInversion, a method that optimizes random noise into high-fidelity class-conditional images given just a pretrained CNN (teacher). Further, we introduce Adaptive DeepInversion, which utilizes both the teacher and application-dependent student network to improve image diversity. Using the synthesized images, we enable data-free pruning, introduce and address data-free knowledge transfer, and improve upon data-free continual learning.

Method: Knowledge distillation&DeepDream

Knowledge distillation

Given a trained model p_T and a dataset X , the parameters of the student model, W_S, can be learned by $\min_{\mathbf{W}_S} \sum_{x \in X} KL(p_T(x), p_S(x))$, where $p_T(x) = p(x, \mathbf{W}_T)$ and $p_S(x) = p(x, \mathbf{W}_S)$ are the output distributions produced by the teacher and student model.

DeepDream

DeepDream is also suitable for optimizing noise into images. Given a randomly initialized input and an arbitrary target label y, the image is synthesized by optimizing $\min_{\hat{x}} L(\hat{x}, y) + R(\hat{x})$, where $L(\hat{x}, y)$ is a classification loss and $R(\hat{x})$ is an image regularization term. DeepDream uses an image prior to steer $\hat{x}$ away from unrealistic images with no discernible visual information: $R_{prior}(\hat{x}) = \alpha_{tv} R_{TV}(\hat{x}) + \alpha_{l_2} R_{l_2}(\hat{x})$, where R_{TV} and R_{l_2} penalize the total variance and l_2 norm of $\hat{x}$, respectively, with scaling factors $\alpha_{tv}, \alpha_{l_2}$.

DeepInversion

DeepInversion

The feature distribution regularization term can be formulated as:

$$R_{feature}(\hat{x}) = \sum l||\mu_l(\hat{x}) - \mathbb{E}(\mu_l(x)|X)||_2 + \sum l||\sigma_l^2(\hat{x}) - \mathbb{E}(\sigma_l^2(x)|X)||_2,$$

where $\mu_l(\hat{x})$ and $\sigma_l^2(\hat{x})$ are the batch-wise mean and variance estimates of feature maps corresponding to the l^{th} convolutional layer:

$$\mathbb{E}(\mu_l(x)|X) \simeq BN_l(prunning\_mean),$$

$$\mathbb{E}(\sigma_l^2(x)|X) \simeq BN_l(prunning\_variance).$$

$R(\cdot)$ can thus be expressed as

$$R_{DI}(\hat{x}) = R_{prior}(\hat{x}) + \alpha_f R_{feature}(\hat{x})$$

Adaptive DeepInversion

Adaptive DeepInversion

Introduce an additional loss $R_{compete}$ for image generation based on the Jensen-Shannon divergence that penalizes output distribution similarities,

$$R_{compete}(\hat{x}) = 1 - JS(p_T(\hat{x}), p_S(\hat{x})).$$

$$R_{ADI}(\hat{x}) = RDI(\hat{x}) + \alpha_c Rcompete(\hat{x})$$

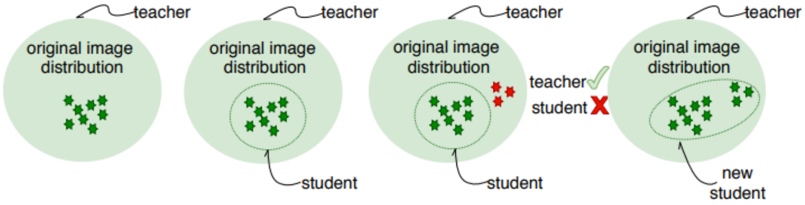

Figure: Illustration of the Adaptive DeepInversion competition scheme to improve image diversity.

Results on CIFAR-10

Teacher Network	VGG-11	VGG-11	ResNet-34
Student Network	VGG-11	ResNet-18	ResNet-18
Teacher accuracy	92.34%	92.34%	95.42%
Noise ($\mathcal{L}$)	13.55%	13.45%	13.61%
$+\mathcal{R}_{\text{prior}}$ (DeepDream [46])	36.59%	39.67%	29.98%
$+\mathcal{R}_{\text{feature}}$ (DeepInversion)	84.16%	83.82%	91.43%
$+\mathcal{R}_{\text{compete}}$ (ADI)	**90.78%**	**90.36%**	**93.26%**
DAFL [7]	–	–	92.22%

Figure: Data-free knowledge transfer to various students on CIFAR-10.

