Visualising Deep Neural Networks

Project report for WiDS Program, Analytics Club, Dec-Jan'21

Shivam Patel
Dept. of Electrical Engineering
IIT Bombay
200070077@iitb.ac.in

***Abstract*—The goal of this guided project is to learn and explore the inner functionings of Deep Neural Networks. Different mechanisms and algorithms such as Occlusion Sensitivity Maps, Saliency Maps, Class Activation Maps(CAM), Grad-CAM are utilised for this purpose. Deeper analysis of neural networks and their functioning can help us pinpoint the classes/labels or image specific attributes that can cause our neural networks to give unwanted and undesired class labels and outputs. This project will take you through the basic understanding of above mentioned mechanisms, and enable you to perform well defined error analysis and take informed decisions on changing/tuning your deep neural networks.**

Keywords - Occlusion Sensitivity Maps, Saliency Maps, Class Activation Map

I. INTRODUCTION

Deep neural networks are the leading Machine Learning and Artificial Intelligence frameworks. Simplicity of single node neural network is well understood and mathematically defined. But after scaling the neural networks to multiple layers, and many different type of operations such as convolutions, transpose convolutions etc. makes it difficult and often counter-intuitive to understand and comprehend. Mankind has achieved accuracies of 97%, when it comes to image classification, speech recognition etc. But the 3% error in them makes them unsuitable for sensitive tasks, such as utilisation in military, banking etc.

Neural Networks are improving day-by-day, in terms of size, accuracy and other target parameters. One important aspect of improving deep neural networks is manually assessing them and trying to find errors in the framework, and work on them to improve the overall model accuracy. But when it comes to image classification, manual examination of trainable parameters is very difficult, with millions of learned weights and biases.

We overcome this question by using 4 main analysis pipelines-

1) Saliency Maps
2) Occlusion Sensitivity Maps
3) Class Activation Maps
4) Grad CAM - Gradient Weighted Class Activation Mapping

There is a vast amount of labelled image data, and there are multiple models trained on each of the datasets. It takes considerable amount of time and computational resources to carry out training on image data, especially with millions of labelled images as training set. Thus, our effort will be directed towards creating analyses pipelines and methodologies rather than training such models.

II. PRIOR WORK

Given the necessity and requirement of creating an error analysis algorithm for image classification and CNNs, many university research groups and individuals have done immense research with remarkable results in this direction:

1) **Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps** - This paper addresses the visualisation of image classification models, learnt using deep Convolutional Networks (ConvNets). There are two visualisation techniques in consideration, based on computing the gradient of the class score with respect to the input image. The first one generates an image map which maximises the class score, and the second one creates a Saliency map, specific to image and ground truth class label.
2) **Visualizing and Understanding Convolutional Networks:** This paper introduces a novel visualisation technique that gives insight into the function of intermediate feature layers and the operation of the classifier. Such a visualising scheme helps us to create a model which can possible beat the maximum ImageNet classification accuracy score.
3) **Learning Deep Features for Discriminative Localization:** This research paper focuses on how we can reverse interpret MAX-POOL layers, and how does that lead to loss of image data and we need gradient ascent to recover/reach the best possible approximation to the gradient descent method.
4) **Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization:** This paper aims at making visual explanations to CNNs more transparent and human-friendly. The approach - Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting important regions in the image for predicting the concept.

III. CONVNETS

Convolutional neural networks are just like any other neural networks, but with the added functionality of convolutional filters that extract features and reduce dimensions of the image and make it easier to train a model on. There are different types of filters which can have preset weights and parameters, while others can be trained to fit the task as can be observed in Figure 1 and 2.

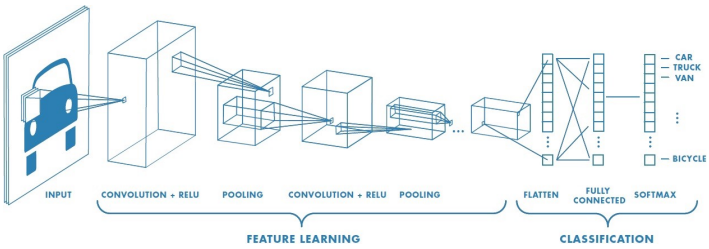

Fig. 1. Basic Convolutional Neural Networks

Stucture:

There are three main parts of any CNN framework-

- **Convolutional Layer-** These layers form the basis for the CNN structure, and this is where trainable filters are applied, learnt and used. This layer gives CNN their name
- **Max Pooling-** These layers are used to reduce dimensions of images in the networks. They are of two types- max and average pool, and as the name suggests, they pool over the maximum value in the filter size and the average value over a filter size.
- **Fully Connected(FC)**- These layers are generally implemented by using a 1x1 convolution, which individually connects each cell in a convolutional network to the further layers, acting like a conventional neural network instead of CNN.

