GEOMETRICALLY GUIDED SALIENCY MAPS

Md Mahfuzur Rahman, Noah Lewis & Sergey Plis *
Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS)
Georgia State University, Georgia Institute of Technology, Emory University
Atlanta, GA, USA
{mahfuz.gsu, lhd231, s.m.plis}@gmail.com

ABSTRACT

Interpretability methods for deep neural networks mainly focus on modifying the rules of automatic differentiation or perturbing the input and observing the score drop to determine the most relevant features. Among them, gradient-based attribution methods, such as saliency maps, are arguably the most popular. Still, the produced saliency maps may often lack intelligibility. We address this problem based on recent discoveries in geometric properties of deep neural networks' loss landscape that reveal the existence of a multiplicity of local minima in the vicinity of a trained model's loss surface. We introduce two methods that leverage the geometry of the loss landscape to improve interpretability: 1) "Geometrically Guided Integrated Gradients", applying gradient ascent to each interpolation point of the linear path as a guide. 2) "Geometric Ensemble Gradients", generating ensemble saliency maps by sampling proximal iso-loss models. Compared to vanilla and integrated gradients, these methods significantly improve saliency maps in quantitative and visual terms. We verify our findings on MNIST and Imagenet datasets across convolutional, ResNet, and Inception V3 architectures.

1 INTRODUCTION

Deep Learning (DL) has significantly advanced in different application areas, especially in computer vision and natural language processing (Chai et al., 2021; Young et al., 2018). One of the defining factors of such advancements is less reliance on domain-specific expert knowledge compared to classical machine learning approaches since DL models can learn directly from the data (LeCun et al., 2015). In other words, DL diminishes the need for feature engineering, which predominantly relies on strict assumptions of the often unknown underlying process that generates the data. Arguably, strict assumptions restricting a model may prevent the natural emergence of the significant features, while DL is only restricted by the training data and mostly weak inductive bias (Battaglia et al., 2018). However, the improved performance of DL comes at the cost of intelligibility, i.e., its decision-making process is quite inscrutable to human beings (Ras et al., 2022). In particular, previous studies have shown that models can identify and rely on spurious correlations (Geirhos et al., 2020). This lack of intelligibility prevents the widespread deployment of DL models in safety-critical domains such as healthcare, medicine, neuroscience, and self-driving cars, to name a few. As such, the rapid progress of DL models and their spread across diverse disciplines have encouraged researchers to understand the rationale behind DL models' decisions.

Recent studies have attempted to tackle deep learning interpretability by explaining the model's overall decision-making process or generating post-hoc explanations for each prediction. Generating a post-hoc explanation is comparatively easy to interpret a model's behavior. To this end, people have proposed different methods (Samek et al., 2020) to generate post-hoc explanations. However, the performance varies widely across different architectures and domains. Moreover, except IG, very few attribution methods satisfy desirable properties like explanation continuity (Montavon et al., 2018), implementation invariance (Sundararajan et al., 2017), sensitivity-n (Ancona et al., 2017), completeness and so on.

Saliency methods focus primarily on the sensitivity of the model's prediction to the input. Generally, vanilla gradients (GRAD) (Baehrens et al., 2010; Simonyan et al., 2013), which measure the sensi-

*https://trendscenter.org/

tivity of the output score for each input, and integrated gradients (IG) (Sundararajan et al., 2017) are noisy (Montavon et al., 2017; Samek et al., 2016; Smilkov et al., 2017; Sturmfels et al., 2020). To this end, we propose an algorithm called *Geometrically Guided Integrated Gradients (GGIG)* that guides traditional IG to reduce noise in the saliency maps. GGIG, akin to *DeepDream* (Mordvintsev et al., 2015), follows the steepest path in the locality of each interpolation point of the linear IG path. The intuition behind following the steepest path comes from the concept of class activation maximization (Couteaux et al., 2019). Indeed, we follow the path that maximizes activation for the class-specific logit because this is when the model is most performant. We show that guiding IG using local loss geometry improves the quality and robustness of the saliency maps.

