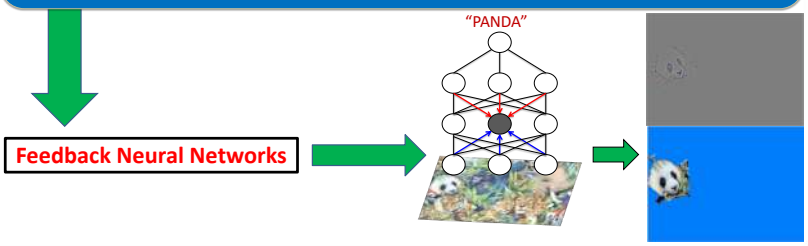
Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks

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

1 Motivation

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.

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 .



2 Principle

We formulate the feedback mechanism as an optimization problem by introducing an addition control gate variable z.

Given an image I and a neural network with learned parameters w , we optimize the target neuron output by jointly inference on binary neuron activations z over all the hidden feedback layers. In particular, if the target neuron is a *k-th* class node in the top layer, we optimize the class score s_k by re-adjusting the neuron activations at every neuron *(i, j)* of channel *c*, on feedback layer l.

$$\max_{\mathbf{z}} \quad s_k(I,\mathbf{z}) - \lambda||\mathbf{z}||$$
$$s.t. \quad z^l_{i,j,c} \in \{0,1\},\ \forall\, l,i,j,c$$

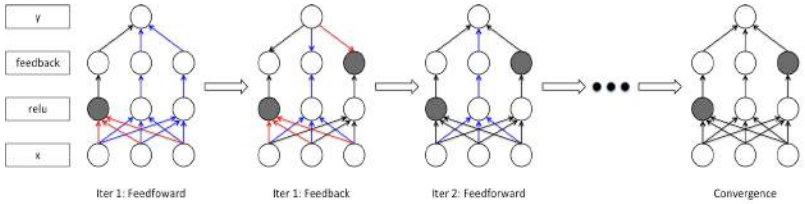
applying a linear relaxation

$$\max_{\mathbf{z}} \quad s_k(I,\mathbf{z}) - \lambda||\mathbf{z}||_1$$
$$s.t. \quad 0 \le z^l_{i,j,c} \le 1,\ \forall\, l,i,j,c$$

Update rule

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \alpha \cdot (\frac{\partial s_k}{\partial \mathbf{z}}|_{\mathbf{z}_t} - \lambda)$$

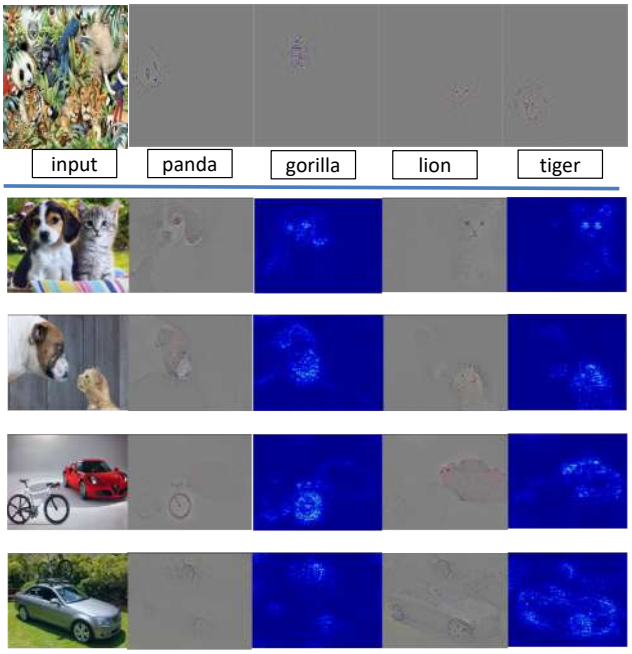
3 The iterative process



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.

4 Experimental Results

➢ **Qualitative results**

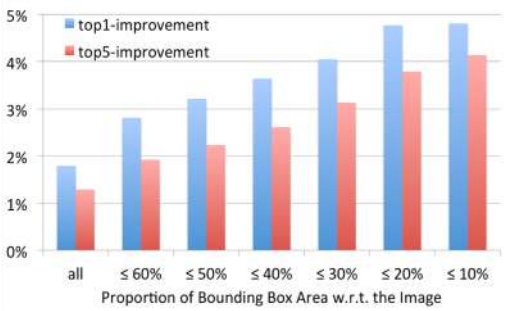


➢ **Weakly Supervised Object Localization**

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

➢ **Image Re-Classification with Attention**

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



5 Conclusions

➢ Achieved the top-down selectivity of neuron activations.
➢ Captured high level semantic by salience maps.
➢ Built a unified deep neural network for both recognition and object localization tasks.