Sanity Checks for Saliency Maps

CS 689 Project Report

Tathagat Verma 180050111

Code for experiments of this report can be found here.

Introduction

Saliency maps are tools that help explain features of the input, generally an image, on which a model puts more weight while making predictions. The applications of such explanation methods have led to much excitement and significant contributions. However, the problem that remains unaddressed thoroughly, due to a dearth of principled guidelines for evaluations, is the assessment of saliency maps.

Contributions

In this paper, the authors propose two, easy to implement, sanity checks - (i) *Model parameter randomization test* and (ii) *Data randomization test*. These tests are necessary but not sufficient, for a saliency map to be considered fit for a task. The motivation for these tests is that the model's and hence the saliency map's outputs must be dependent on the data the model is trained on and also on its weights. If it happens that a saliency map is independent of these, then we can conclude directly that the saliency map does not provide an appropriate explanation of the model's decision.

The authors evaluate several saliency methods - Gradient, Gradient$\odot$Input, Integrated Gradients (IG), Guided Back-prop (GBP), Guided GradCAM, SmoothGrad (SG) on these two tests and find out that Guided Back-prop and Guided GradCAM *fail* their tests.

Further, the authors show through experiments that relying on visual assessment of explanation methods is unjustified as visual assessments rely only on absolute values of explanations and not on their multiplicative sign.

In addition to the authors work, I have proposed another variant of the model parameter randomization test which involves addition of noise to parameter weights.

Metrics

In all evaluations, we want to compare similarity between explanations for the original model and for the modified model (modified as per the test). For this purpose, the authors use the following metrics:

- Structural similarity index (SSIM)
- Pearson correlation of histogram of gradients (HOGs)

- Spearman correlation rank with absolute value (absolute)
- Spearman correlation rank without absolute value (diverging)

Model Parameter Randomization Test

Since a model prediction depends on its weights, an explanation method must also be dependent on the model parameters. Thus, for a given explanation method, if we randomize the weights of a model, we must expect a significant difference in the explanation when compared to that for the original trained model. This test is useful for checking an explanation method's validity for model debugging, if a model has weights different from a trained model then their explanations must differ appreciably.

Specifically, there are two variants of this test:

1. Cascading Randomization: In this test, parameters of the model from the top layer until an intermediate layer are all randomized. This is done successively for all layers.
2. Independent Randomization: In this test, parameters of only a single layer are randomized, while keeping other layers unchanged.

Data Randomization Test

In classification tasks, an explanation method must give maps explaining the decisions of a model. If we completely randomize the labels for all inputs while training a model, then it can do nothing better than random guessing. Thus, if an explanation method is sound, it must have a significantly different explanation for the model which was trained with randomized weights. If not, it means that the explanation method did not capture the correct explanation for a model's decision in the first place.

Additional Work

Taking motivation from the model parameter randomization test, I have tried out another variant of model parameter randomization - *adding noise to parameters* instead of complete randomization. Specifically, I add zero mean Gaussian noise to the parameters. Intuitively, at a higher noise level, the explanation must change significantly and at the lower noise level, the difference must be minimal. The noise level in this case is controlled by the variance of the Gaussian noise we add to the weights.

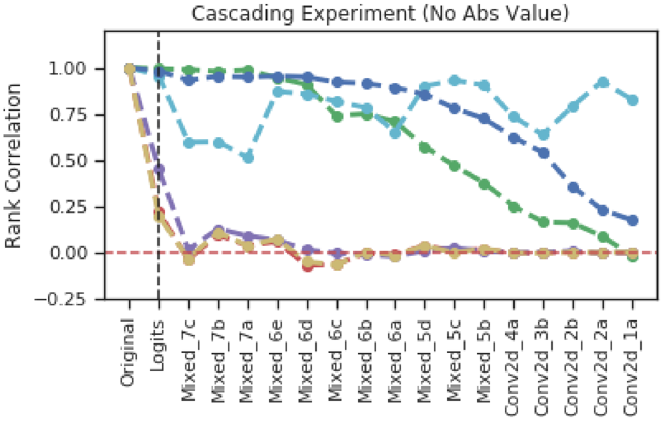
At a fixed noise level, the methods which have lower similarity to the original explanations are more sensitive to model parameters and can, hence, be said to give better explanations for model decisions. So this test can serve two purposes - (i) sanity check, same use case provided by the authors' model parameter randomization test and (ii) sensitivity of an explanation method to model parameters. Sensitivity to model parameters can be particularly useful in some scenarios while not so important in others, depending on the task at hand. For example, in model debugging, it can be useful to have more sensitive explanations if we want precise model parameters.

