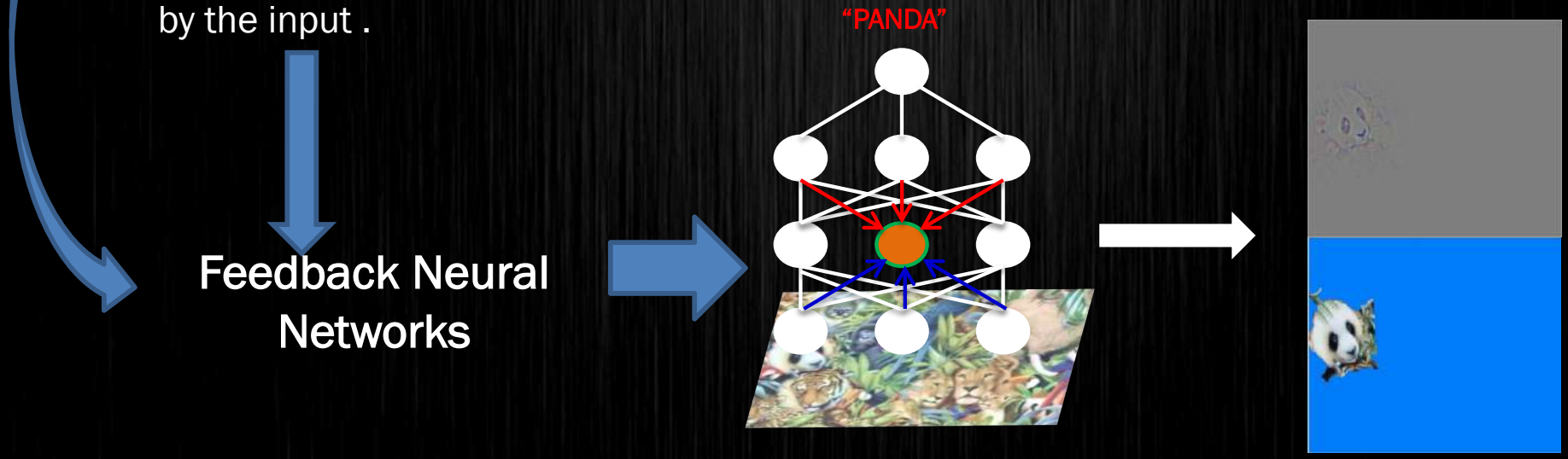
Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks

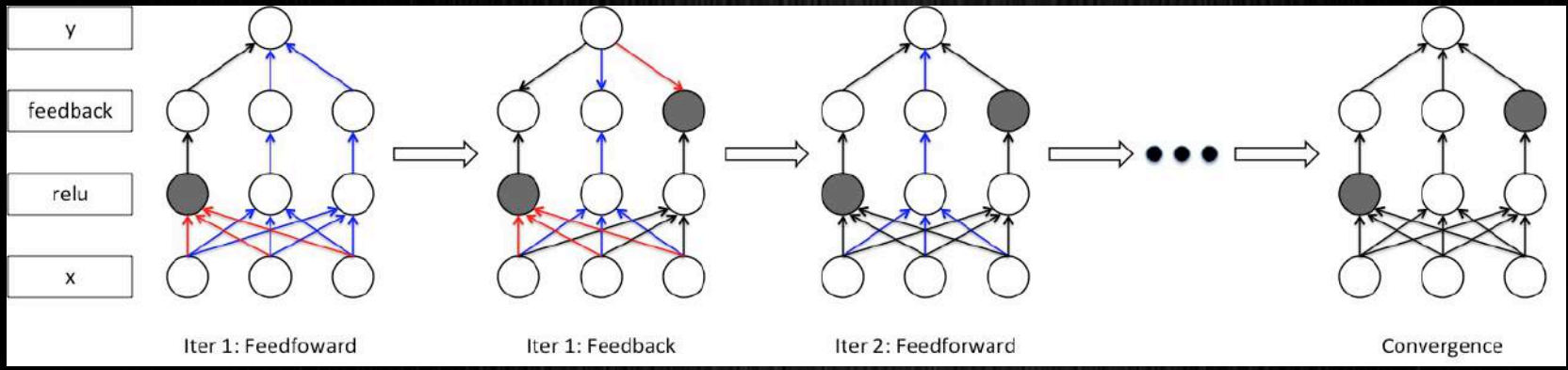
Chunshui Cao , Xianming Liu ,Yi Yang ,
Yinan Yu, Jiang Wang ,Zilei Wang,Yongzhen Huang ,Liang Wang ,
Chang Huang, Wei Xu ,Deva Ramanan ,Thomas S. Huang

