Locally Smoothed Attribution for Genomic CNNs

Ian Nielsen#1

#Electrical and Computer Engineering, Rowan University
201 Mullica Hill Road, Glassboro, NJ 08028, United States
1nielseni6@rowan.edu

Abstract— **This paper proposes a new technique of generating attribution maps and applies it to a Convolutional Neural Network (CNN) for the task of classifying genomic data. The proposed method is called Shift Smoothed Gradients. This method averages the gradients of multiple shifted samples to generate a "smoothed" version of the gradient. This approach differs from previous approaches because it accounts for localized changes in the gradient. This approach provides transparency into the learned features which are most important to the neural network. The method is tested for two metrics: shifting invariance, and whether the map highlights features of interest. The proposed method outperforms other recent methods on the two metrics when tested using a modern CNN architecture trained on human and goldfish genome data. When applied to genomic data, this method provides insight into what differentiates two or more different samples of DNA. This method can also be extended to other tasks and network architectures besides CNNs.**

This paper also comes with code:
https://github.com/nielseni6/ShiftSmoothedAttributions

Keywords— **Explainability, XAI, Bioinformatics**

I. Background and Introduction

This project focuses on attribution, which is a subset of explainability methods which are used to explain the predictions of a neural network. Attribution can be defined as a method of mapping a score, often referred to as attribution score, to each input feature. In the case of image classification these features would be pixels, and in the case of genomic data classification the features would be the letters representing each nucleotide. The attribution map that is created must have the same dimensions as the input to the network, since we must show the attribution score of all input features.

Much of explainable artificial intelligence has largely focused on image processing tasks. While this application is very useful, there has not been much research into explainability using genomic data. Some notable recent methods which focus on image classification include SmoothGrad [1], Integrated Gradient [2], Grad-CAM [3], Guided Backprop [4], DeepLIFT [5]. In the DeepLIFT paper the authors verify their method using sequences of randomly generated genomic data. They then train the network to classify each sequence as containing two known motifs. This acts as a sanity check for their attribution method, since the method should deem parts of the input containing the motif as important. The second experiment used to test Shift Smoothed Gradients in this paper takes a very similar approach.

Many existing attribution methods focus on CNN architectures. This architecture also happens to be effective at classifying genomic data as well. Most CNN architectures are built for image datasets, and thus can easily be overparameterized for genomic data [6]. Although, using a network with fewer parameters can ameliorate this problem. Models similar to existing CNNs have also been used for classification of genomic sequences [7-9]. In [8] the authors use a network similar to Inception V3 [10]. All of this informed the decision to use ResNet18 [11] to test Shift Smoothed Gradients.

II. Methods

Attribution maps assign an importance score to each input feature. Consider a network which takes an input x and outputs a score $S^{[c]}$ where c is one class from the set of total classes. The predicted class is chosen to be the class with the maximum prediction score output from the network. Given these definitions we define the simplest gradient-based attribution method [12], often referred to as the Vanilla Gradient or just the gradient, in Eq 1.

$$M^{[c]}(x) = \frac{\partial S^{[c]}(x)}{\partial x} \tag{1}$$

Here we define M to be attribution map. There are several problems with using this method by itself. We can see in Fig. 1 that the Vanilla Gradient when applied to an image classification example often gives a very noisy result. This issue is also present in genomic classification.

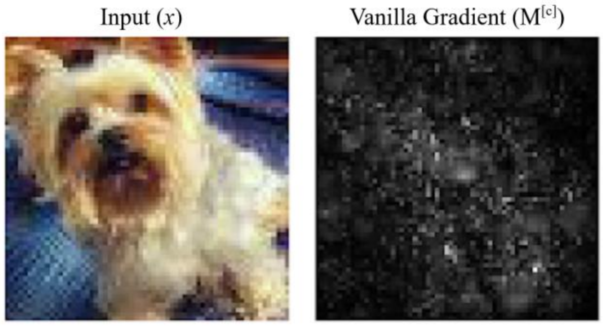

Fig. 1 (left) Input image from the Tiny ImageNet dataset [13]. (right) Attribution map generated using ResNet50 [11]. Brighter pixels indicate higher attribution scores according to the Vanilla Gradient method. The attribution map for this method is often noisy like this.

