National University of Singapore
School of Computing
CS3244: Machine Learning
Tutorial 09

RNN and Explainable AI(XAI)

1. **RNN and BPTT** Here, we'll be computing gradients via *Backpropagation Through Time* (BPTT). The Forward Pass of a RNN can be characterised as follows(Here σ denotes *softmax* function):

$$\mathbf{h}_t = g^{[h]}((\mathbf{W}^{[xh]})^\top \mathbf{x}_t + (\mathbf{W}^{[hh]})^\top \mathbf{h}_{t-1}) \tag{1}$$
$$\hat{\mathbf{y}}_t = g^{[y]}((\mathbf{W}^{[hy]})^\top \mathbf{h}_t) \tag{2}$$
$$\hat{\mathbf{o}}_t = \sigma(\hat{\mathbf{y}}_t) \tag{3}$$

   The loss L is the Cross Entropy Loss:

$$L = -\sum_t^T \mathbf{y}_t \cdot \log(\hat{\mathbf{o}}_t) \tag{4}$$

   For simplicity, let's call the final time step loss, $E_T = -\mathbf{y}_T \log(\hat{\mathbf{o}}_T)$. The objective of BPTT is to update the parameters $\mathbf{W}^{[xh]}$, $\mathbf{W}^{[hh]}$, and $\mathbf{W}^{[hy]}$.

   (a) Use Chain Rule to find an expression for $\frac{\partial E_T}{\partial \mathbf{W}^{[hh]}}$. (Note, there is no need to expand the term $\frac{\partial \mathbf{h}_{T-1}}{\partial \mathbf{W}^{[hh]}}$ further)

   (b) Out of the various terms above, which one do you think is responsible for the *Vanishing Gradient Problem*?

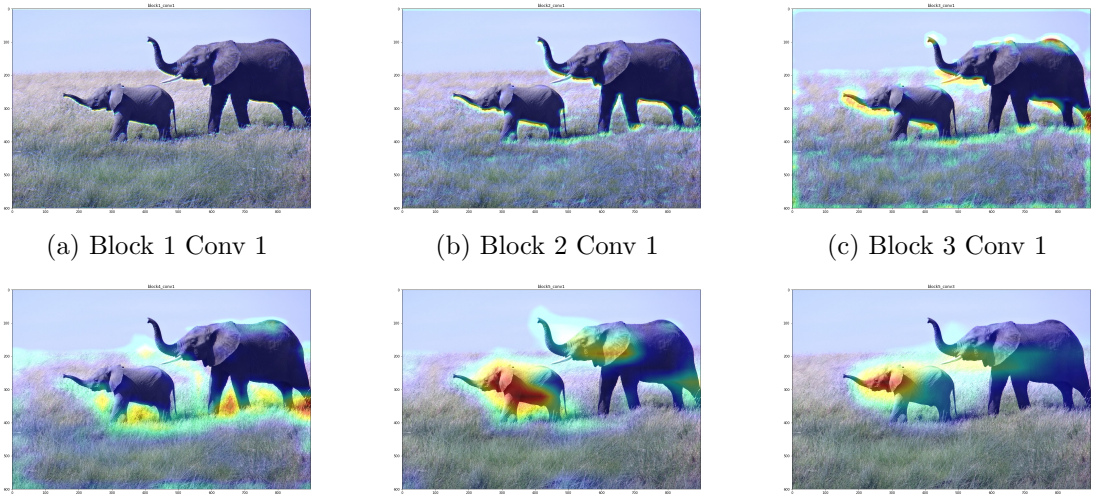
2. **LIME**

   (a) In LIME, we use a predictor which is a quadratic function rather than a line. How do you sample points for the decision boundary? Show how quadratic LIME can improve over a linear LIME.

   (b) What are the disadvantages of using LIME for explanation?

3. **XAI** Fig 2 illustrates CAM (Class Activation Mapping) / the heatmaps of various convolutional layers of the input image shown in Fig 1. Here, we use a large neural network, VGG16 which has 5 blocks of convolutions and each block has 3 convolutional layers. VGG16 is used to classify the images into number of classes. Elephant is one of those classes.

Figure 1: Input Image



(a) Block 1 Conv 1 (b) Block 2 Conv 1 (c) Block 3 Conv 1

(d) Block 4 Conv 1 (e) Block 5 Conv 1 (f) Block 5 Conv3

Figure 2: Heatmap through various Convolutional layers of VGG16 - Image Credit

(a) Comment on the heatmaps with respect to the image classification.

(b) In each layer, different segment of the image is activated. Why those activations are different? You can assume that *Block5 Conv3* is the decision making layer.