Motivation

1. In human's brain, visual attention typically is dominated by "goals" from our mind easily in a top-down manner, especially in the case of object detection or attention. Cognitive science explains this in the "Biased Competition Theory" , that human visual cortex is enhanced by top-down stimuli and non-relevant neurons will be suppressed in feedback loops.
2. The states of relu and max pooling dominate everything. But for most of popular convolutional neural networks, the states of relu and max pooling are determined only by the input .



The iterative process



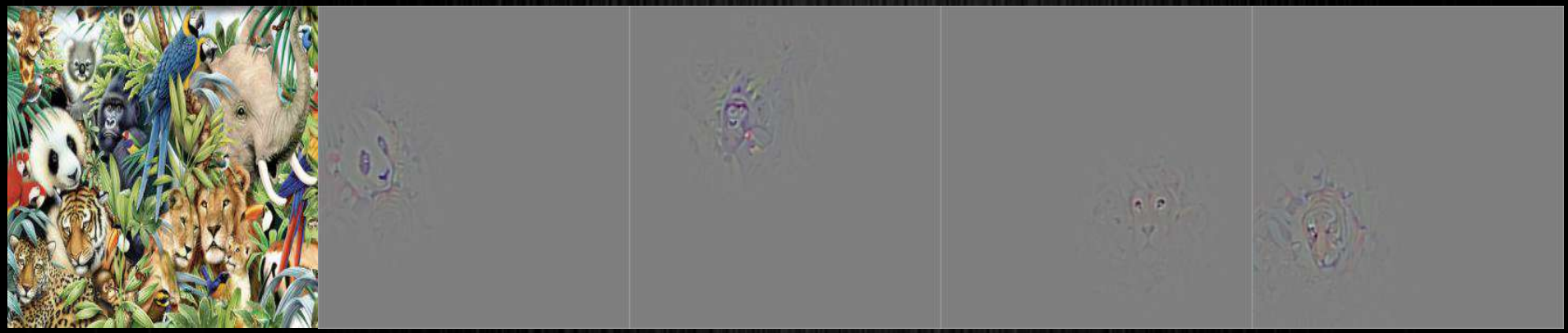
At the first iteration, the model performs as a feedforward neural net.
Then, the neurons in the feedback hidden layers update their activation status to maximize the confidence output of the target top neuron.
This process continues until convergence