Parameter Sharing and Sparse Connections make CNNs the appropriate choice for large dimensional data such as high resolution images over conventional deep neural networks.

IV. BOTTLENECKS IN CNN VISUALISATIONS

CNNs, unlike simple neural networks, can't be traced back to the original inputs using the outputs and trained weights. Due to layers such as convolutions, max pools etc. a lot of original data and input parameters gets lost, and can't be traced back fully. Some of the major bottlenecks in re-tracing outputs to input is -

- **Non-Bijective Activation Functions -** Activation functions such as ReLU, Binary-Step are non-bijective, and thus the input cannot be predicted from their outputs, as their inverse functions don't exist. There are multiple ways of overcoming this, such as making experimentally verified assumptions and following intuitive reverse operations of the activation functions.
- **Convolutional Filters -** Convolution in image processing is an irreversible process as the total dimensions in input and output tensors(/matrices) differ, and hence, there does not exist a non-singular inverse operation to reverse the convolution. This is overcome by using padding(one out of many other reasons) in convolutional networks, and hence we can then reverse the output, as we are concerned only with the non-padded input, which is of smaller dimensions than the zero-padded input.
- **MAX/AVG Pool Layers -** Pooling layers are explicitly used for dimensionality reduction, and depending on the size of the pool filter and its stride, the reduction can be 4-10 fold. A very naive, but efficient way of overcoming this is storing the keys of the image pixel in a new tensor(called Switch Matrix), where the key value is 1 for the pixel if it was the maximum in the filter, and 0 otherwise. This gives us the locations of the pixels which were fed into the layers after pooling, and we can trace their value back from the output layer. As far as the other pixels are concerned, we fill in zero value or the image average value to estimate them.
- **Subsampling -** As the name suggests, subsampling is a method of dimensionality reduction, where a sub sample of image is chosen for feeding into further networks. This leads to data loss, and no approximations can be made for this method. Hence, it is generally avoided in modern day CNNs, as we have more computation power and other better methods to reduce dimensions.

V. ANALYSIS PIPELINE

A. *Saliency Maps*

Saliency maps are used to represent the parts/sections of images which human eye first focuses on when it looks at any given image. This represents the areas of an image which a biological eye uses to classify an image. These areas represent the most important sub-parts of an image. Saliency maps generated by using computer vision and neural networks don't always replicate human saliency maps.

The major parameter used to create a saliency map is the pixel brightness and contrast. In most of the cases, the object to be detected has different contrast distribution as compared to its surroundings. If the contrast is similar, then object detection becomes tough even for humans, such as in case of natural camouflage. Discriminative image and video resolution exploits this fact, and compresses visual data by reducing resolution of image parts far away from high saliency regions.

B. *Occlusion Sensitivity Maps*

Occlusion Sensitivity maps are generated from the image itself, without any backprop algorithm or approximations. The image is divided into subparts, depending upon the fineness and resolution of the OS-Map we desire. Then, a particular region of the image is covered/gray-ed by a filter of desired size, and then image is given as input to the pre-trained model. The final probability of classification for the ground truth label corresponding to the Soft-Max output is noted down for all such filter positions. The regions where the filter produces