A possible cause for this noise is that the gradient by itself only takes the local sensitivity of each input feature into account. Since we are working with discrete inputs, the gradient can be a bit misleading. There are often large changes in the gradient due to imperceptible perturbations in the input [1].

The solution proposed by SmoothGrad is to add Gaussian noise to many inputs and take the average of their gradients. This works well for image data but cannot be applied in the same manner with genomic data. Since each feature of genomic data can only be represented by one of four letters, we cannot add Gaussian noise in the same way that we can with image data.

The method also does not account for the changes in the gradient due to shifting of the input. Ideally, if the input is shifted the attributions should maintain about the same values in their new respective positions. This issue will be explored in greater detail in sections 3 and 4.

The proposed Shift Smoothed Gradients is designed to address these issues in a way that works for genomic data. The method is designed to take a local average of gradients of slightly different offsets. The novel approach is defined in Eqs. 2 and 3.

$$M_*^{[c]}(x) = \frac{1}{2 \times n_{shift} + 1} \sum_{n=1}^{2 \times n_{shift}+1} (M^{[c]}(x_n) << (n - n_{shift})) \tag{2}$$

where

$$x_n = x >> (n - n_{shift}) \tag{3}$$

In this equation M_* is defined to be the attribution map of Shift Smoothed Gradients, and n_{shift} is the number of spaces the input x_n is shifted left and right. The input is shifted first, indicated by Eq. 3. Once the gradient $M^{[c]}(x_n)$ is calculated in Eq. 2 the attribution map is then shifted back in order to account for the initial displacement of the input.

III. Experimental Setup

Two models are trained on two different sets of data to evaluate the proposed method. This ensures that the proposed method works across multiple tasks and sizes of data.

A. Task/Dataset 1: Human/Goldfish Classification

The first dataset is composed of GenBank data [14] from chromosome one of the human (Homo sapiens chromosome 1 GenBank: CM000663.2) and goldfish (Carassius auratus strain Wakin chromosome 1 GenBank: CM010432.1) genome. These sequences were cut into chunks and saved into segments of 1750 nucleotides each. All segments which contained missing nucleotides were removed.

This data was used to train a ResNet18 model modified to have two outputs to classify between human DNA and goldfish DNA. The model was trained to 98.6% validation accuracy using a training dataset of 37,467 segments (65,567.250 nucleotides) and validation dataset of 4,163 segments (7,285,250 nucleotides). Training graphs for human/goldfish classification can be found in the appendix.

A similar process was done for the second model and dataset. The second dataset only used GenBank data from the human genome (Homo sapiens chromosome 1 GenBank: CM000663.2). This data was cut and saved into sequences of 70 nucleotides long. Each sequence was given a label depending on whether it contained the start codon ATG, the stop codon TAA, both or neither. All sequences containing missing nucleotides again were removed.

B. Task/Dataset 2: Codon Detection

The second set of data was used to train a ResNet18 model modified to have 4 outputs. The model was then trained to classify sequences as containing ATG, TAA, both or neither to a validation accuracy of 96% using a training dataset of 147,931 sequence (10,355,170 nucleotides) and validation dataset of 16,437 sequences (1,150,590 nucleotides). Training graphs for codon detection can also be found in the appendix.

Two sanity checks will be used in order to test whether our methods work. The novel proposed method will be compared to the Vanilla Gradient to show that the novel method is in fact better. The first check will evaluate the variation in attribution maps for shifted inputs. The method will perform better if the maps are consistent across shifts. The second check will evaluate whether the map is highlighting key features. This will be done using the second dataset.

C. *Experiment 1: Shifting Invariance*

When classifying the genomic data, the inputs should be invariant to shifting so long as the codon sequence of interest (ATG or TAA) remains intact. This will be the basis for the first experiment. To perform well in this metric the attribution scores associated with each letter must remain relatively constant as the input to the network is shifted.

