A Peek Into the Reasoning of Neural Networks: Interpreting with Structural Visual Concepts

Yunhao Ge[1,2], Yao Xiao[2], Zhi Xu[2], Meng Zheng[1], Srikrishna Karanam[1],
Terrence Chen[1], Laurent Itti[2], and Ziyan Wu[1]
[1]United Imaging Intelligence, Cambridge MA
[2] University of Southern California, Los Angeles CA
{first.last}@united-imaging.com,yunhaoge@usc.edu,yxiao915@usc.edu,zhix@usc.edu,itti@usc.edu

1. Implementation Details

Network structure The network architectures of Graph Neural Network G and Embedding Network E are shown in Table. 1. G takes n hypotheses $\mathbf{h} = \{h_1, h_2, ...h_n\}$ (each hypothesis h_i is in the form of Structural Concept Graph (SCG)) as input, and output n feature vectors ($G(h_i)$) which concatenate all updated node and edge features of h_i. In G, we use class-specific e^c_{ji} for different hypotheses in each GraphConv layer. E concatenates all n feature vectors from all the hypotheses into a long vector and maps the vector (1 $\times$ (188 $\times$ n)) into n dimensional vector (1 $\times$ n) with a MLP, where n is the number of classes of interest. "node" denotes node feature, "edge" denotes edge feature, "GraphConv" is graph convolutional layer, "ReLU" denotes ReLU activation function, "BN" denotes batch normalization, and "FC" denotes fully connected layer.

Part	Input $\rightarrow$ Output Shape	Layer Information
G	node:(2048 $\rightarrow$ 64); edge:(4 $\rightarrow$ 5)	GraphConv-(e^c_{ji}), ReLU, BN
	node:(64 $\rightarrow$ 32); edge:(5 $\rightarrow$ 5)	GraphConv-(e^c_{ji}), ReLU, BN
	node:(32 $\rightarrow$ 32); edge:(5 $\rightarrow$ 5)	GraphConv-(e^c_{ji}), ReLU, BN
E	(188 $\times$n)$\rightarrow$(n)	FC-(188 $\times$n,n)

Table 1. Network architectures of Graph Neural Network G and Embedding Network E.

Training details In Section 3.3 of the main paper, Eq.2 explains the knowledge distillation we used to imitate the reasoning process of Xception on ImageNet dataset. We train G and E in a end-to-end manner. Below are the details: we use Adam with β_1=0.9 and β_2=0.999, batch size 128, learning rate 0.01 for the first 100 epochs and use a decay rate of 0.5 for the next 200 epochs. For each interested class, we use 400 images to train and other 900 images to test.

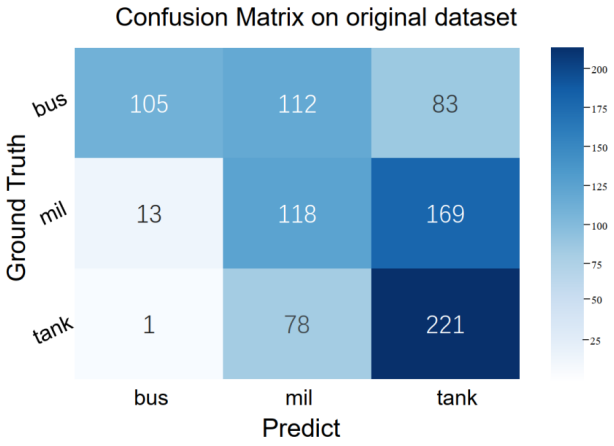

Figure 1. Confusion matrix of 3 class vehicle classification on test dataset with Resnet-18 trained on the original training set.

2. Model Diagnosis with VRX Details

As we mentioned in mian paper Section 4.4, Fig. 1 demonstrates the confusion matrix of 3 class vehicle classification on original dataset with Resnet-18.