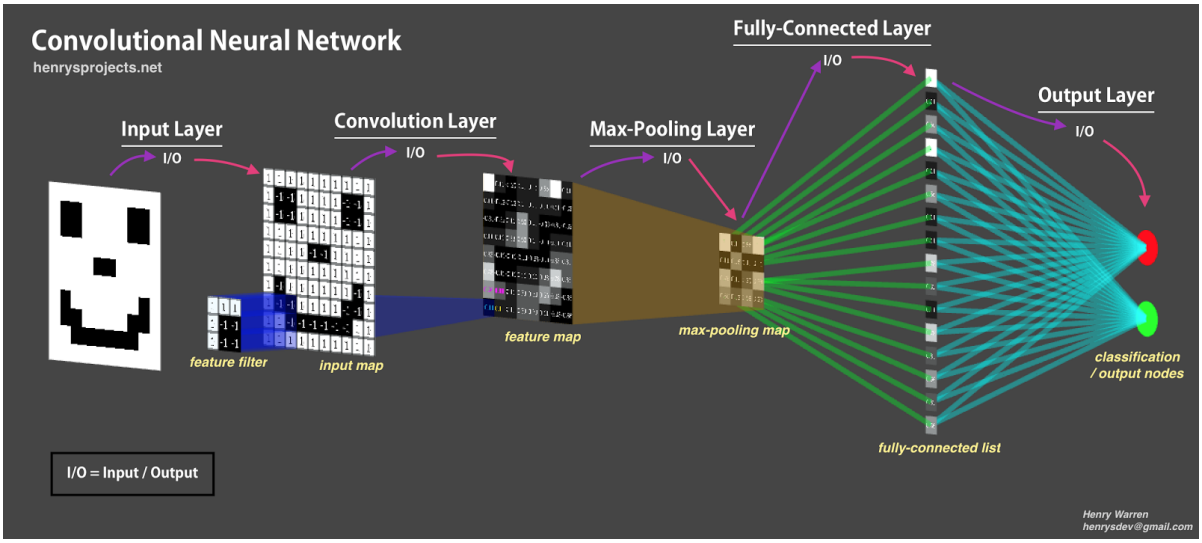

Fig. 2. CNN with major layers

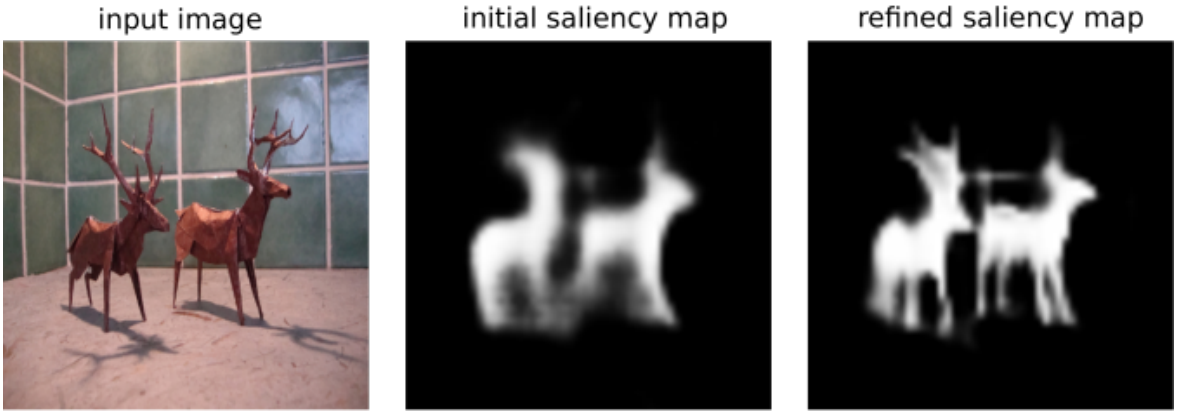

Fig. 3. Computer generated Saliency Maps

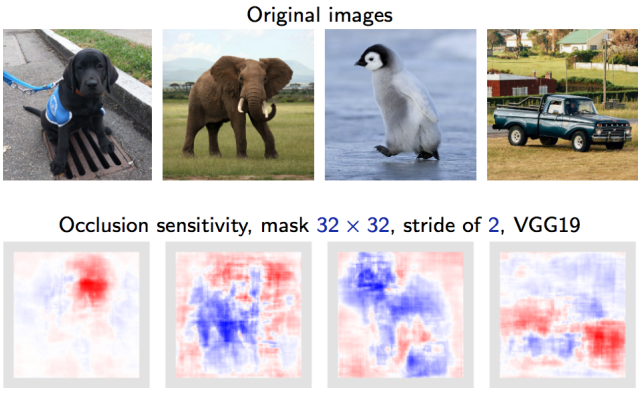

Fig. 4. Occlusion Sensitivity Maps

maximum deviation from the natural/unfiltered output is given a higher intensity in the OS-Map and vice-versa.

The resulting Occlusion Sensitivity map is then plotted, which gives us the relative importance of different pixel values in determining the ground truth labels in any pre-trained neural network, as can be seen in Figures 3 and 4.

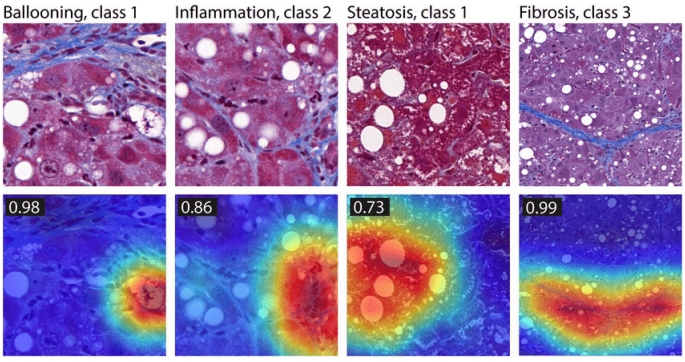

Fig. 5. CAM used in Medical Image Processing and Diagnosis

C. *Class Activation Maps*

The Class Activation Maps are methods used for determining the highest activated layers of a CNN, which are