The iterative process demo

input

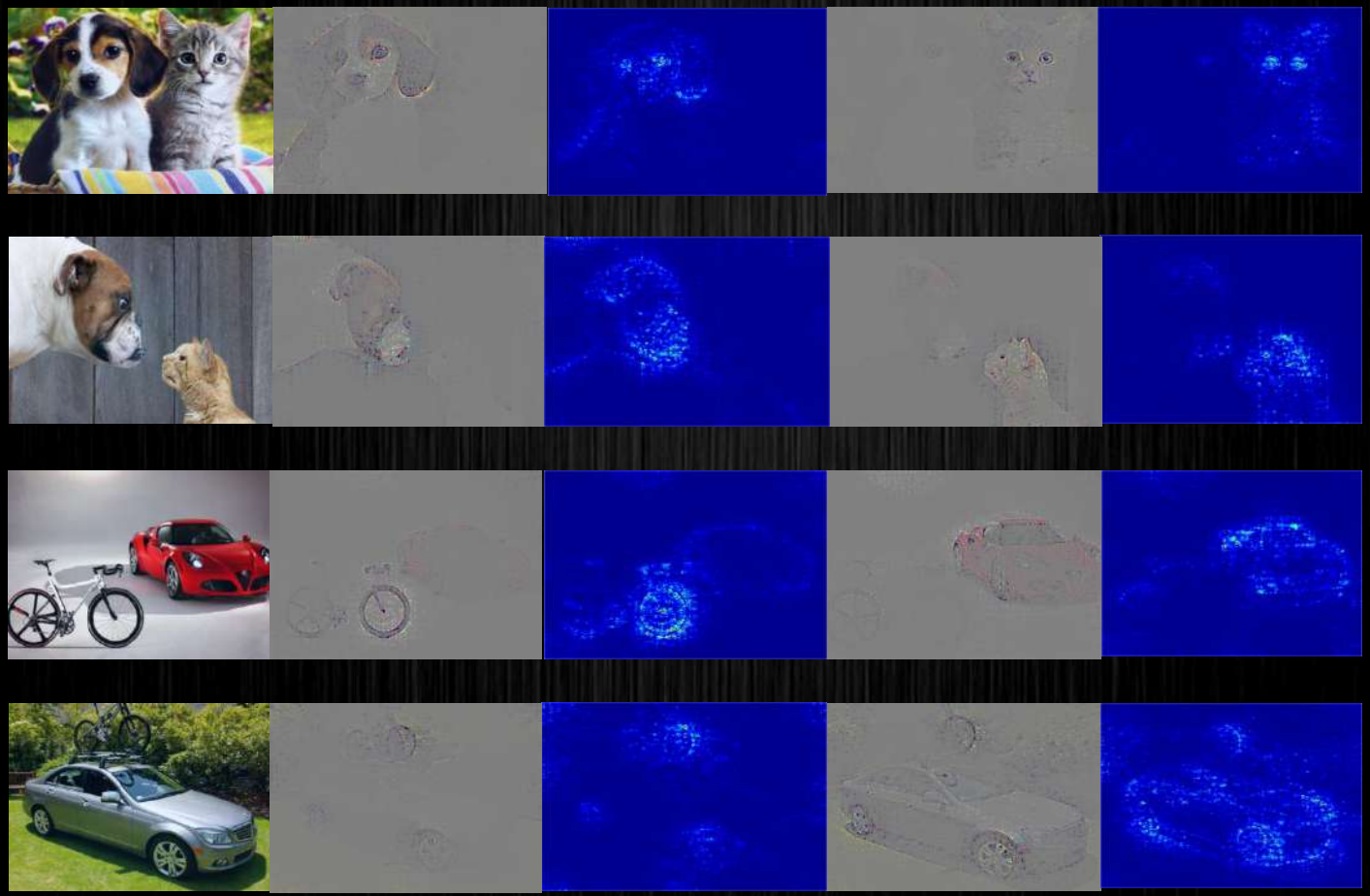
panda

For different targets(googlenet)



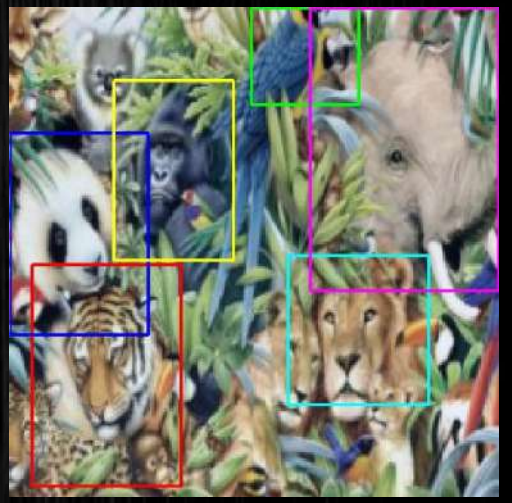
input | panda | gorilla | lion | Tiger

More visualization examples(vgg)



