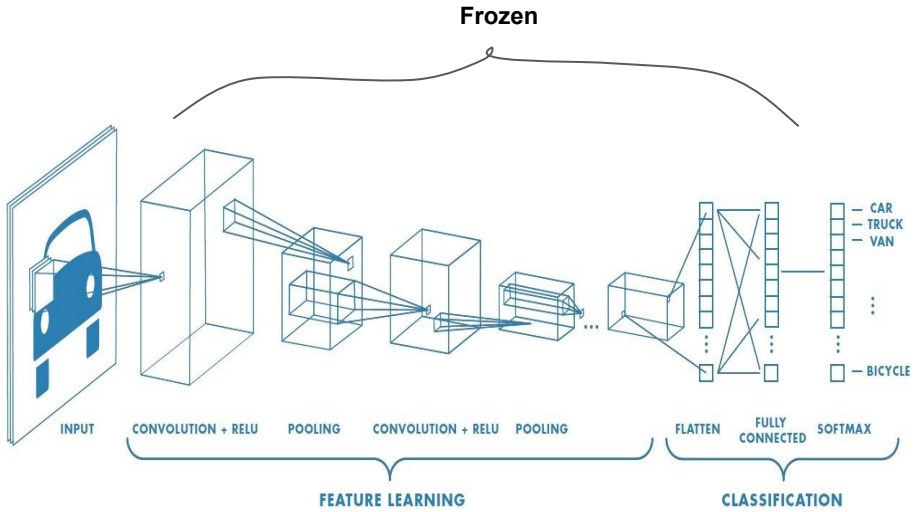
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman
VGG, Oxford

Presenter: Uday Kusupati

Cons

#1: Novelty



- Simple idea
- Already proposed by

Erhan, Dumitru & Bengio, Y & Courville, Aaron & Vincent, Pascal. (2009).
Visualizing Higher-Layer Features of a Deep Network. Technical Report, Univeristé de Montréal

- Input fixed + Modify weights = Training
- Weights fixed + Modify Image = Visualization!

#2: What is Influence?

- No clear definition
- The motivation is from linear models which is a pretty big approximation
- No theoretical justification to the approximation
- Change in score per change in pixel?
- Variance is a better measure?

$$S_c(I) \approx w^T I + b$$

$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0}$$

#3: First order derivative doesn't always work

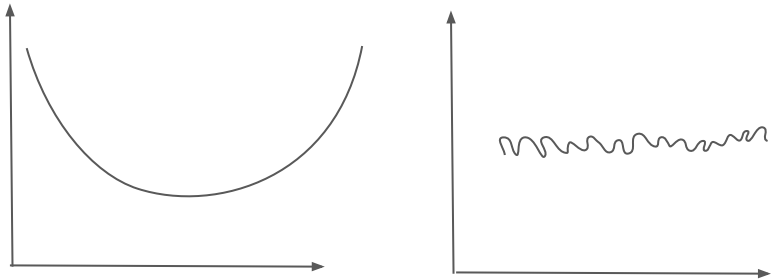
$\frac{\partial S_c}{\partial I} \simeq 0 \Rightarrow$ Very confident classification

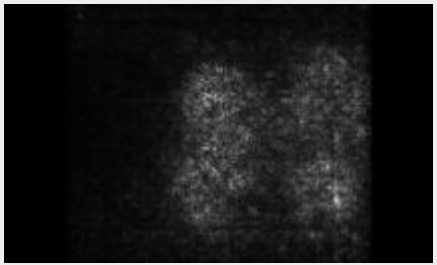
Saliency Map for the visualized image?

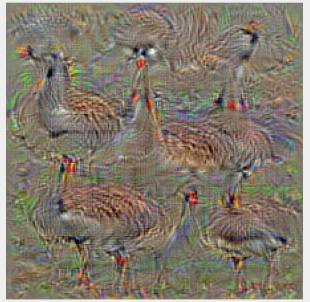
Derivatives close to zero near minima, if the curve is locally convex and well optimized

Fix:

- Higher order derivatives, More terms of Taylor series?
- Variance as influence?