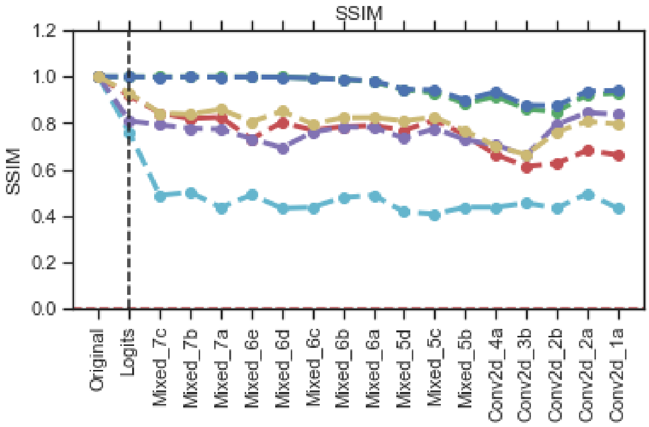
There can be more types of noise, e.g. Poisson noise, that one can add to the parameters for mimicking task specific scenarios and giving better sensitivity measures.

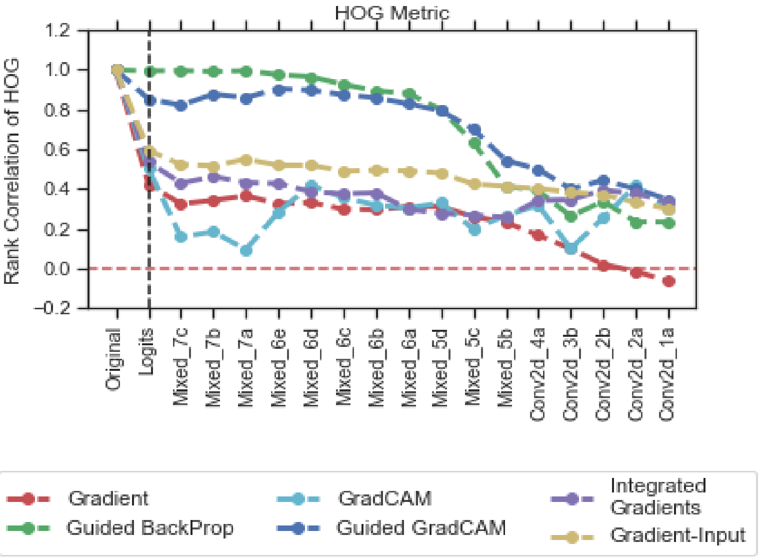
Experiments

In this section, first, results of the experiments for the model parameter randomization test using cascading randomization are presented (replicating the paper), which have convincing quantitative results. We do not show the data randomization test results since these are only qualitative. Then we show results of the newly proposed noise addition test.