Saliency methods do not address the local behaviors within the loss landscape, even though several studies (Garipov et al., 2018; Gotmare et al., 2018; Izmailov et al., 2018; Entezari et al., 2021) have shown that geometric ensemble model is possible by combining multiple local optima in the parameter space. We predict that there is a vast scope to build a novel interpretability method either by improving the loss landscape during training or leveraging the loss geometry during post-hoc explanations. Based on this intuition, we propose a second interpretability method, called *Geometric Ensemble Gradients (GEG)*, that analyzes the model's behavior for each input by sampling from the proximity of the model in weight space. Our quantitative analysis and visual inspection of GEG maps revealed that leveraging loss geometry has excellent potential for improving model interpretability.

Our main contributions are as follows:

- We propose an interpretability method, called *Geometrically Guided Integrated Gradients*, that guides the traditional IG by using the local loss geometry of the model.
- A second interpretability method, *Geometric Ensemble Gradients*, which, for the first time, uses a geometric ensemble for model interpretability.
- We show that the proposed methods offer better saliency maps for different datasets and architectures when assessed through visual inspection and quantitative metrics.

2 Related Work

DL models are often difficult to interpret due to many parameters and non-linearity. Fortunately, many studies have so far proposed a large number of interpretability methods (Baehrens et al., 2010; Simonyan et al., 2013; Shrikumar et al., 2017; Smilkov et al., 2017; Sundararajan et al., 2017; Selvaraju et al., 2017; Springenberg et al., 2014; Bach et al., 2015; Montavon et al., 2017; Hooker et al., 2019; Adebayo et al., 2018a; Zeiler & Fergus, 2014) with their relative costs and benefits.

Gradient-based methods, also referred to as *visualization methods* (Ras et al., 2022), are fast, easy to implement, and readily applicable (Sundararajan et al., 2017) to existing models compared to perturbation-based methods. However, the saliency maps generated using gradient-based methods (e.g. GRAD and IG) are usually affected by the high-frequency variations, resulting in very different attributions for the neighboring pixels, violating the *explanation continuity* principle. This lack of *explanation continuity* increases the problem of shattered gradients (Balduzzi et al., 2017)—a problem usually caused by a high number of piecewise non-linearity in the learned function. Consequently, the generated explanations may become highly sensitive to the small changes in input and may not represent the model's decision-making process. IG has its own specific problem, it depends heavily on the choice of the baseline (Sahoo et al., 2020). Regardless of the chosen baselines, any feature within the baseline that is numerically close to the target sample receives very low attributions and may not be proportional to the true importance. To reduce the inherent noise in the saliency maps, we may utilize some useful loss landscape properties as observed in several studies (Garipov et al., 2018; Gotmare et al., 2018; Izmailov et al., 2018; Entezari et al., 2021) to design reliable interpretability methods.

3 Proposed Methods

We propose two methods, *Geometrically Guided Integrated Gradients (GGIG)* and *Geometric Ensemble Gradients (GEG)*. In GGIG, we incorporate the idea of *path methods* (Sundararajan et al.,

2017) and enhance the quality of the attribution by analyzing the local loss behavior. Like IG, GGIG starts from the baseline $\boldsymbol{x}'$. However, instead of following a linearly interpolated path, it finds the point that offers the steepest slope in the proximity of the next linear interpolation point, similar to gradient ascent. GGIG thus maximizes class activation for each of the linearly interpolated points. **Algorithm** 1 (Appendix A) describes how we traverse the path in the steepest direction of the slope. The procedural steps for GGIG are shown in **Algorithm** 2 (Appendix B). We hypothesize that, based on the *explanation continuity* axiom of attribution methods, the prediction curve in the vicinity of $\boldsymbol{x}$ holds important information about the interaction between model f and input $\boldsymbol{x}$. For GEG, we take into account the model's local behavior for the input $\boldsymbol{x}$ by sampling around the model using i.i.d $\mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$ and consider only those model samples that produce original or larger prediction scores for the input. We use GRAD of the input computed for each sampled model output as a candidate map. We finally ensemble them to produce the final saliency map for the input. The key steps involved in GEG are shown in **Algorithm** 3 (Appendix C). Figure 1 shows the schematic diagram illustrating the mechanisms of the GGIG and GEG.

