Supplementary Material for: Interpret Neural Networks by Identifying Critical Data Routing Paths

Yulong Wang Hang Su Bo Zhang Xiaolin Hu
Tsinghua National Lab for Information Science and Technology
Beijing National Research Center for Information Science and Technology, BNRist Lab
Department of Computer Science and Technology, Tsinghua University
{wang-yl15@mails,suhangss@mail,dcszb@mail,xlhu@mail}.tsinghua.edu.cn

The following document shows further results which are not included in the submitted paper. We apply our proposed method *Distillation Guided Routing* (DGR) on AlexNet [2] and ResNet-50 [1] models. For AlexNet, we identify the critical routing nodes in the *Critical Data Routing Paths* (CDRPs) at the 5 convolutional layers' output channels. For ResNet, we identify the critical routing nodes in CDRPs at the 16 bottleneck block layers' output channels. The hyperparameter settings are the same with the experiments of VGG-16 model [4]. We perform SGD on the same input $\boldsymbol{x}$ for $T = 30$ iterations, with learning rate of 0.1, momentum of 0.9 and no weight decay. Balanced parameter γ in Equation (1) in the main body is set to 0.05.

1. Quantitative Analysis

In this section, we report classification accuracy results of the subnetwork outlined by identified critical data routing paths. We compare our method against two baseline methods, *Adaptive Weight Routing* (AWR) and *Adaptive Activation Routing* (AAR), both of which trim out the CDRPs iteratively based on weights norms and activations norms. For ResNet model, since we only perform at the end of each bottleneck block layer's output, in which there are no explicit corresponding weights involved, we only compare against with AAR baseline method.

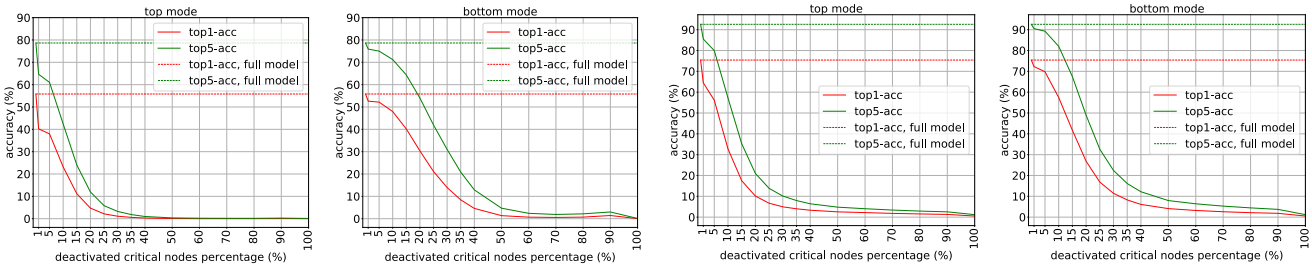
Table 1 summarizes the performance of our method in terms of top-5 accuracy and sparsity. Since the best routing paths selection criterion requires the top-1 prediction to be same with the full model, all the methods achieve the same top-1 accuracy. However, our method achieves the highest top-5 accuracy compared to other baseline methods, and only suffers minor top-5 accuracy degradation compared to the full model. Our method also achieves far more sparse routing paths compared to the baseline methods.

Ablation study We also further validate the CDRPs by partially deactivating the critical nodes on the identified CDRPs in the original full model, while keeping other non-critical nodes unchanged.

Figure 1a and 1b show the model accuracy degradation with different fractions of critical nodes being deactivated in *Top Mode* and *Bottom Mode*, which deactivates the critical nodes starting from *larger* and

Table 1: Adaptive routing methods comparison with same top-1 prediction requirement. For sparsity, lower is better

Methods	Top-1 Acc.	Top-5 Acc.	Sparsity
AlexNet Full Model (%)	55.81	78.65	100.00
AWR (%)	55.81	73.68	93.20 ± 0.31
AAR (%)	55.81	71.71	88.13 ± 1.03
DGR (Ours) (%)	55.81	**77.21**	$\mathbf{22.43} \pm \mathbf{4.67}$
ResNet-50 Full Model (%)	75.42	92.58	100.00
AAR (%)	75.42	87.32	72.32 ± 1.34
DGR (Ours) (%)	75.42	**89.14**	$\mathbf{12.02} \pm \mathbf{4.50}$



(a) AlexNet (b) ResNet-50

Figure 1: The accuracy degradation when critical nodes are deactivated in the original full model, with *Top Mode* and *Bottom Mode*. Only small fractions of critical nodes being deactivated will lead severe performance degradation

smaller values of control gates respectively. With only small fractions of critical nodes on CDRPs pruned out in the network, the model performance deteriorates severely, which validates the CDRPs identified by our method are effective.