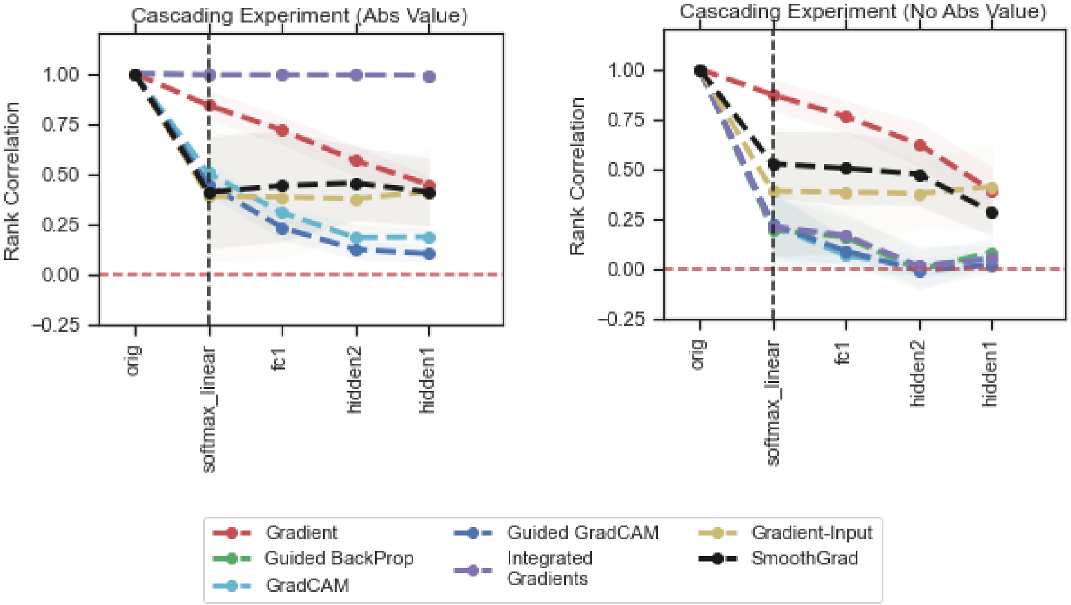
1. In the model parameter randomization test, we observe that Guided Back-prop and Guided GradCAM are insensitive to parameters of the higher layers which is indicated by high similarity of the explanation to that of the original model.
   The figures below are for the Inceptionv3 model trained on the ImageNet dataset. The numbers plotted are - (i) Spearman correlation rank without absolute value, (ii) SSIM and (iii) rank correlation of HOGs. We notice in all three plots that for Guided Back-prop and Guided GradCAM, the similarity/correlation is close to 1 for the initial (top) layers, showing that these explanation methods are invariant to the top layer weights. For the other explanations, the correlation falls appreciably from the top layer itself, hence passing this test.







