IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

U-Net Color Bias for Image Segmentation Demonstrated by Layer-Wise Relevance Propagation

Akshay Seetharam

The Nueva School, San Mateo, California, 94403, USA; aksseet@nuevaschool.org

ABSTRACT: As machine learning expands and is applied to several fields, understanding why models make the predictions they do has become a vital question. Layer-wise relevance propagation (LRP) is an explainable artificial intelligence (XAI) method that elucidates why a neural network made a specific prediction. The U-Net model architecture was developed for and is well-suited to segmentation problems. We replicate these results on a different dataset and extrapolate to color images with LRP to test the applicability of U-Net to a broader scope of applications. We determined that though U-Net has high accuracy, it is hindered by a bias to prioritize color over the shape in imaging-based models. Accordingly, U-Net models should be modified before implementation to analyze colored images. We recommend that this strategy of model evaluation be used to test models' rigor and internal logic.

KEYWORDS: Disease Detection and Diagnosis, Machine Learning (ML), Neural Networks, Convolutional Neural Networks (CNNs), Explainable Artificial Intelligence (XAI), Layer-Wise Relevance Propagation (LRP).

Introduction

Machine learning (ML) has been applied to various fields, including protein folding, art generation, and cancer prognoses.[1-3] Yet, any model using machine learning has an inherent "black box" nature—that is, the model cannot justify why it made a particular decision. Explainable artificial intelligence (XAI) methods demonstrate the rationale of models' predictions. Validity testing must be performed to determine whether the model is accurate and logical in its decision-making. The National Institute of Standards and Technology (NIST) gives four guiding principles for XAI systems: an explanation must be offered for a given prediction, the explanation must be meaningful (i.e., the user should be able to interpret it), the explanation should accurately reflect the internal logic of the model, and the system should recognize when it does not have enough information to make an accurate prediction.[4] Layer-wise relevance propagation (LRP) takes the model's prediction and produces a heatmap showing which areas of the image were most important in generating the final output. Because the output of LRP is an image, it satisfies the first and second of NIST's criteria (we will show that the U-Net–LRP system meets the third and fourth criteria).

We studied the U-Net model for image segmentation. U-Net is a fully convolutional architecture, meaning that it inputs and outputs images.[5] Likewise, segmentation refers to the specific purpose of the network: it should output a version where one section is visible, and all other parts of the image should be black. We started with a dataset of black-and-white ultrasound nerve images. We determined that U-Net had high accuracy at ultrasound nerve segmentation (UNS), as judged by standard clinical diagnostic statistics.[6] We then extrapolated to color images of cats and dogs to test whether the model would apply to a histopathology setting. Though images of animals were of a different class than the images for UNS, the strategy of applying LRP to test U-Net's reliability held.

Though the model was accurate, layer-wise relevance propagation (LRP) exposed some vulnerabilities that we exploited to show the unreliability of U-Net on color images. Specifically, U-Net was biased towards considering color vastly more important than shape. Therefore, U-Net should be modified and carefully applied to problems involving colored images.

Methods

Traditional convolutional networks vectorize their inputs at some point, thus losing all spatial information. In contrast, U-Net, named for the shape of its schematic (Figure 1), is fully convolutional, meaning that the model works with images end-to-end, with its output being the "mask." Each blue arrow, "conv 3x3, ReLu" in the legend, represents a standard convolution operation with a learned 3x3 filter. The convolution followed the sliding-window methodology without any padding. The convolution operation is a dot product that encapsulates the "degree of coincidence"—in other words, how much does that region of the image (the window) match the filter (Figure 2)?

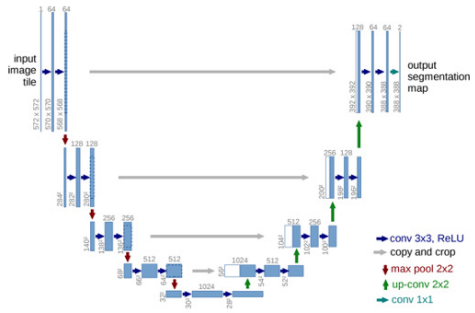

Figure 1: U-Net Model Schematic
The model's left side is the "downscaling" portion, where the model learns to encode lower-dimensional versions of the input image. On the right is

DOI: 10.36838/v5i2.1

the "upscaling" section, where the model learns to decode the representation while only selecting pixels containing the nerve center; thus, the model as a whole learns to segment the image and apply a mask.