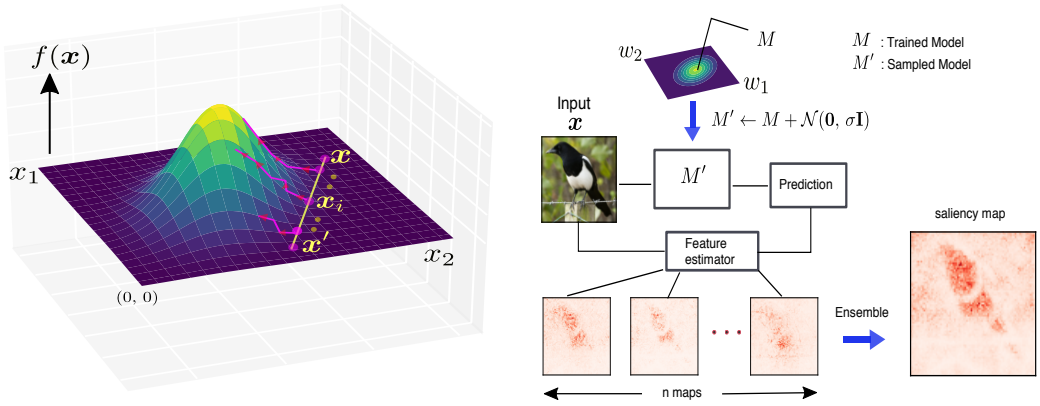

Figure 1: **Left:** The basic idea of "Geometrically Guided Integrated Gradients." It takes a baseline $\boldsymbol{x}'$, creates a linearly spaced path to the actual input $\boldsymbol{x}$. From each interpolation point, it follows a trajectory in the loss landscape that maximizes class activation. **Right:** The main intuition behind "Geometric Ensemble Gradients." It perturbs the trained model and considers only the sample-specific optimal model to build an ensemble explanation for the sample.

4 Experiments

Training on MNIST Dataset For MNIST, the model architecture was a CNN that consisted of two convolutional layers with (32, 64) filters of sizes (5, 5). Each convolutional layer is followed by a $2x2$ max pooling layer and a ReLU activation. We fed the final convolution output to a fully connected network with 1024 input and 10 output units (softmax). We optimized the model using stochastic gradient descent (SGD) with a learning rate of 0.0004 and momentum of 0.9. The model was trained for 400 iterations with a mini-batch size of 64 and finally achieved an accuracy of 99.2%.

Post hoc explanation experiments For GGIG, we used a learning rate of 0.0001 for gradient ascent from each linear interpolation point between input $\boldsymbol{x}$ and baseline $\boldsymbol{x}' = \mathbf{0}$. For MNIST, we iterated the gradient ascent for 200 steps and noted the maximum sensitivity along the gradient ascent trajectory for each input. For GEG, we sampled 200 models M' around the trained model M using i.i.d $\mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$ that produced the same or larger prediction score compared to the original score. We assigned the maximal possible sensitivity within the model's local landscape as feature attributions.

The traditional integrated gradients scale raw attributions by $\boldsymbol{x} - \boldsymbol{x}'$ (i.e., difference between input $\boldsymbol{x}$ and baseline $\boldsymbol{x}'$). However, the input and the baseline greatly affect the feature attributions. Furthermore, this point-wise multiplication was initially justified to sharpen the gradient explanations; however, it is better justified when a measure of salience is a priority over mere sensitivity (Ancona et al., 2019). In this paper, we were more interested in the sensitivity of features rather than their marginal salience to the target score. So we omitted the scaling factor $\boldsymbol{x} - \boldsymbol{x}'$ while generating

