Interpretability of Deep Neural Networks in Healthcare

Arshdeep Singh
IIT Delhi
cs5160625@iitd.ac.in

Aditya Jain
IIT Delhi
cs1160335@iitd.ac.in

1. Introduction

Artificial Intelligence, especially deep learning has made major breakthrough in solving various complex tasks such as computer vision, speech recognition and natural language understanding. Despite widespread success they are still considered as a "black box" solution. Better predictive power comes with stacking more layers therefore making them harder to interpret. In a sensitive area like health-care where any decision comes with a huge responsibility, qualitative and quantitative evaluation of how decisions are made come with a huge responsibility. Interpretability is not only essential to evaluate decisions but also to improve the model where it might be making errors, thereby showing where it fails, making more robust by uncovering the corner cases.

Usually the predictions of a deep learning model are quantified with a metric. However, notion of trust also depends on the visibility a human has into the working of the machine. In other words, if neural networks should be able to provide human interpretable justifications for its output leading to insights about the inner workings. In this work, we try to interpret a neural network by visualising which part of the input image made the network make a prediction by identifying patches in an image.

2. Related Work

Various kinds of work has been done in understanding how a neural network makes decision. Some profound ways are by visualizing the feature maps of the network or by using adversarial attacks. For visualization deconvnet[5], guided backpropagation were proposed early to find the part of the images which impact the decision most . In our method we will be using Grad Cam [6] along with the before mentioned methods to understand the model. Some other methods like attention in the network are also popular to observe the behaviour of the network. Attention [7] after being successful in NLP , is also found to be useful in the computer vision tasks. Residual attention networks [8]for the image classification is one such network architecture.
Using these methods to explain neural networks in healthcare has been tried by many different researchers in different problem domains. One such example is in interpretability in Alzheimer's disease [3] detection using MRI.

3. Proposed Method

We will be using the diabetic retinopathy eye image dataset for our experiments. We create an neural image classifier for the different stages of retina in a diabetic patient and then try to interpret the decisions made by the neural network. So for this task we will be using two classifier , first one being a Resnet [4] model and other one an residual attention network [8] . The reason for choosing these network is to see if the network with attention module installed is capable of showing some extra insights about the classification.
After classification we use GradCam , guided backpropagation , guided gradcam and the attention layer feature maps for visualizing the results. Then finally we compare the results and see what interpretation can be concluded and how the visualization differ for different model.

We also repeat our experiments on another dataset for Histopathologic Cancer Detection, using Residual Attention Network and then using GradCam to visualise features from the last attention module which is discusses later in the model architecture.

4. Model Architecture

4.1. Residual Attention Network

We used a residual attention network, which is a convolutional neural network using attention network which is built by stacking Attention Modules which generate attention aware features. Residual attention network is constructed by stacking multiple attention modules. Each attention module has two parts: mask branch and trunk branch. Trunk branch performs input feature preprocessing. Given trunk branch output $T(x)$ with input x, the mask branch uses bottom up, top down structure to learn the same size mask $M(x$ that softly weighs output features $T(x)$. This mimics the fast feedforward and feedback attention process. The output mask is used as control gates for neurons of the trunk branch similar to Highway Network. The output of

the Attention Network Module H

$$H_{i,c}(x) = (1 + M_{i,c}(x)) * T_{i,c}(x)$$

where i ranges over all spatial positions and $c \in \{1, ..., C\}$ is the index of the channel.

5. Pre-processing

The diabetic retinopathy dataset contained a total of around 35k images with highly unbalanced classes. The class 0 that is the class with no signs of any problem was alone having a total of 25k images.