leading to maximum probability of classification across the softmax layer. They can be comprehended through various model parameters, such as weights, relative weights, and most importantly gradients. CAM based CNN debugging architectures are the most widely used frameworks for Deep Networks advancement. Regionally localised weights and gradients enable the algorithm to pin point the exact location of the activated classes, along with their probability densities/intensities. They are mostly used in models where Global Average Pooling layer is present at the end of the sequential CNN model.

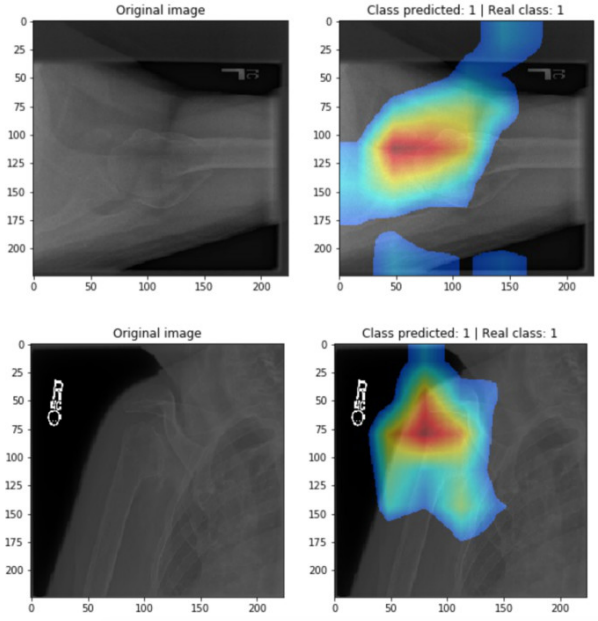

Fig. 6. CAM used in Orthopaedic Diagnosis

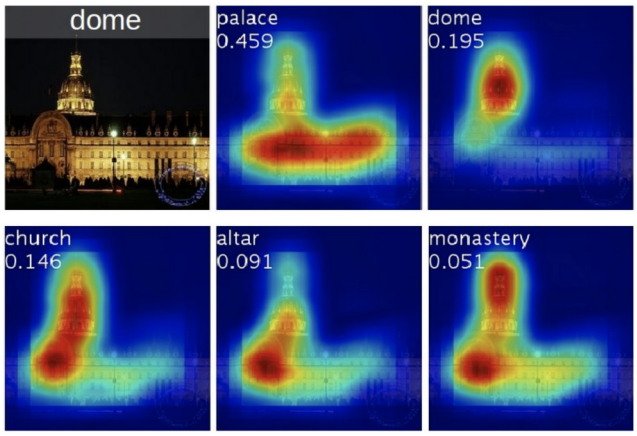

Fig. 7. CAM for Architectural Classification

D. *GRAD CAM*

The Gradient Weighted Class Activation Mapping is an Improvement to the already existing CAM maps. The biggest setback of using CAM maps was that they could be applied only to CNN architectures having Global Average Pooling (GAP). In the case of GAP in the last layer of CNN, it is noted that the last layer receives the same gradient as does the previous layer before GAP. This can be extended to other CNN architectures by assuming that the average gradient received by the last layer of the feature map can be used as the corresponding weight for defining the Class Activation Map for non GAP architectures.

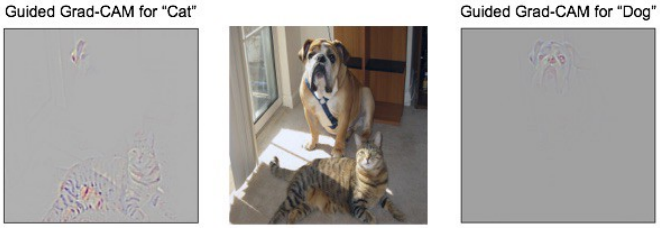

Fig. 8. GRAD CAM

VI. Conclusions

We have seen various visualisation techniques, with different attributes and varying uses. Each algorithm differs in their complexity, preferred feature of utilisation and the mathematical underlyings. We saw that Saliency Maps are one of the earliest visualisation techniques, which draw motivation from natural biological systems of simpler complexities. We then explored Occlusion Sensitivity Maps, which are the most intuitive and easiest to implement visualisation techniques. Later, with advancements in Deep Learning and Artificial Intelligence, researchers found a more effective way to channelise trained model parameters and gradients to effectively visualise model functionings.