Weakly Supervised Object Localization

Method	Localization Error (%)
Oxford	44.6
Feedback(ours')	38.8

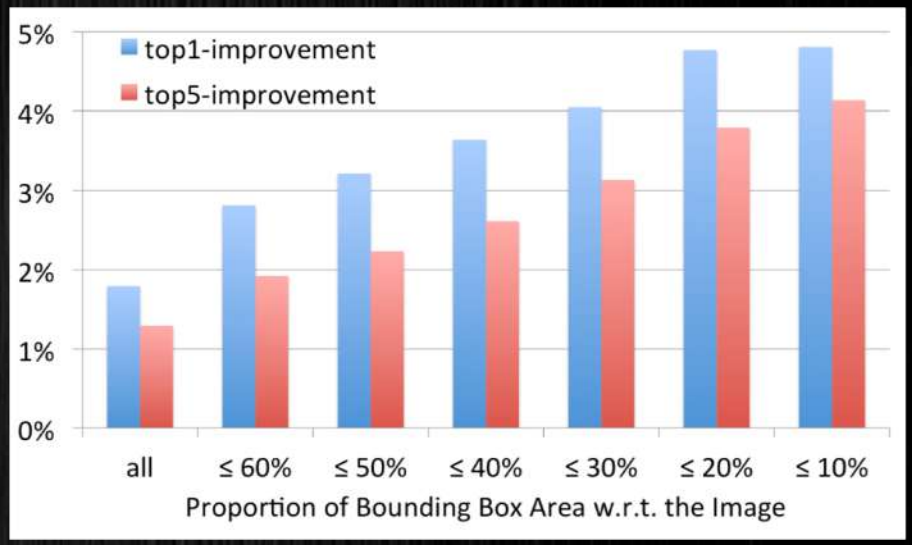


Comparison of our weakly supervised localization results on ImageNet 2014 validation set with the simplified testing protocol: the bounding box is predicted from a single central crop of images and the ground truth labels are provided. We show that our feedback method significant outperforms the baseline method (error rate 44.6%) that uses the original image gradient to localize , both on GoogLeNet architecture.

Image Re-Classification with Attention

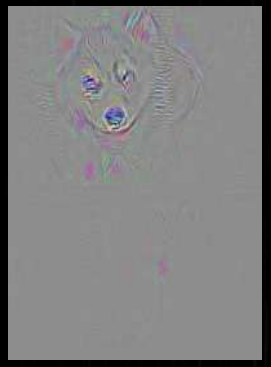
Method	Top 1 (%)	Top 5 (%)
GoogleNet	32.28	11.75
GoogleNet Feedback	30.49	10.46

Classification errors on ImageNet 2014 validation set : the first row is the performance of GoogleNet given a single central crop of images, the second row shows classification results of the same GoogleNet given the attention cropped images.



We divide the ImageNet 2014 validation set based on the proportion of the object size in the image. Classification accruacy using feedback crop for images increases with smaller objects.