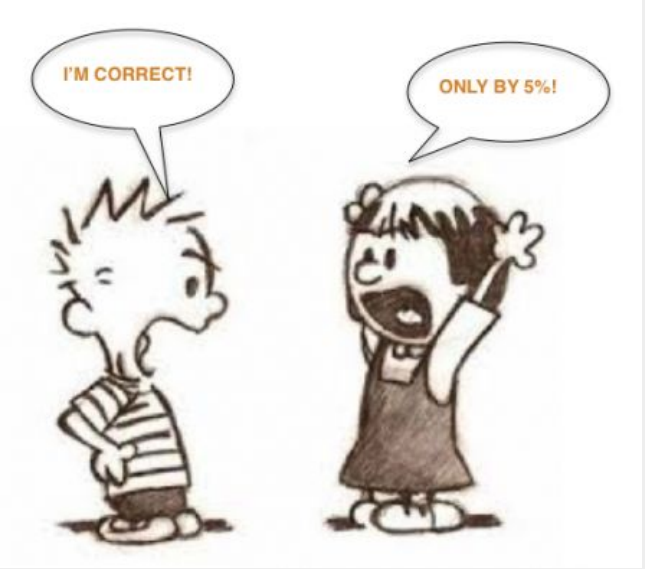
?

#4: Very Few Experiments

- Comparison with results when softmax output is maximised?
- Cross saliency?
- Image editing through back-prop?

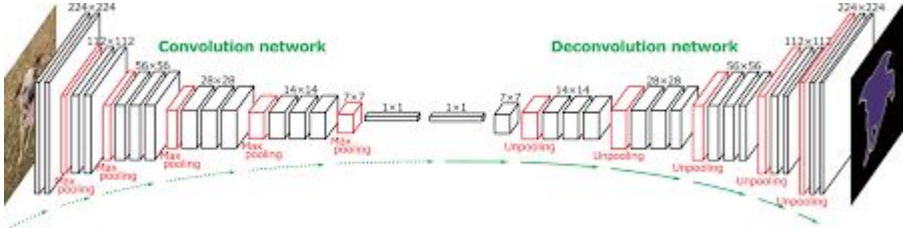
#5: Quantitative results

- Compare scores of real images with the visualized images. How close are they?
- Check if the visualizations make sense to a human sample set.



#6: Ambiguities/Incomplete

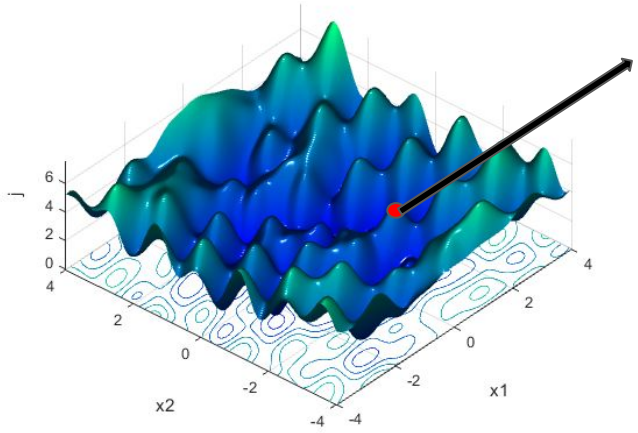
- Deconv and backprop- equivalent or similar? Conclusion?
- Better results than AlexNet?
- Why is the image we construct being regularized?

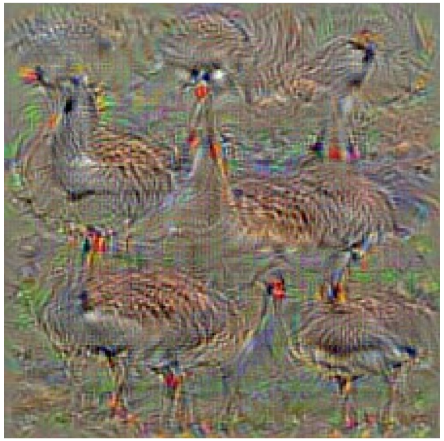


||||

Back Propagation?

#6: Ambiguities/Incomplete





Locally optimal image = Some random goose? or Multiple geese?

Is it because training images consist of multiple geese?

More discussion on the discriminative nature of the model