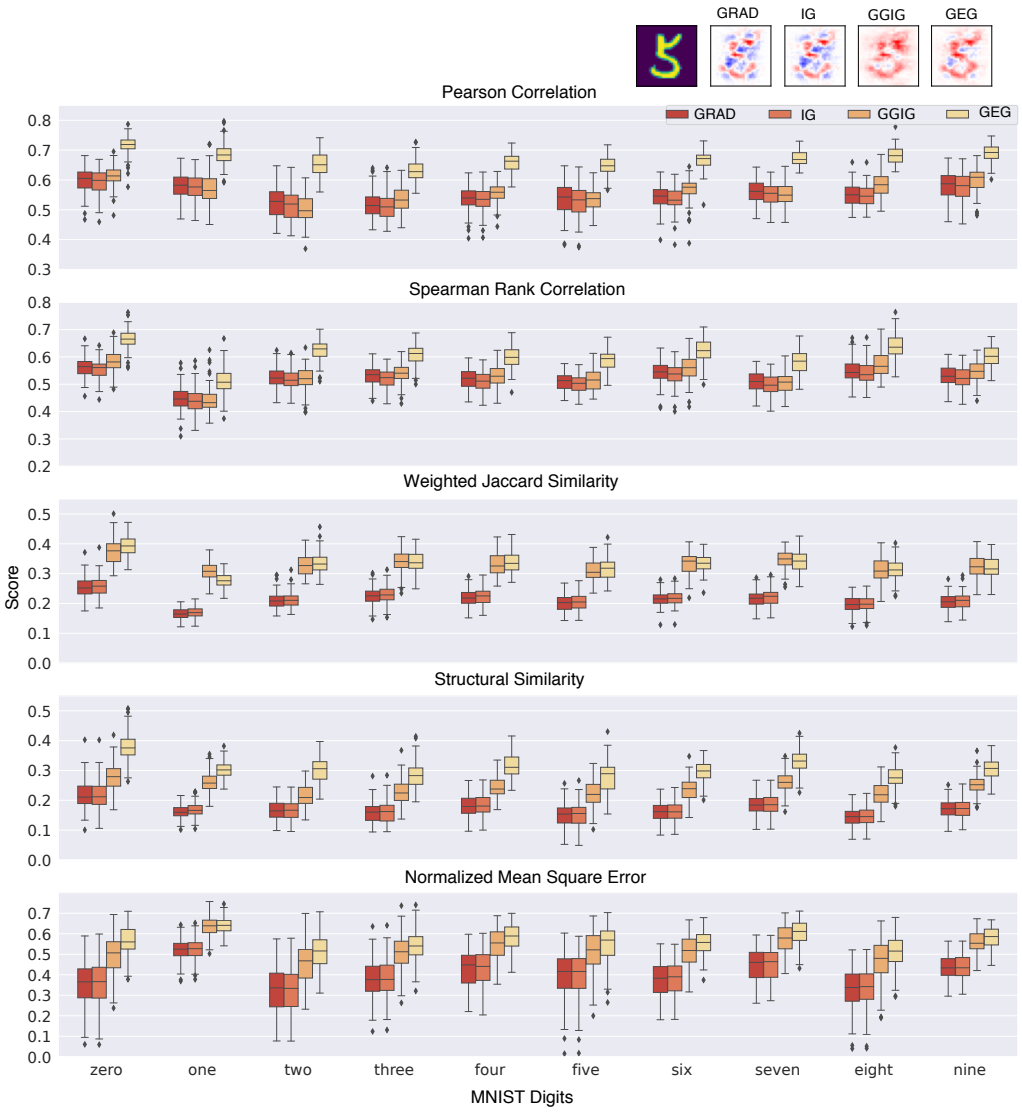

Figure 2: **Top:** Selected maps for MNIST samples generated using different methods. **Bottom:** Quantitative evaluations using different correlation and similarity metrics. It is obvious that the maps produced by GGIG and GEG have higher structural and numerical correlations with the input. Precisely, GEG, as an explanation method, performs the best by a large quantitative margin. GGIG also performs better than GRAD and IG.

integrated gradients for comparison across the methods. We display the saliency maps obtained on the MNIST dataset in Figure 2.

We also performed a *Model Randomization* test (Adebayo et al., 2018b) to verify the sensitivity of the methods to the model parameters. To this end, we randomly reinitialized the weights and generated post-hoc explanations using the randomized model. The proposed methods, GGIG and GEG, were as sensitive as GRAD and IG, suggesting that both of our methods reflect information learned by the model.

Quantitative Evaluation on MNIST Dataset Visual inspection of explanation methods can be unreliable as it is possible to create adversarial samples (Goodfellow et al., 2014; Szegedy et al., 2013) that can fool the human eye, totally changing the model predictions. We perform different similarity measures between the maps and the input to understand the quality of the proposed methods on MNIST, namely, *Pearson Correlation Coefficients, Spearman Rank Correlation, Weighted Jaccard*