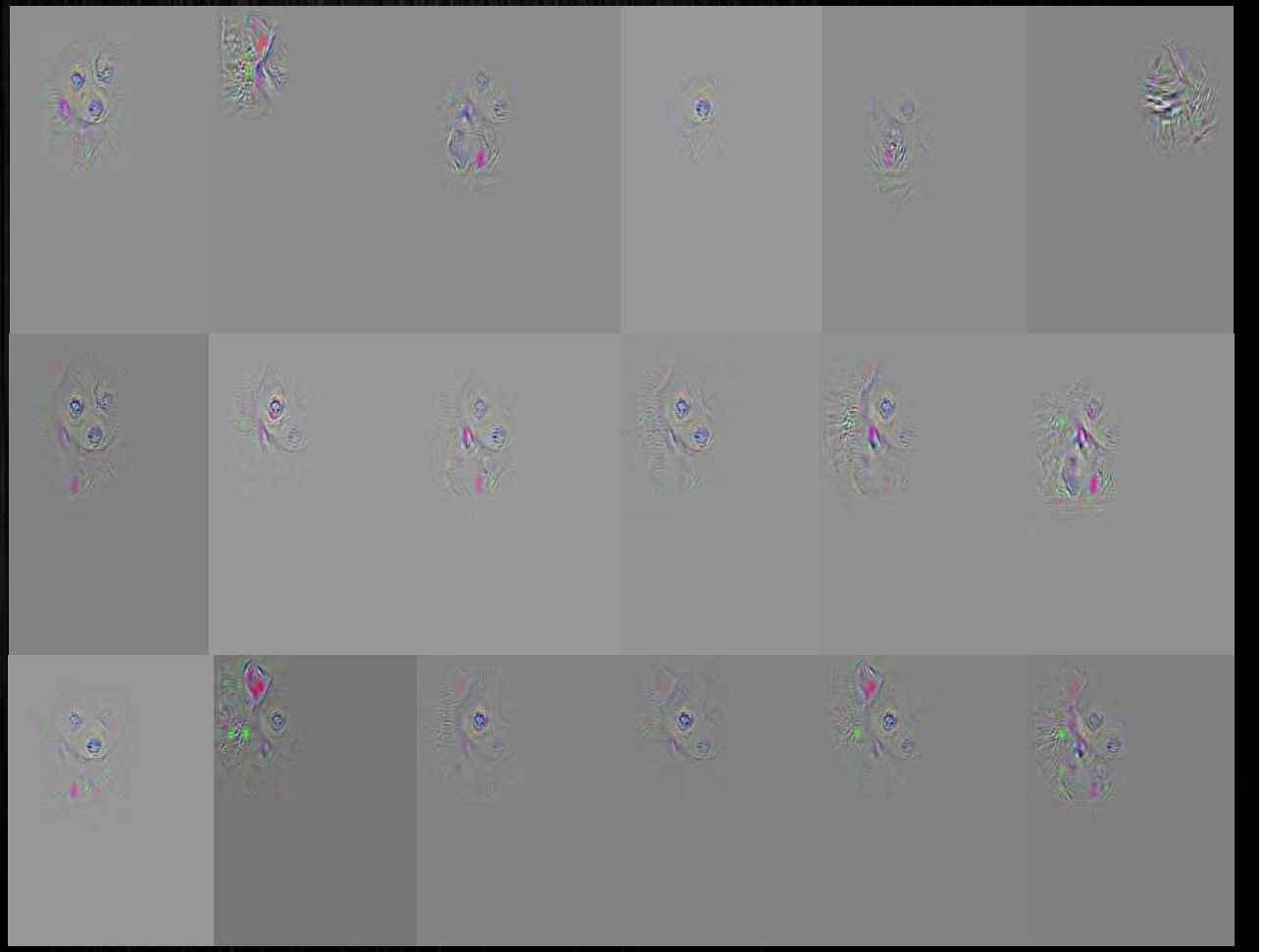
Visualization of feedback neurons(alexnet)