2. Semantic Concepts Emerge in CDRPs

Functional process of intra-layer routing nodes We use the t-SNE [3] method to display 50,000 ImageNet validation images' intra-layer routing nodes representations in 2D embedding. We regard all the individual critical nodes in a certain layer composing the intra-layer routing nodes. The encoding representation is simply the optimized control gates $\boldsymbol{\lambda}_k^*$ for the k-th layer.

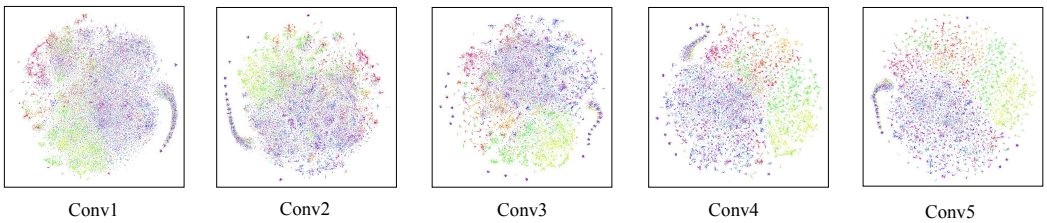


Figure 2: t-SNE 2D embedding for AlexNet network.

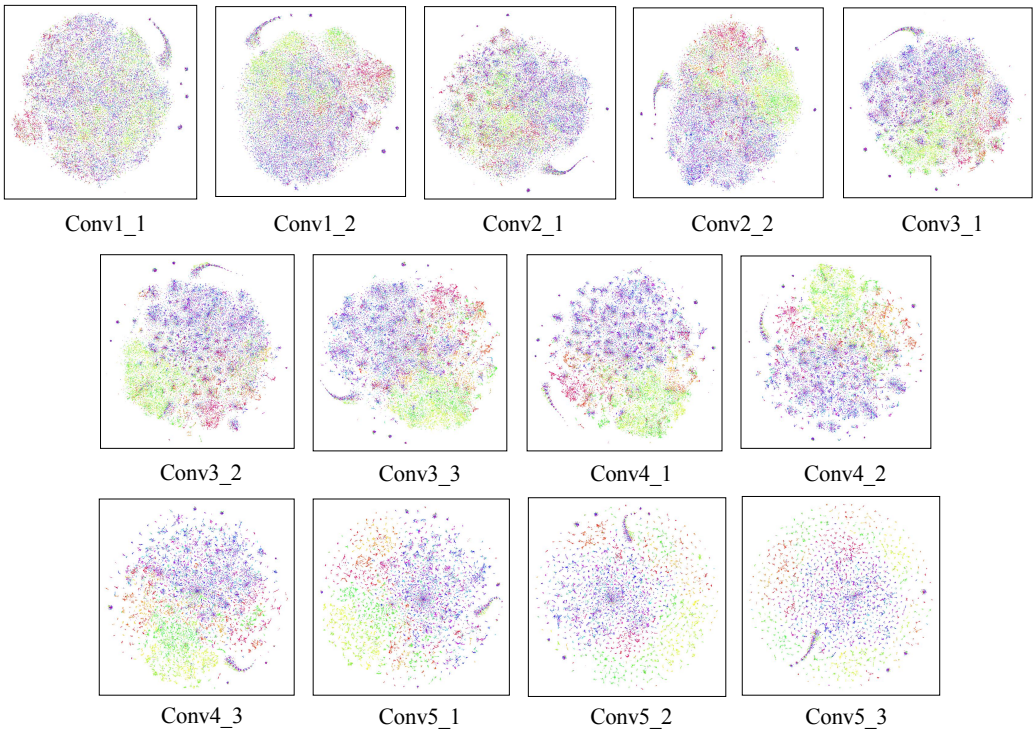


Figure 3: t-SNE 2D embedding for VGG-16 network.