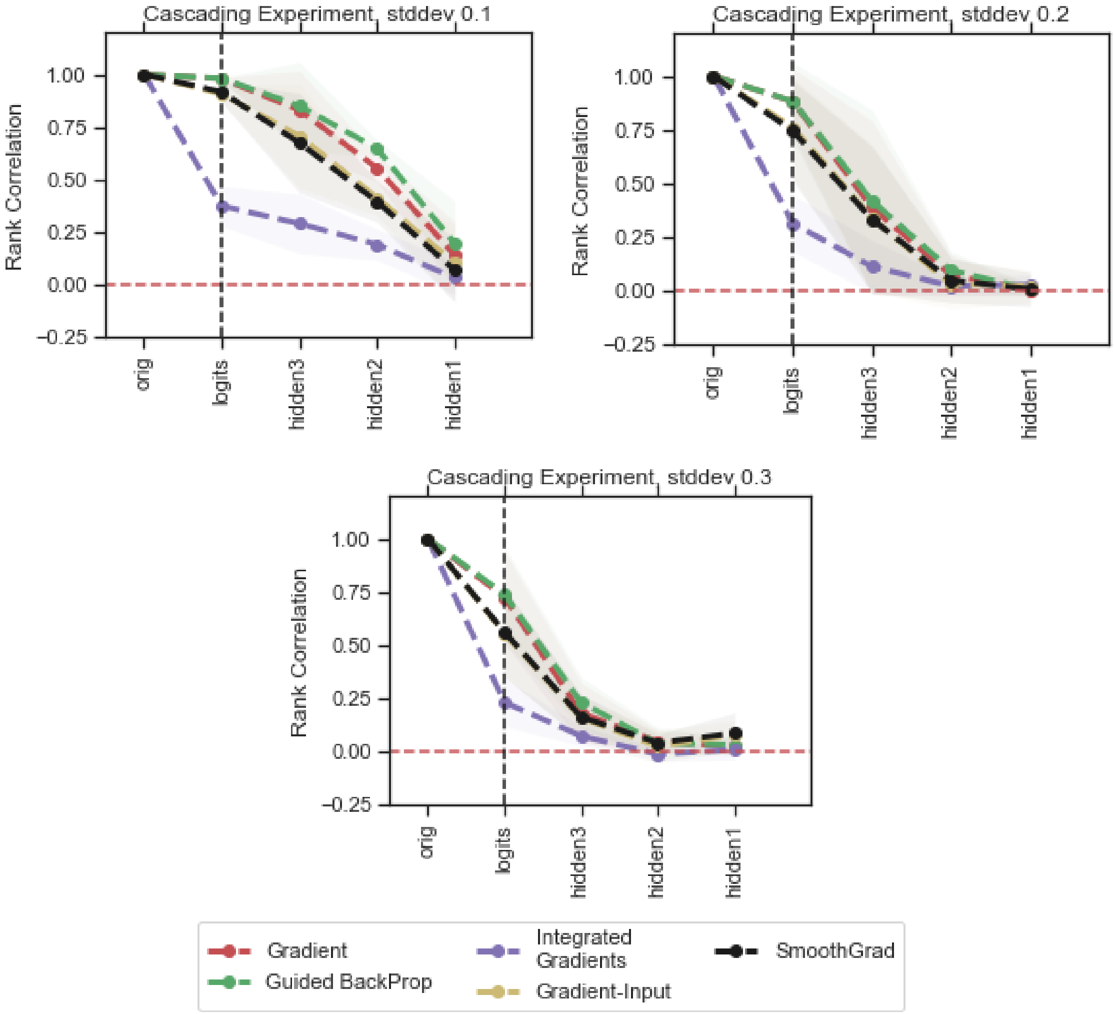
2. Performing the same cascading randomization experiment in the case of a CNN model trained on the MNIST dataset, we get the two figures (scores for absolute and non-absolute values of the explanations) shown below. Looking at the absolute value scores, we might say that Integrated Gradients is similar to the output since its correlation is close to 1. Since visual assessment is made using these absolute values, one would come to the same conclusion by looking at the images for explanations (which are indeed similar, not shown due to the limited size of the report). However, only by looking at the non-absolute values can we know that Integrated Gradients actually have very low correlation values since the explanation method had changed signs of the map upon randomization of weights. Thus Integrated Gradients does pass the test which cannot be predicted by only seeing the images of explanations.



3. Now we add noise to model parameters and observe sensitivity of explanation methods at varying noise levels. We report Spearman rank correlations without absolute values of various explanation methods for an MLP model trained on the MNIST dataset. The standard deviations of Gaussian noise used are - (i) 0.1 (ii) 0.2 and (iii) 0.3.

   From these figures we can observe a clear trend of reducing rank correlation with increasing noise level. Another interesting observation is convexity-concavity of these plots. At the lower noise level (standard deviation 0.1) the rank correlation graph is concave while at higher noise levels, these curves become increasingly convex. This is justified since at low noise levels, the top layers do not have enough stochasticity to generate appreciably different explanations. As we cascade the layers, the overall stochasticity increases, leading to lower correlation explanations. At higher noise levels, the model has enough stochasticity at the top layers itself to produce low similarity explanations.

   In addition to this, across all noise levels, we observe integrated gradients to be the most sensitive to noise in parameters and Guided Back-prop to be the least sensitive. Hence, for tasks where we want high sensitivity to model parameters, integrated gradients can be the best suited method.



Conclusion

In this report, we summarized the contributions of authors, viz. two sanity checks for saliency maps - model parameter randomization test and data randomization test. In addition to this, another model noise test is introduced in which we add noise to model parameters instead of complete randomization. This new test also acts as a sensitivity comparison method.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292.