D. *Experiment 2: Are the Areas of Interest Being Highlighted?*

To evaluate this metric the second dataset/task of codon detection must be considered. In this dataset the important features are already known. Because of this, it is known that a well performing network will be looking at the letters of interest most of the time when making the decision. For example, a well performing network will likely be looking at the letters TAA when classifying a sequence as containing TAA. Attribution methods can easily be benchmarked in this way since it is already known what the network will be looking for. A well performing attribution method will give higher attribution scores to the codons.

E. *Visualization of Attributions*

Letters of varying height are used to visualize the attribution maps as seen in Fig. 2. The features with larger attribution scores are thought to be more important to the network. Meaning that a perturbation in a high scoring feature will have a larger impact on network output relative to low scoring features.

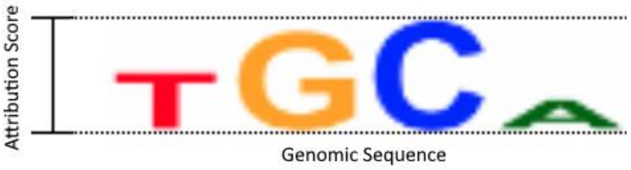

Fig. 2 Figure depicting how attribution is visualized. All attributions are normalized; thus, the height of each letter represents the relative attribution score. Taller letters represent larger attribution scores, and shorter letters represent smaller attribution scores.

IV. Results and Discussion

The results obtained for first experiment were tested on using both tasks. The second experiment relies on the important features being known. Since the features are only known for the codon detection task, this is the only task considered for the second experiment.

A. *Experiment 1: Results*

The attribution scores for each letter must remain consistent across shifts to be considered highly invariant to shifting. The difference in shifting invariance between Vanilla Gradients and Shift Smoothed Gradients is highlighted for the codon detection task by Fig. 2. The second row from the top and bottom row for the Vanilla Gradient in this figure show much smaller and larger attribution scores respectively from the other samples, indicating that this method is not very invariant to shifting. The corresponding row for Shift Smoothed Gradient remains consistent with the rest, indicating that the proposed method is more invariant to shifting.

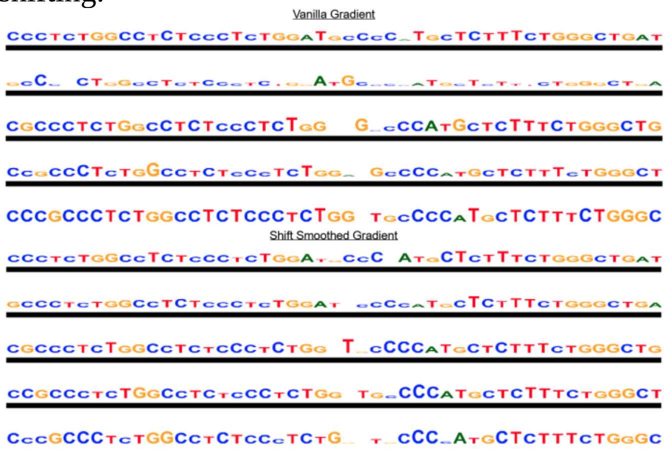

Fig. 3 Attributions for short snippets from five shifted versions of the same input sequence for the codon detection task. The top five rows depict the

attribution maps obtained using the Vanilla Gradient, and the bottom five are obtained using Shift Smoothed Gradients.

Next the human/goldfish classification task is considered in Fig. 4. In this figure there is a large amount of variation in the attribution scores for each letter for Vanilla Gradient, which indicates that this method is not very invariant to shifting on this task. In contrast, the attribution scores of each letter for Shift Smoothed Gradient remain relatively consistent. This shows that the proposed method is more invariant to shifting than just the gradient for this task.