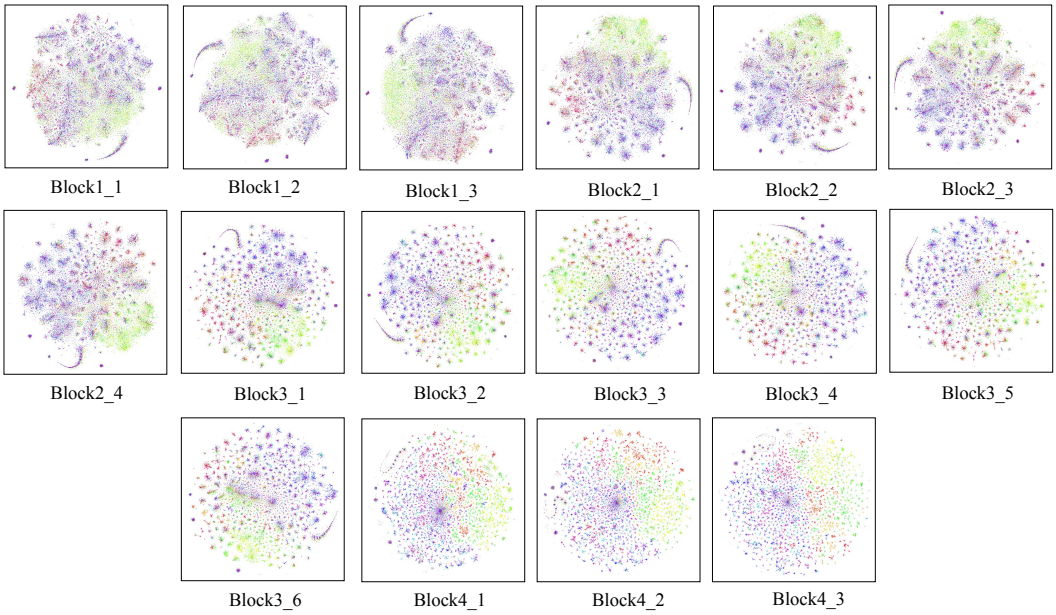


Figure 4: t-SNE 2D embedding for ResNet-50 network.

Figure 2, 3 and 4 show all the intra-layer routing nodes of convolutional layers and bottleneck block layers in AlexNet, VGG-16 and ResNet-50 models. Each point stands for a single image. Points with the same ground-truth labels are painted in the same color for visual effect. The more scattered

Table 2: The Area-Under-Curve (AUC) score for different binary classifier on adversarial sample detection by discriminating CDRPs of real and adversarial image. Higher is better.

	Num. of training samples	1	5	10
AlexNet	random forest	0.7220	0.7325	0.7770
	adaboost	0.7415	0.7505	0.7630
	gradient boosting	0.7525	0.7590	0.7845
Resnet-50	random forest	0.9120	0.9190	0.9210
	adaboost	0.9145	0.9195	0.9200
	gradient boosting	0.9255	0.9315	0.9345

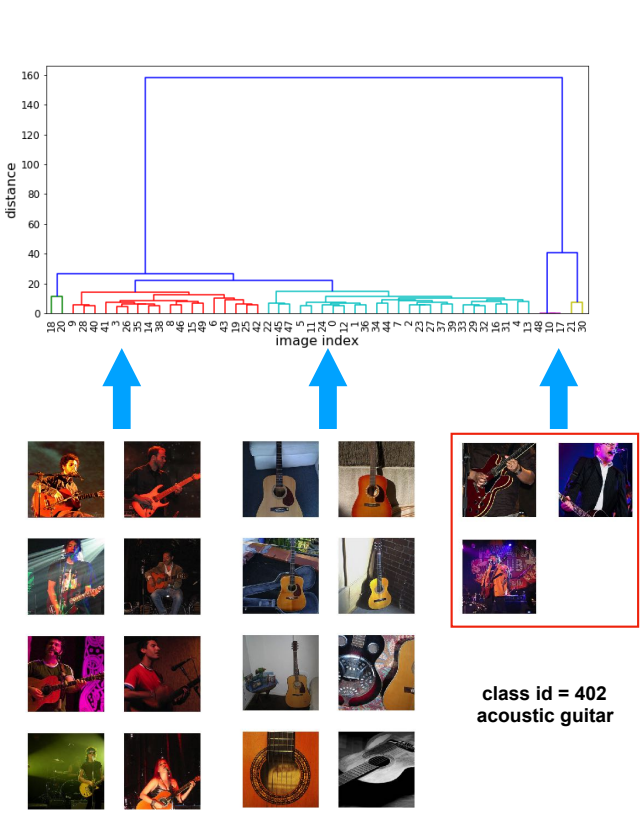
points in the embedding indicate more discriminative ability of the corresponding intra-layer routing nodes representation. From the figures, we can discover the degree of embedding discriminative ability increases with ascending layer levels throughout three models. Moreover, for high level layers of deeper models like VGG-16 and ResNet-50, the discriminative ability is more significant that those in the shallower model like AlexNet.