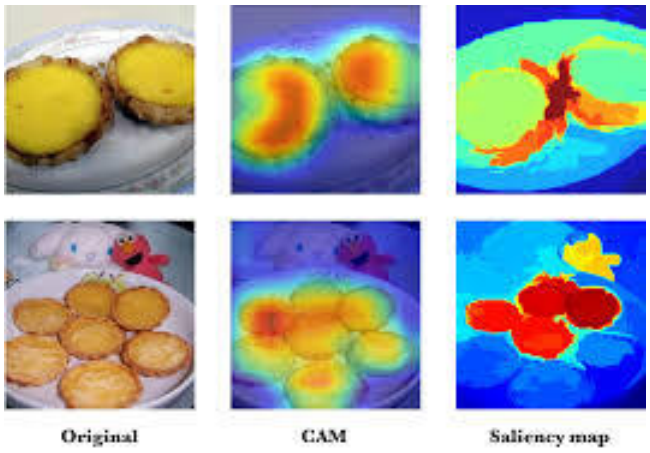

Fig. 9. CAM vs Saliency Maps

Models can be applied to the same images side-by-side to get a deeper understanding about their workings and uses, as in Figure 9.

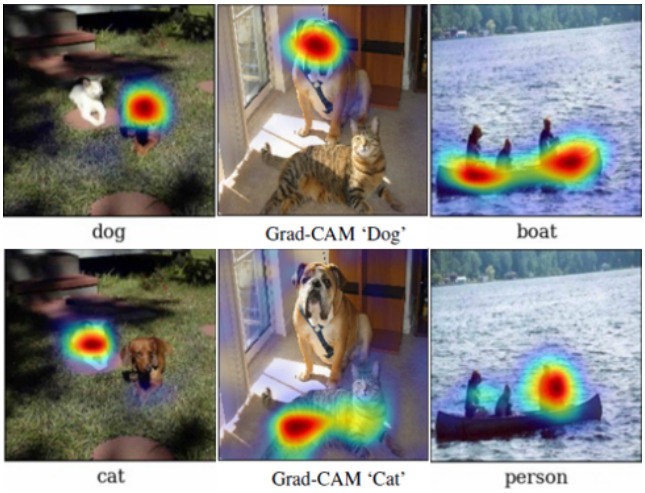

Fig. 10. Multi Class CAM

VII. ACKNOWLEDGEMENTS

I would like to thank Saikiran Akkapaka for his support and guidance throughout the project. I would also like to extend my gratitude towards Analytics Club @IITB, which made such a project possible, under the guidance and expertise of our seniors.

VIII. APPENDIX

A. *Deep Dream by Google AI*

Deep Dream by Google AI Team is a novelty analysis model, which is used to extrapolate the working of Grad-CAM model. Gradient Ascent checking is disabled while reversing the neural network functioning. This results in psychedelic looking images being outputed by the model. This is very much similar to humans looking at figures or real life objects being formed in clouds.

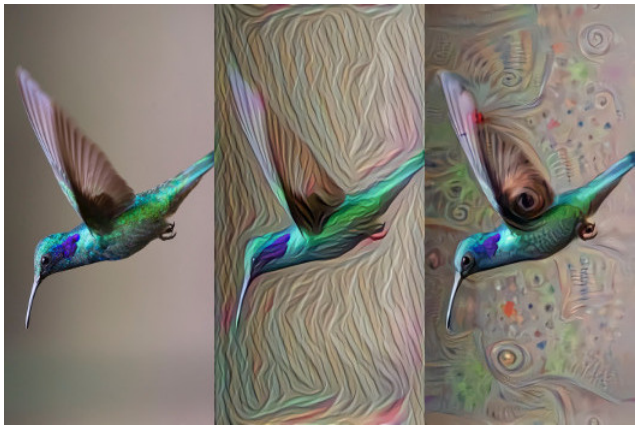

Fig. 11. Deep Dream Hummingbird

REFERENCES

All the references to papers and images used in this project are mentioned as text links at respective places. The research papers referred for this project are -

[1] Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman: https://arxiv.org/abs/1312.6034
[2] Visualizing and Understanding Convolutional Networks, Matthew D Zeiler, Rob Fergus: https://arxiv.org/abs/1311.2901
[3] Learning Deep Features for Discriminative Localization: https://arxiv.org/abs/1512.04150
[4] Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization: https://arxiv.org/abs/1610.02391

Fig. 12. Deep Dream Polar Bear

Fig. 13. One more Deep Dream Artwork