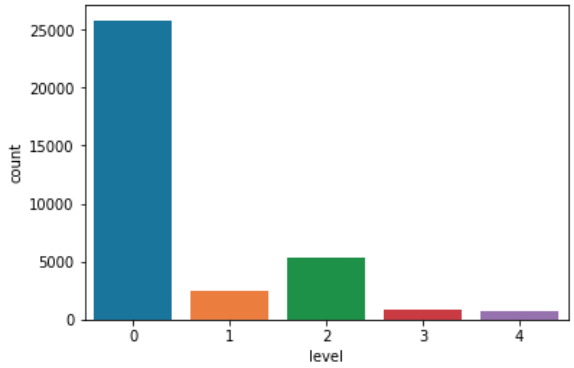

Figure 2. The initial distribution of dataset

To remove this problem oversampling of the low frequency classes was done such that after oversampling each class had around 2.6k images and the total number of dataset was 14k. Apart from class imbalance all the images were of very resolution and so were resized to the size of 224 .

6. Datasets and Experiments

6.1. Datasets

For our experiments we used following two datasets, Diabetic Retinopathy Detection [1] which has a large set of high-resolution retina images taken under a variety of imaging conditions. A left and right field is provided for every subject. Images are labeled with a subject id as well as either left or right. A clinician has rated the presence of diabetic retinopathy in each image on a scale of 0 to 4, according to the following scale: 0 - No DR, 1 - Mild, 2 - Moderate, 3 - Severe, 4 - Proliferative DR. We also used Histopathologic Cancer Detection dataset [2] which consists of 220k cancer and non-cancer images.

6.2. Experiments

As discussed before the primary task of our experiment was of image classification. The severity of the disease increasing as the class number for the diabetic retinopathy dataset. In total we ran three experiments as follows :

1. Resnet on diabetic retinopathy dataset
2. Residual attention network on retinopathy dataset
3. Residual attention network on the histopathologic cancer detection dataset

The resnet model was finetuned on an already imagenet pretrained model and was trained for 8 epochs. While the residual attention networks were trained for around 10 epochs each. Batch size of 8 was used for all the training processes with learning rate of 0.001 . A validation set of around 1k images were used for the diabetic retinopathy dataset.

We perform experiments using residual attention network on the Histopathologic Cancer Detection dataset. We used Adam optimizer with initial learning rate of 0.001. We initialised weights of out network using xavier initialisation. After training the network for 10 epochs we achieved F-Score of 0.95 at threshold 0.51. For interpreting the model, we applied GradCam to the feature maps of the last attention layer of the network.

6.3. Results

The classification accuracy of resnet vs the residual attention network are shown in the table:

Architecture	Accuracy
Resnet	90.2
RAN	91.1

Table 1. Accuracy on diabetic retinopathy dataset

The improvement in accuracy shows that residual attention network indeed learn something which might help it in classifying better.