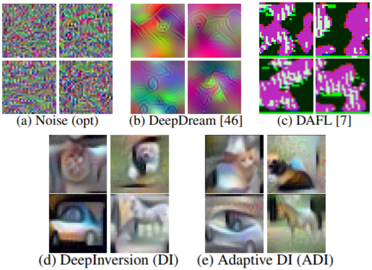

Figure: 32×32 images generated by inverting a ResNet-34 trained on CIFAR-10.

Results on ImageNet

Figure: Class-conditional 224 ˆ 224 samples obtained by DeepInversion, given only a ResNet-50 classifier trained on ImageNet and no additional information. Note that the images depict classes in contextually correct backgrounds, in realistic scenarios.

Results on ImageNet

Model	DeepDream top-1 acc. (%)	DeepInversion top-1 acc. (%)
ResNet-50	100	100
ResNet-18	28.0	**94.4**
Inception-V3	27.6	**92.7**
MobileNet-V2	13.9	**90.9**
VGG-11	6.7	**80.1**

Figure: Classification accuracy of ResNet-50 synthesized images by other ImageNet-trained CNNs.

Method	Resolution	GAN	Inception Score
BigGAN [5]	256	✓	178.0 / 202.6^{+}
DeepInversion (**Ours**)	224		60.6
SAGAN [75]	128	✓	52.5
SNGAN [43]	128	✓	35.3
WGAN-GP [16]	128	✓	11.6
DeepDream [46]*	224		6.2

Figure: Inception Score (IS) obtained by images synthesized by various methods on ImageNet.

Data-free Knowledge Transfer

Figure: Class-conditional 224 $\times$ 224 images obtained by DeepInversion given a ResNet50v1.5 classifier pretrained on ImageNet. Classes top to bottom: (left) daisy, volcano, quill, (right) cheeseburger, brown bear, trolleybus.

Data-free Pruning & Data-free Continual Learning

Image Source	Top-1 acc. (%) −50% filters −71% FLOPs	−20% filters −37% FLOPs
No finetune	1.9	16.6
Partial ImageNet		
0.1M images / 0 label	69.8	74.9
Proxy datasets		
MS COCO	66.0	73.8
PASCAL VOC	54.4	70.8
GAN		
Generator, BigGAN	63.0	73.7
Noise (**Ours**)		
DeepInversion (DI)	55.9	72.0
Adaptive DeepInversion (ADI)	60.7	73.3

Figure: ImageNet ResNet-50 pruning comparison with prior work.

Method	Top-1 acc. (%) Combined	ImageNet	CUB	Flowers
ImageNet + CUB (1000 → 1200 outputs)				
LwF.MC [56]	47.64	53.98	41.30	–
DeepDream [46]	63.00	56.02	**69.97**	–
DeepInversion (**Ours**)	**67.61**	**65.54**	69.68	–
Oracle (distill)	69.12	68.09	70.16	–
Oracle (classify)	68.17	67.18	69.16	–
ImageNet + Flowers (1000 → 1102 outputs)				
LwF.MC [56]	67.23	55.62	–	78.84
DeepDream [46]	79.84	65.69	–	**94.00**
DeepInversion (**Ours**)	**80.85**	**68.03**	–	93.67
Oracle (distill)	80.71	68.73	–	92.70
Oracle (classify)	79.42	67.59	–	91.25
ImageNet + CUB + Flowers (1000 → 1200 → 1302 outputs)				
LwF.MC [56]	41.72	40.51	26.63	58.01
DeepInversion (**Ours**)	**74.61**	**64.10**	**66.57**	**93.17**
Oracle (distill)	76.18	67.16	69.57	91.82
Oracle (classify)	74.67	66.25	66.64	91.14

Figure: Continual learning results that extend the network output space, adding new classes to ResNet-18.

Literature

1. **Main article** Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion.

Leo Breiman and Nong Shang.
Born again trees.
University of California, Berkeley, Berkeley, CA, Technical Report, 1(2):4, 1996.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner.
Data-free knowledge distillation for deep neural networks.
arXiv preprint arXiv:1710.07535, 2017.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart.
Model inversion attacks that exploit confidence information and basic countermeasures.
In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka.
Inceptionism: Going deeper into neural networks.
Google research blog, 20(14):5, 2015.