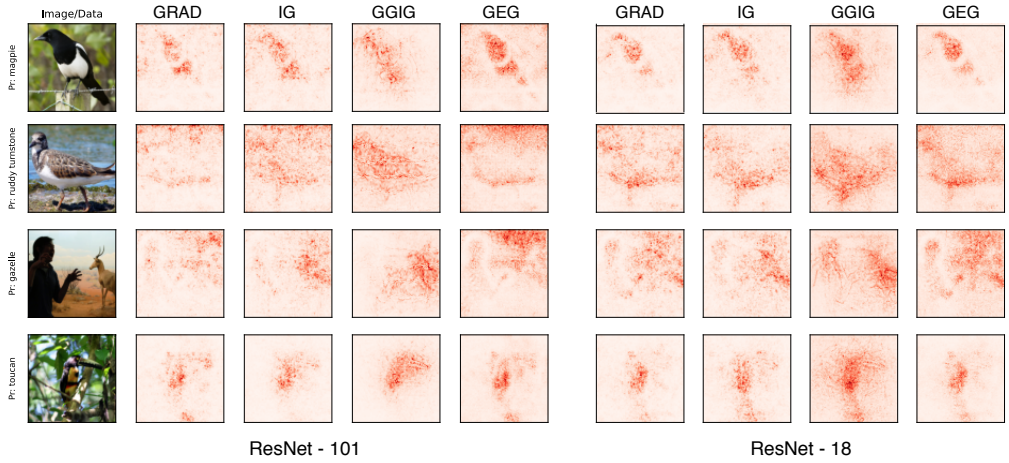

Figure 3: Comparative saliency maps for ImageNet dataset generated by GRAD, IG, GGIG, and GEG methods. The maps obtained using GGIG and GEG are more discriminative and more clearly reveal the underlying structure of the class-associated objects.

Similarity, Structural Similarity, and Normalized (Reverse) Mean Square Error (the more, the better). For quantitative evaluation, both data and maps were rescaled in the range [0, 1]. We assumed that the amount of information a method can capture about the structure and distribution of the input in the saliency maps directly determines its quality as an explanation method. Figure 2 shows sample maps generated by different interpretability methods and the detailed results of quantitative evaluation. As we can observe, when assessed quantitatively, the GEG method produces the most relevant interpretation of the model's prediction. GGIG is the second best performing method. Comparatively, GRAD and IG retain little information about the numerical and structural association to the input.

Experiments on ImageNet: We also evaluated the proposed methods using a small subset of images from the ImageNet dataset (Krizhevsky et al., 2012) and different pretrained models, namely ResNet-18, ResNet-34, ResNet-50, ResNet-101 (He et al., 2016), and Inception V3 (Szegedy et al., 2016). Though we found meaningful maps in every case, maps still vary in quality possibly due to their architectural differences. Figure 3 shows some maps (more results are in Appendix) produced using different saliency methods.

5 Conclusion

In this paper, we propose two interpretability methods inspired by the recent discoveries of the loss landscape properties. Our first method, GGIG, uses gradient ascent to guide traditional IG and enhance the quality of the feature attributions. This method maximizes class activation to modify the linear path implemented in IG. The second method, GEG, perturbs the model, optimal to the input $\boldsymbol{x}$, to build an ensemble explanation for each prediction. With GEG, we show, for the first time, that it is possible to ensemble geometric properties of the loss surface for interpretability. We show the effectiveness of the proposed methods through visual inspection and several quantitative metrics. We verify our findings across convolutional, ResNet (ResNet-18, ResNet-34, ResNet-50, ResNet-101), and Inception V3 architectures on two benchmark datasets (MNIST and ImageNet). As demonstrated, we expect that this work is a stepping stone toward building a more rigorous interpretability method leveraging the interactivity of the model's loss geometry with the input.

References

Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*, 2018a.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31:9505–9515, 2018b.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Gradient-based attribution methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 169–191. Springer, 2019.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, pp. 342–350. PMLR, 2017.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

Vincent Couteaux, Olivier Nempont, Guillaume Pizaine, and Isabelle Bloch. Towards interpretability of segmentation networks by analyzing deepdreams. In *Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support*, pp. 56–63. Springer, 2019.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8803–8812, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. Using mode connectivity for loss landscape analysis. *arXiv preprint arXiv:1806.06977*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 9737–9748, 2019.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015.

Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397, 2022.

