Signal processing based Interpretability of Machine Learning Models

Indian Institute of Technology Jodhpur

Report on LRP implementation

- LRP
- VGG-16 model
- Code blocks
- Results
- References

LRP

Layer-wise Relevance Propagation (LRP) is one of the most prominent methods in explainable machine learning, if our network predicts a class for the input image, then the explanation given by LRP would be a map of which pixels in the original image contribute to that class and to what extent. This method does not interact with the training of the network, so you can easily apply it on already trained classifiers.

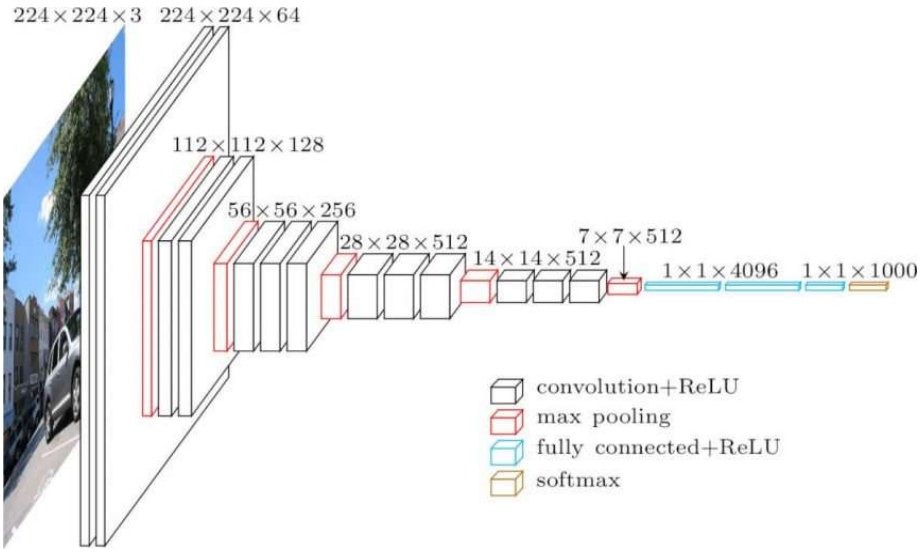
Intuitively, what LRP does is that it uses the network weights and the neural activations created by the forward-pass to propagate the output back through the network up until the input layer. There, we can visualize which pixels really contributed to the output. We call the magnitude of the contribution of each pixel or intermediate neuron "relevance" values *R.* In the output layer, we pick one neuron, or class, that we want to explain. For this neuron, the relevance is equal to its activation, the relevance of all other neurons in the output layer is zero. For example, if we want to use LRP to find out the relevance of the network's neurons and inputs with respect to predicting class c, we start with the output neuron for class c and only look at how the network arrived at this neuron's values. From there on we go backwards through the network by following this basic LPR rule:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

It is important to note that there are different propagation rules for LRP and that you will probably want to combine several of them to get the best results.

VGG-16 MODEL

LRP rules could be easily expressed in terms of matrix-vector operations. In practice, state-of-the-art neural networks such as VGG-16 make use of more complex layers such as convolutions and pooling. In this case, LRP rules are more conveniently implemented by casting the operations of the four-step procedure. We take the VGG-16 trained on ImageNet data for image classification. ImageNet is a dataset of over 15 million labeled high-resolution images belonging to roughly 22,000 categories.



The input to the cov1 layer is of fixed size 224 x 224 RGB image.

Code Block

Code is divided into 2 main parts one for the calculation of results and other for having the important utilities for the code.

Utilities for the code :

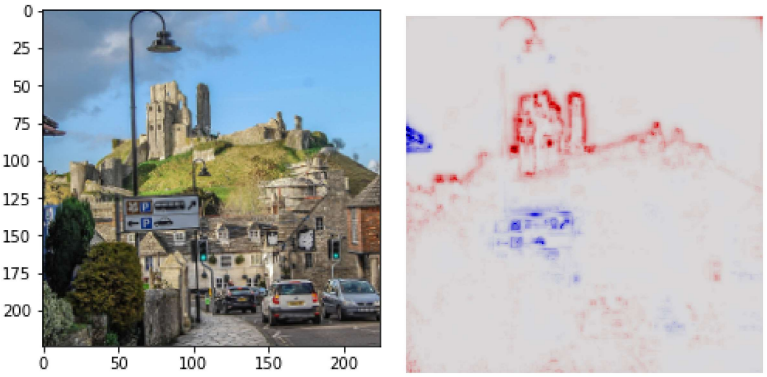
- Def toConv : Function for converting the last dense layers into convolutional layers.
- Def newLayer : Clone a layer and pass its functions.
- Def heatmap, digit, image : Visualizing the data.
- Def loadparams, load data : For loading the parameters and the data.
- Imgclasses : Dictionary of all the class names.

Main code : The image is inputted with the help of cv2 and the array formed is converted to 224x224x3 for inputting into the vgg16 model. The image is normalized and fed into the pretrained model. The input can then be propagated in the network and the activations at each layer are collected, activations in the top layer are the scores the neural network predicts for each class.

This code iterates from the top layer to the first layer and applies propagation rules at each layer. Top-layer activations are first multiplied by the mask to retain only the predicted evidence for the class. This evidence can then be propagated backward in the network by applying propagation rules at each layer (treating max-pooling layers as average pooling layers in the backward pass). As each layer is composed of a collection of two-dimensional feature maps, relevance scores at each layer can be visualized as a two-dimensional map

Results

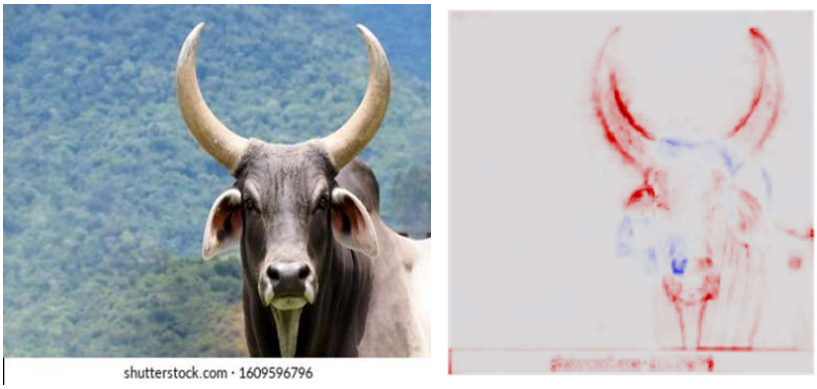
The following results were obtained on different images:



Prediction : “castle”
The outer boundary of the caste is red that means that these pixels have the highest relevance among all. The blue colored section had the negative or no relevance.

Prediction : “Piggy bank”
The outer boundary of the pig’s face are highlighted and the slot for coin is highlighted distinguishing it from predicting real pigs.

Prediction : “golf ball”
The dotted pattern on the golf ball is highlighted, distinguishing it from the 2nd most favourable ping pong ball.



Prediction : “Ox”
Horns, shape of nose and plated skin on the neck region have the most relevance, hence highlighted the most.

Prediction : “jersey, T-shirt, tee”
The model is highlighting the human face features but the prediction is T-shirt since human face or person is not an output class for the model.

References

- https://machinelearningmastery.com/use-pre-trained-vgg-model-classify-objects-photographs/
- https://git.tu-berlin.de/gmontavon/lrp-tutorial
- http://heatmapping.org/

- http://iphome.hhi.de/samek/pdf/MonXAI19.pdf
- https://towardsdatascience.com/indepth-layer-wise-relevance-propagation-340f95deb1ea