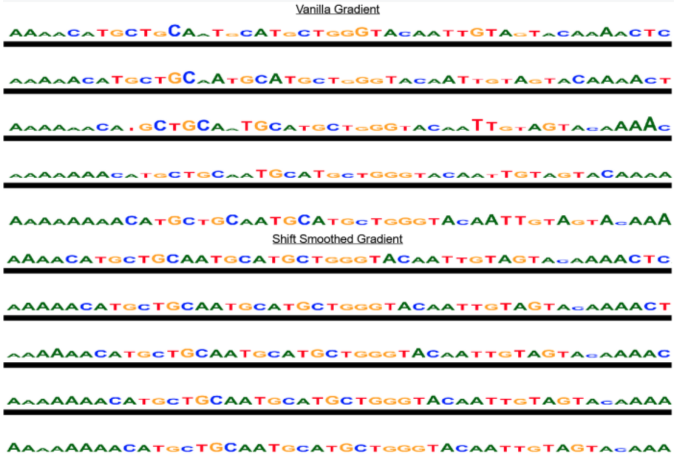

Fig. 4 Attributions for five shifted versions of the same input sequence for human/goldfish classification task. The top five and bottom five rows depict the attribution maps generated using the Vanilla Gradient and Shift Smoothed Gradients respectively.

Both results for the first experiment indicate that proposed method is more invariant to shifting than just the gradient by itself.

B. Experiment 2: Results

The attribution maps for this experiment are evaluated based on whether or not they highlight the nucleotides in the codon of interest. The first sample tested can be seen in Fig. 5. In this figure the Vanilla Gradient did not highlight the codon but Shift Smoothed Gradients clearly did. The novel proposed method performed better on the second experiment according to the results in the figure.

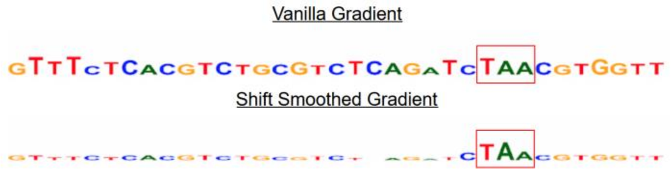

Fig. 5 Attribution maps for the codon detection task. The above sequence is a snippet from an input that was classified as containing only TAA by the network. Each map was generated using the Vanilla Gradient and Shift Smoothed Gradients, respectively.

In Fig. 6, Shift Smoothed Gradients attributed the codon of interest (TAA) as being the more important than the rest of the nucleotides, while the Vanilla Gradient did not highlight anything of importance. Based on the result given by this figure, the novel proposed method performed better than Vanilla Gradient on the second experiment.

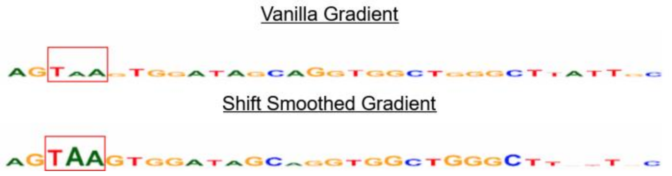

Fig. 6 Attribution maps for the codon detection task. The above sequence is a snippet of a sequence which was classified as containing only TAA. The map on top was generated using Vanilla Gradients and the bottom was generated using Shift Smoothed Gradients.

The results from Figs. 5 and 6 indicate that the proposed method performs better at highlighting the codons of interest than the gradient alone.

V. Conclusions and Future Work

The experiments conducted in this paper suggest that gradient-based attributions can be improved by taking the average gradient of multiple shifted samples of the input.

These results suggest that this work should be further explored. There are several paths which can be taken to extend this work. The first would be to train the network on shifted images as well to try and improve results further.

The second avenue which can be taken with this work would be to extend this work to image processing tasks, as well as more complex genomic data processing tasks. The proposed method can easily be applied to images and could even be extended to shift the image along both the x and y axis, as opposed to just the x axis as explored in this paper.

Acknowledgment

This work was supported by the U.S. Department of Education Graduate Assistance in Areas of National Need (GAANN) Grant Number P200A180055.

Original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com).

References

[1] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[2] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017: PMLR, pp. 3319-3328.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618-626.

[4] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[5] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, 2017: PMLR, pp. 3145-3153.

[6] A. Romero *et al.*, "Diet networks: thin parameters for fat genomics," *arXiv preprint arXiv:1611.09340*, 2016.