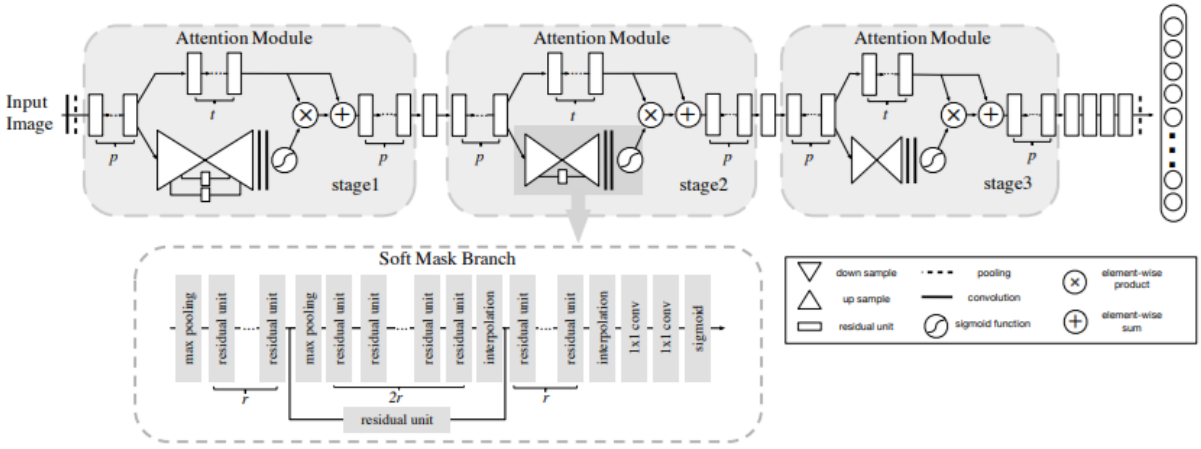

Figure 1. Architecture for residual attention network taken from paper

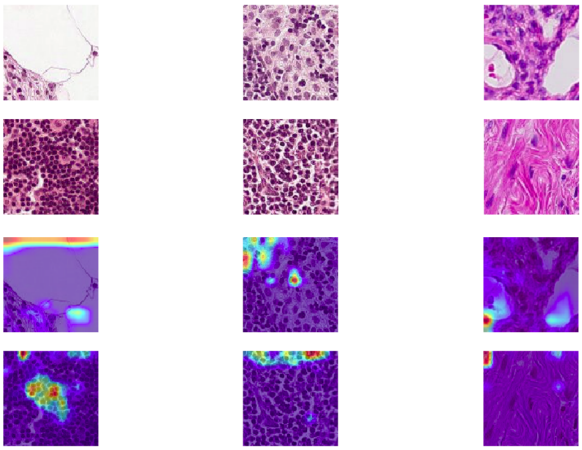

Figure 3. Residual Attention Network, Feature Maps : Attention Stage 1, Histopathologic Cancer Detection Dataset

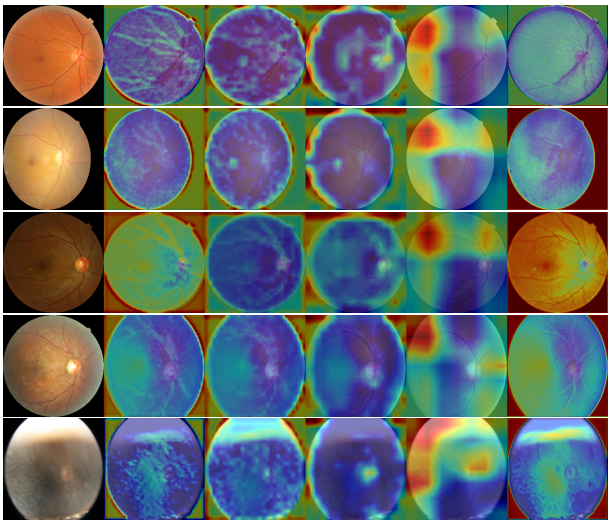

Figure 4. Resnet GradCAM at different layers namely: ReLu, Layer1, Layer2, Layer3, Layer4

Comparing both the networks we can see that the explicit attention layer grad cam is more sharp and better than that of the layers of the normal Resnet. This indicates the better role of attention modules in understanding the model and making it better. However grad cam proves to be good heuristic of how the network learns and makes decisions.

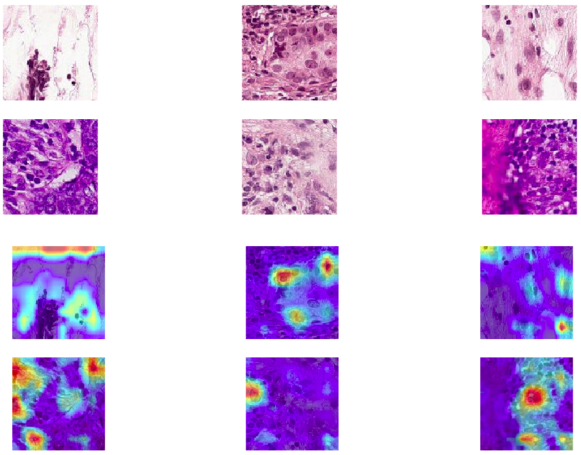

Figure 6. Residual Attention Network, Feature Maps : Attention Stage 6, Histopathologic Cancer Detection Dataset