Intra-class sample clustering The CDRPs also reflect the input data layout patterns. Figure 5a and 5b show the agglomerative clustering results on the intra-class samples using the whole CDRPs representations obtained from VGG-16 network. We can discover that the clustering results correspond to input layout patterns strongly. For example in Figure 5b, for the class 'car wheel', we find three typical clusters. The second cluster mainly consists of the car body and wheels. The third cluster shows a single wheel in the front view. The first cluster mainly consists of hard examples to classify. Particularly, there is an image of the whole sports car, resulting in semantic ambiguity. Figure 5a also shows similar patterns. These results indicate that the identified CDRPs reflect input patterns, and help to find out hard examples or complex samples in the dataset.

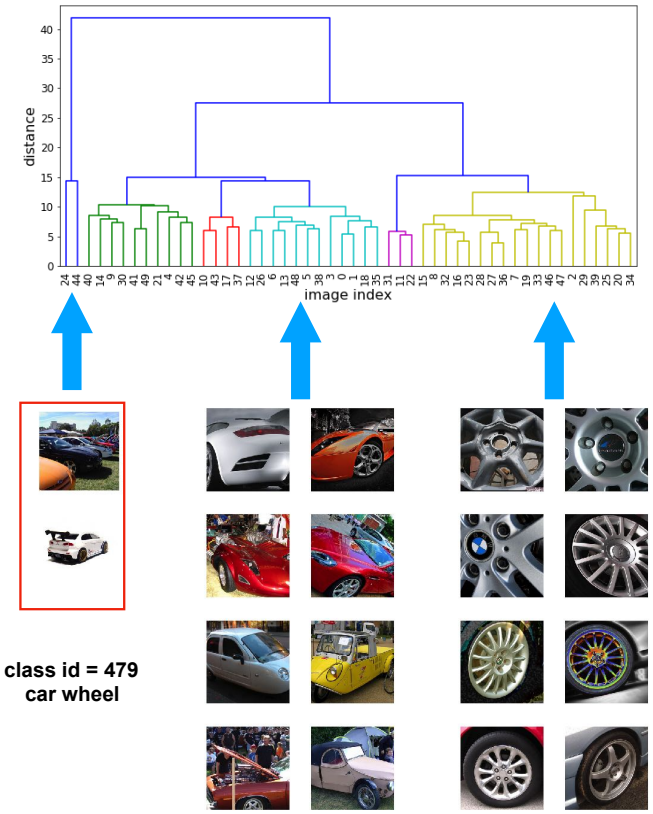
3. Adversarial Sample Detection

We apply the proposed adversarial sample detection scheme in Section 3 on the CDRPs representations obtained from AlexNet and ResNet-50 to discriminate whether the CDRPs are from real or adversarial samples. The experiment settings are the same with VGG-16 network, in which we randomly sample 1/5/10 images and 1 image of each class from ImageNet training and validation datasets as training and test samples respectively. Each sample is accompanied by an adversarial image, which is generated by Equation (5). The target classes are from a random permutation of original classes. Table 2 summarizes the binary classifiers' performance in AUC score. Compared to VGG-16 network, the CDRPs representations obtained from AlexNet are less discriminative when dealing with real and adversarial images. The CDRPs representations from ResNet-50 are more discriminative than those of VGG-16 network. Our results demonstrate that without complicated algorithm, the adversarial attacking can be defended based on the discriminative CDRPs representation.

We also apply our method on large scale dataset. We use all the 50,000 validation images in ImageNet dataset as test samples, and evenly sample 50,000 training images from 1,000 classes in ImageNet training dataset. Table 3 summarizes the results. Our method can achieve higher defense success rate even with larger scale dataset, which validates that our method's scalability and effectiveness in detecting the



(a) The first cluster consists of singers playing guitar on the stage, and the second cluster shows the front view of guitars. Red bounding box indicates samples hard to classify, which include an image with little area to show the guitar on the corner.



(b) The second cluster mainly consists of the car body and wheels. The third cluster shows the front view of car wheels. Red bounding box indicates samples hard to classify, which include an image of the whole sports car

Figure 5: Intra-class sample clustering helps identify hard examples in the dataset. The top picture shows agglomerative clustering results. The typical images of each cluster are shown below. Red bounding box indicates hard examples.

Table 3: The Area-Under-Curve (AUC) score for different numbers of training and test samples in our proposed adversarial sample detection scheme. All the CDRPs representations are obtained from VGG-16 network.

Num. of training samples	Num. of test samples	random forest	adaboost	gradient boosting
1	1	0.8792	0.8877	0.9051
5	1	0.8942	0.9057	0.9189
10	1	0.9041	0.9104	0.9147
50	50	0.9289	0.9084	0.9220

adversarial samples.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.