Subham Sekhar Sahoo, Subhashini Venugopalan, Li Li, Rishabh Singh, and Patrick Riley. Scaling symbolic methods using gradients for neural model explanation. In *International Conference on Learning Representations*, 2020.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

A GEOMETRICALLY GUIDED GRADIENTS

Algorithm 1 Geometrically Guided Gradients (G^3)

Input: model M, sample $\boldsymbol{x}$, iterations n, learning rate lr
Output: G^3 explanation $\boldsymbol{e}$

function GRADASCENT(M, $\boldsymbol{x}$, n, lr)
  gradients $G \leftarrow \{\}$
  for $k \leftarrow 1$ to n **do**
    $f \leftarrow \text{score}(M(\boldsymbol{x}))$
    $\boldsymbol{e}_k \leftarrow \nabla_{\boldsymbol{x}} f$
    $G \overset{+}{\leftarrow} \boldsymbol{e}_k$
    $\boldsymbol{x} \leftarrow \boldsymbol{x} + lr \times \boldsymbol{e}_k$
  end for
  $\boldsymbol{e} \leftarrow \max(G)$ ▷ pixel-wise maximum attribution
  return $\boldsymbol{e}$
end function

B GEOMETRICALLY GUIDED INTEGRATED GRADIENTS

Algorithm 2 Geometrically Guided Integrated Gradients (GGIG)

Input: model M, sample $\boldsymbol{x}$, steps n, baseline $\boldsymbol{x}'$, iterations i, learning rate lr
Output: GGIG explanation $\boldsymbol{e}$

gradients $G \leftarrow \{\}$
$\alpha = {}^{(\boldsymbol{x}-\boldsymbol{x}')}/_{n}$
for $k \leftarrow 0$ to n **do**
  $\boldsymbol{x}_k \leftarrow \boldsymbol{x}' + k \times \alpha$
  $G \overset{+}{\leftarrow}$ GRADASCENT(M, $\boldsymbol{x}_k$, i, lr)
end for
$\boldsymbol{e} \leftarrow \max(G)$ ▷ pixel-wise maximum attribution
return $\boldsymbol{e}$

C GEOMETRIC ENSEMBLE GRADIENTS

Algorithm 3 Geometric Ensemble Gradients (GEG)

Input: model M, sample x, iterations n
Output: GEG explanation $\boldsymbol{e}$

prediction $p \leftarrow \text{predict}(M(\boldsymbol{x}))$
score $s \leftarrow \text{score}(M(\boldsymbol{x}))$
gradients $G \leftarrow \{\}$
for each of n iterations **do**
  $M' \leftarrow M + \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$
  $s' \leftarrow \text{score}(M'(\boldsymbol{x}))$
  $p' \leftarrow \text{predict}(M'(\boldsymbol{x}))$
  if $s' \geq s$ **and** $p' = p$ **then** ▷ if model performs equally or better
    $G \overset{+}{\leftarrow} \nabla_{\boldsymbol{x}} s'$
  else
    continue;
  end if
end for
$\boldsymbol{e} \leftarrow \max(G)$ ▷ pixel-wise maximum attribution
return $\boldsymbol{e}$

D SALIENCY MAPS (IMAGENET DATASET)

D.1 ARCHITECTURE: RESNET-34

Figure 4: Comparative saliency maps for ImageNet samples generated by GRAD, IG, GGIG, and GEG methods. The underlying structures of the objects are more clearly visible and reasonably consistent in the GGIG and GEG maps.

D.2 ARCHITECTURE: RESNET-50

Figure 5: Explanations (saliency maps) of ResNet-50 predictions for ImageNet samples generated by GRAD, IG, GGIG, and GEG methods. As we can see, the GGIG and GEG maps have strong correlation and similarity with the corresponding input images.