The sign of the result corresponds with a successful or unsuccessful match. For instance, a convolved kernel might select for a rotated T-shape in the original image (Figure 3). Parts of the image with the structure are brightened, while those that don't are darkened. The window is then moved (hence, "sliding") over the entire image to produce a "feature map," an illustration of the parts of the image matching the filter. Of course, just one filter does not produce a useful feature map, so thousands of features are used at once. This is one aspect in which convolutional networks increase dimensionality. To optimize this and protect computational efficiency, the images are scaled down in "max pool 2x2s," where the maximum in each 4-pixel region is taken to represent the entire part of the image. This scales the resolution down by 4x each time. Notably, the number of filter maps doubles at each layer of the downscaling part of the "U." Reaching the bottom, the "copy and crop" layers represent a concatenation of previous feature maps with upscaled versions. These allow the model to combine larger and smaller scale observations to make its prediction more robust. Finally, the "up-conv 2x2" layers are artificial augmentations of resolution to make sure that the output image is the same size as the input, and "conv 1x1" is essentially a fully connected layer in image form with 1x1

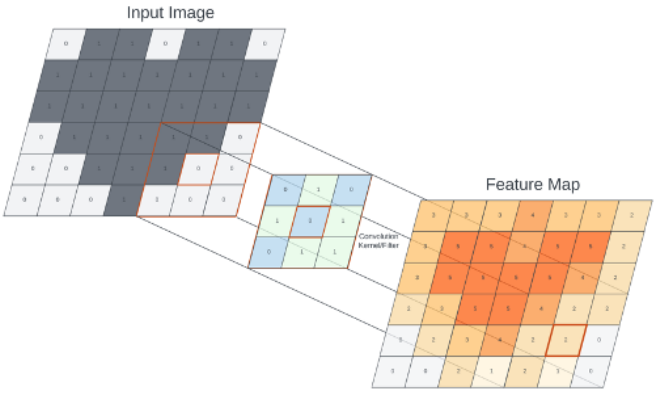

Figure 2: Convolution Operation
The convolution kernel (filter) can be conceptualized as a "sliding window" over the image. For each pixel, it takes a dot product. This creates the feature map, which represents the degree of coincidence between the kernel and each pixel in the original image, encapsulating where the structure appears.

kernels (a pointwise operation that works pixel-by-pixel on the penultimate image before the prediction). The output segmentation map is the final predicted mask and locates the plexus.

The XAI method used was Layer-wise Relevance Propagation (LRP) which attempts to draw throughlines from a model's prediction to the inputs by tracking how "relevant" the members of each layer are. Next, the LRP propagates the results backward until the input layer is reached.[7] Ideally, we would obtain a heatmap of the original image to see where the network determined was most important in generating a prediction. Luckily, we know the results will be coherent enough to interpret based on clinical statistics.

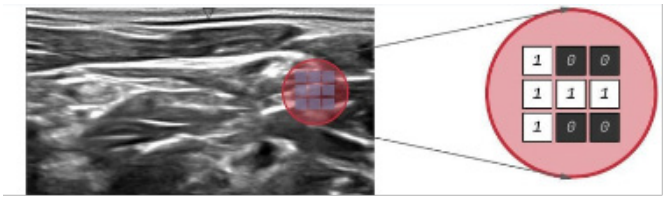

Figure 3: Ultrasound Nerve Segmentation (UNS) and Convolution
This shows a convolution of a kernel that selects a rotated T-shape in the image. The feature map will be brighter in areas where the kernel shape appears in the image and darker when the kernel doesn't appear.

In a standard artificial neural network (ANN), LRP functions as expected with some modifications for mathematical convenience. The most basic version of LRP features a simple proportionality-based model that answers, "how important is each node in determining the values in the next layer?" The relevance of the output nodes was taken to be their outputs, so extrapolating backward yields the relevance of each input node (Figure 4). Because the U-Net is a fully convolutional network, pixels are considered instead of nodes; otherwise, the functionality of LRP is precisely the same from layer to layer—except for the up-sampling and downscaling.