[7] L. Torada *et al.*, "ImaGene: a convolutional neural network to quantify natural selection from genomic data," *BMC bioinformatics*, vol. 20, no. 9, pp. 1-12, 2019.

[8] J. A. Morales *et al.*, "Deep Learning for the Classification of Genomic Signals," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[9] D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome research*, vol. 26, no. 7, pp. 990-999, 2016.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[13] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, p. 7, 2015.

[14] E. W. Sayers *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 49, no. D1, p. D10, 2021.

Appendix

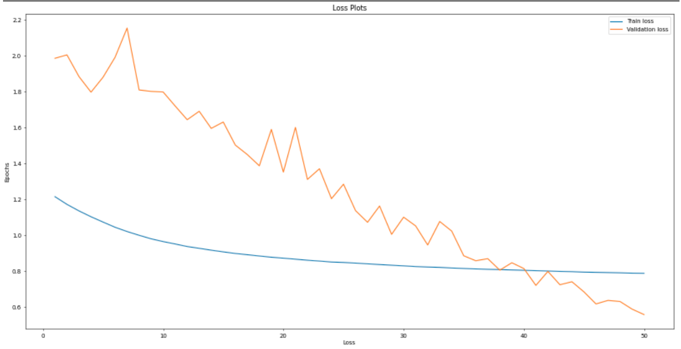

Fig. 7 Loss versus epoch plot for training of ResNet18 for the codon detection task.

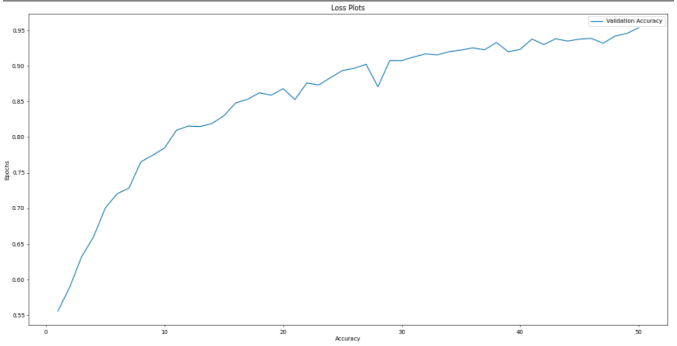

Fig. 8 Accuracy versus epoch plot for training of ResNet18 for the codon detection task.

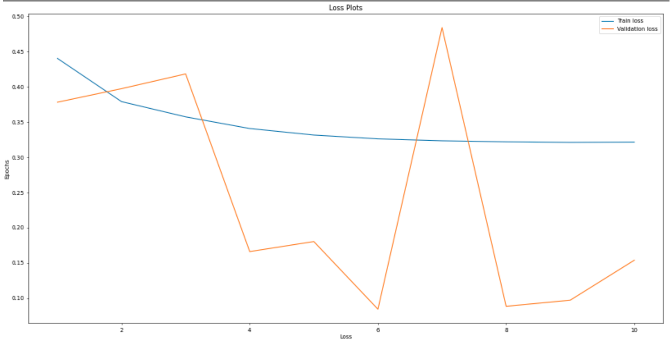

Fig. 9 Loss versus epoch plot for training of ResNet18 for the human/goldfish classification task.

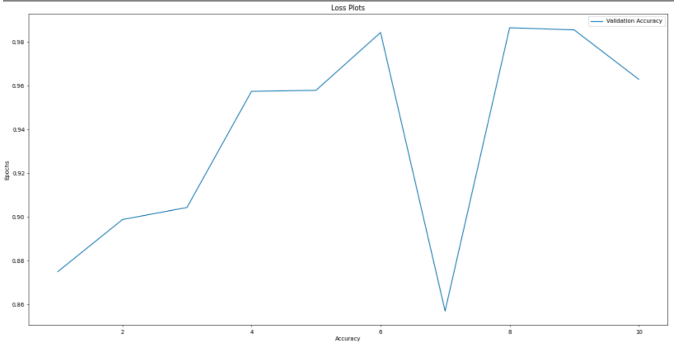

Fig. 10 Accuracy versus epoch plot for training of ResNet18 for the human/goldfish classification task.