D.3 ARCHITECTURE: INCEPTION V3

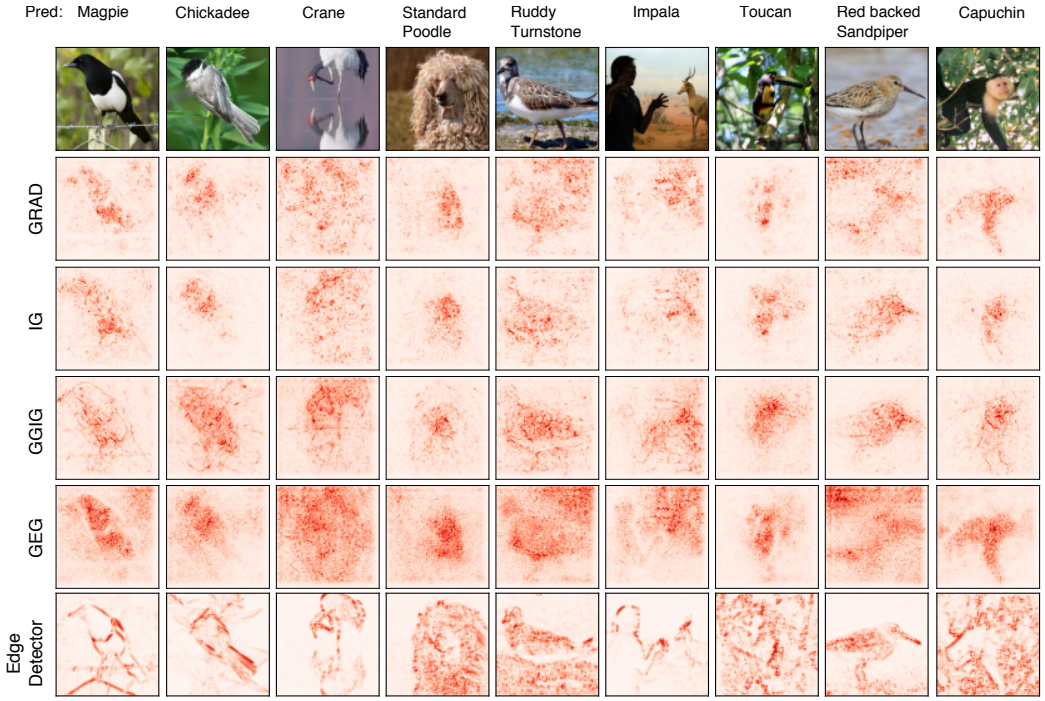

Figure 6: Saliency maps for selected ImageNet samples. It is evident that the GGIG and GEG methods consistently capture the useful concepts as used by the model for predictions.

D.4 BACKGROUND REPLACEMENT EXPERIMENTS

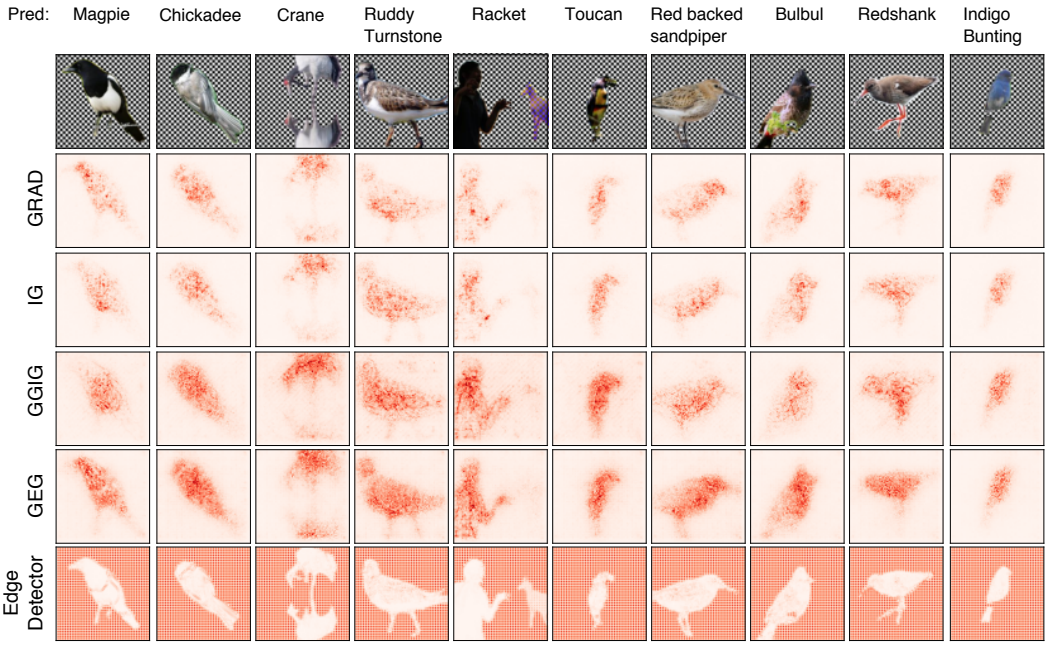

Figure 7: We assigned ImageNet samples a very different fixed background (a black and white checkerboard) and generated post-hoc explanations. It is obvious from the resulting maps that the model learned concepts, not merely edges from the training objects, and all the saliency methods supposedly ignored the background. It is apparent that GGIG and GEG methods retained learned concepts more accurately during post-hoc explanations.