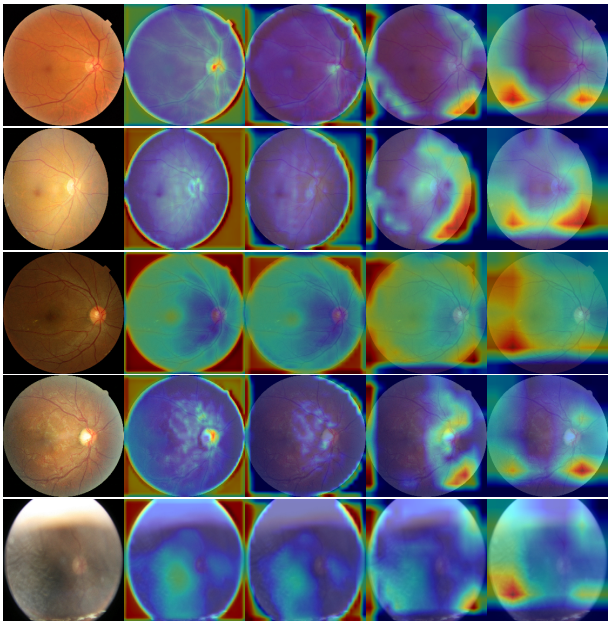

Figure 5. Residual Attention Network GradCAM at 3 different levels of attentions and final feature layer

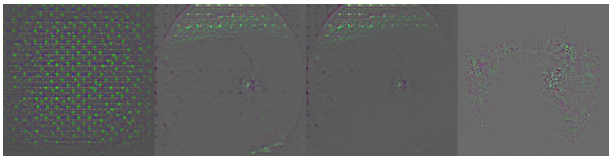

Figure 7. Different backpropagation techniques applied for stage 4 level eye in resnet model : deconv, guided backpropagation , guided gradcam , vanilla backpropagation

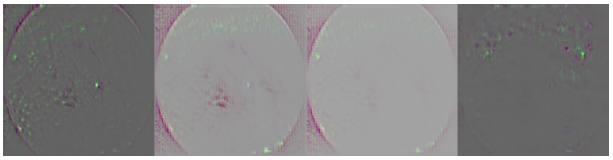

Figure 8. Different backpropagation techniques applied for stage 4 level eye in residual model : deconv, guided backpropagation , guided gradcam , vanilla backpropagation

Other methods such as guided backpropagation , guided gradcam and deconv are not able to find the details required in making the decisions in these images , it may be because the disturbance in the medical images is very low compared to normal images as in imagenet. Therefore these techniques fails to provide better good interpretability.

7. Conclusion

In this work, we see how different visualization techniques can be applied to neural networks in the medical image classification domain.Such interpretations of deep neural network can improve medical facilities and help improve automated deep networks by making them more robust against any kind of errors they make with the aid of humans. We see how grad cam can help in visualizing the network models in this domain but guided backpropagation and deconv fail to provide significant contribution to the understanding. Although it is a good start but still the methods used are not sufficient to completely trust the decisions of a neural network. Future work includes using adversarial attacks to determine the robustness of the classifier.

References

[1] Diabetic retinopathy detection, https://www.kaggle.com/c/diabetic-retinopathy-detection/data. 2

[2] Histopathologic cancer detection dataset, https://www.kaggle.com/c/histopathologic-cancer-detection. 2

[3] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter. Visualizing evidence for Alzheimer's disease in deep neural networks trained on structural MRI data. *arXiv e-prints*, page arXiv:1903.07317, Mar 2019. 1

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1

[5] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. *CoRR*, abs/1505.04366, 2015. 1

[6] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 1

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 1

[8] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. *CoRR*, abs/1704.06904, 2017. 1