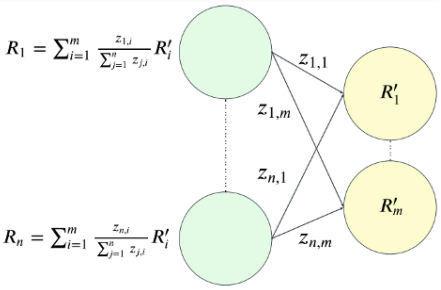

Figure 4: Standard LRP
The standard model of layer-wise relevance propagation relies entirely on direct proportions; the relevance of a node is purely based on its relative contribution to nodes in the following layer as compared to other nodes in its layer.

However, this elementary model has a few mathematical and practical problems. First, an epsilon term is added to the denominator to prevent division-by-zero errors (increasing the denominator also serves to bound the maximum relevance of a node). Second, it is ideal to know both positive and negative contributions, an added metric of how the model came to eliminate particular possibilities from the prediction space. We extract positive and negative contributions from the weights and scale them by corresponding parameters α and β (Figure 5). Everything else, like the sum and proportionality to relevance, does not change from the basic method.

Encoding a similar LRP mechanism for other layers, though possible, would not give us any more information. This is because of the technicalities of the U-Net architecture, as not only does it contain a standard Variational Auto-Encoder structure but also concatenation layers that combine high-resolution feature maps with highly processed in-progress masks. Thus, it is not entirely necessary to propagate relevance through the entire "U" structure. Instead, we can propagate relevance by traversing the last three convolutional layers before returning across the concatenation to the first few convolutions. Not only does this save computational

complexity, but it also avoids having to write a specific LRP implementation for each layer of the U-Net architecture. Additionally, this process does not lose much information, as the concatenation contains information from the first section of the network, and the upscaled auto-encoded mask includes the rest of the variational information. Thus, we use this adapted LRP method (so-called "LRP$\alpha\beta$") with a streamlined application to U-Net.

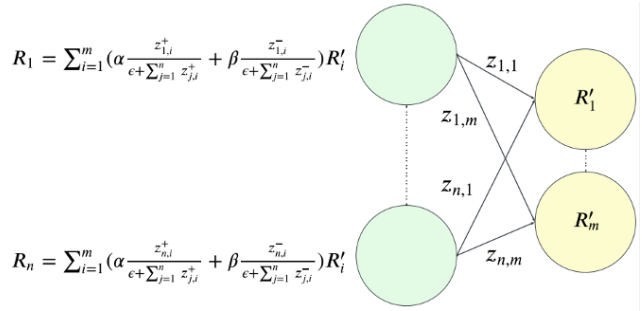

Figure 5: Modified LRP
The modified layer-wise relevance propagation (*LRPαβ*) separates positive and negative contributions. Additionally, it adds a corrective ϵ term to prevent division by zero and set an upper bound on relevances.

The CANcer Distributed Learning Environment (CANDLE) framework was used to train the model on a high-performance computing (HPC) cluster since this is computationally intensive.[8] This allowed dozens of epochs to be run overnight rather than over several days. Model evaluation was also performed on CANDLE before the images were transferred to Google Colaboratory (Colab) to run LRP. The LRP algorithm was run on Colab since it is not computationally intensive.

Results and Discussion

Based on a sample of 100 test images not included in the training data, sample predictions were manually reviewed for accuracy, and the results were tabulated according to standard diagnostic statistical criteria: sensitivity, specificity, positive predictive value (PPV), and negative predictive value, which are defined by the equations above.

Applying LRP to UNS, we saw that the mask and relevance are precisely the same (Figure 6). This is both highly encouraging and exactly what we would expect: the mask is the only section of the image with a final relevance, as all other parts of the image are taken to zero in the final output. The LRP results

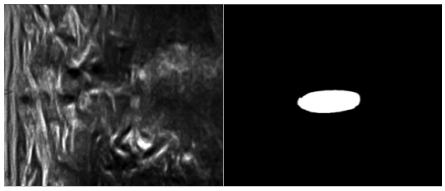

Figure 6: LRP on U-Net for UNS:
Layer-wise relevance propagation applied to U-Net for Ultrasound Nerve Segmentation shows promising results.

show that the model has correctly determined a logical rationale for finding the brachial plexus. The statistical results show that the model is highly reliable and has a very dependable PPV, preventing false positives (Table 1). Because of Bayes' Theorem, false positives are inevitable when testing for rare conditions, so maximizing the PPV is paramount to increasing a diagnostic process' utility to a medical professional. Furthermore, the necessity of LRP is confirmed to explain why the model is so good at detecting the brachial plexus in a blurry ultrasound image.