conv3

conv4

conv5



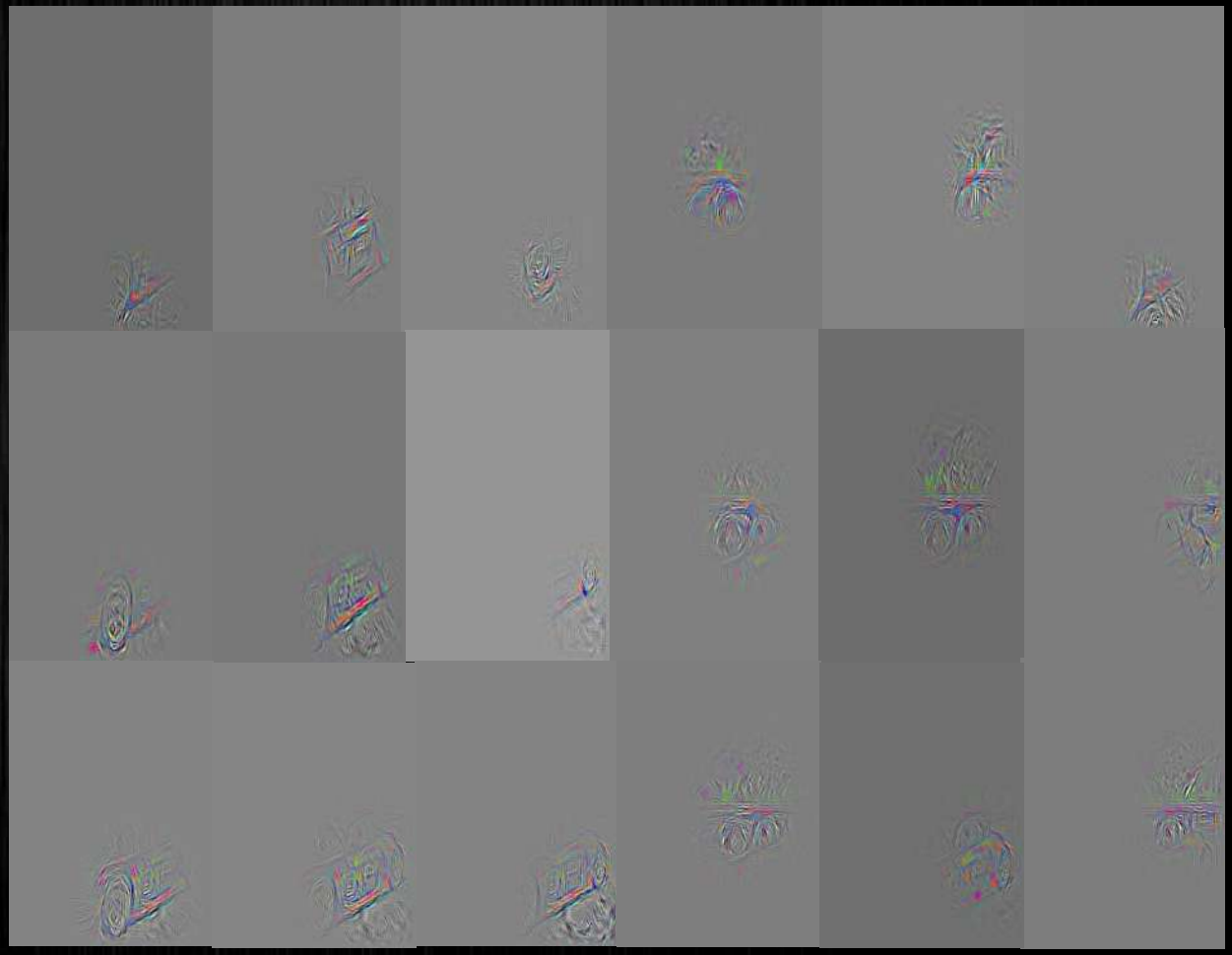
Notes:
1.Feedback for husky.
2.Visualize the remaining neurons of conv3,conv4, conv5. The showing images are corresponding to neurons picked from top30 activations of each layer.

Visualization of feedback neurons(alexnet)

conv3

conv4

conv5



Notes:
1.Feedback for race car.
2.Visualize the remaining neurons of conv3,conv4, conv5. The showing images are corresponding to neurons picked from top30 activations of each layer.