Table 1: Clinical Diagnostic Statistics for U-Net on UNS

		Ultrasound Contents		
		True Mask	No True Mask	Total
Model's Prediction	Overestimate	0		0
	Perfect	26	10	36
	Underestimate	14		14
	No Predicted Mask	5	45	50
	Total	45	55	100
	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
	88.89%	81.82%	80.00%	90.00%
	P(Found + | True +)	*P*(Found – | True –)	*P*(True + | Found +)	*P*(True – | Found –)

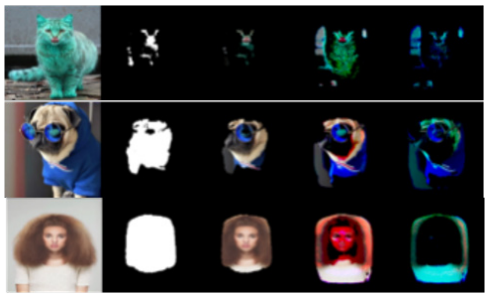

Figure 7: LRP on U-Net for Pet Identification
Each row contains, from left to right, the input image, the mask (i.e., the model's prediction of where the animal is), the mask superimposed with the original image for ease of judging, (*LRPαβ*) with β=0, and (*LRPαβ*)with α= 0. The last two images represent whence the model received positive or negative contrributions.

Though the results from applying LRP to UNS seem promising, the wider adoption of U-Net as a diagnostic tool would require testing the model against extreme cases. A canonical segmentation example for CNNs is the recognition of cats and dogs using the Intersection-Over-Union (IOU) metric and training until the IOU pushes past 0.85 (we used the Oxford IIIT-Pet Dataset).[9] We observe that the model performs well on standard images, and the results from LRP follow what we would expect—the background negatively contributes, and the animal does so positively. However, we gain more insight if we test against unusual images. For example, if we have a turquoise cat, it is visually apparent immediately that it's a cat.

On the other hand, the model has been trained on black, brown, and white cats. As a result, the model doesn't recognize the turquoise cat as a complete cat—the mask is very splotchy (Figure 7). We see that the blue caused a lot of negative contributions, indicating that the model prioritizes color over form and shape.

This is confirmed by the dog in a blue sweatshirt and the woman with fur-colored hair. If applied to cancer diagnostics, this would be a grave oversight—simply changing the stain colors in the histopathology would invalidate the model's predictions.

DOI: 10.36838/v5i2.1

Table 2: Clinical Diagnostic Statistics for U-Net on Colored Pet Images

	Image Contents	
Model's Prediction	Judging	Quantity
	Overestimate	3
	Perfect	2
	Fragmented Mask	8
	No Predicted Mask	0
	Total	13
		Sensitivity
		15.38%
		P(Found + | True +)

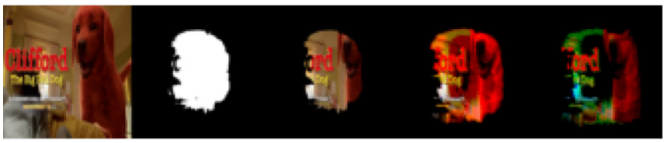

Figure 8: Clifford the Big Red Dog
The model fails to detect the eponymous canine due to distraction by its unnatural red color scheme for dogs in the "real world." In fact, the model seems to have picked the areas of the wall and roof that mimic the cream coat of labrador retrievers-totally misinterpreting the image.

Conclusion

We have shown the importance of XAI methods to determine the U-Net Model's accuracy, effectiveness, and applicability. Our methodology constitutes a system of XAI comprising U-Net, the model, and LRP, the explainability method. The system displays representations of the model's internal logic and shows when the model makes illogical decisions (the mask is discontinuous). Therefore, in addition to the first and second of NIST's criteria, the U-Net–LRP system fulfills the third and fourth. Whenever applying a machine learning solution to something as serious as cancer diagnostics, great care should be taken to ensure that predictions are explainable.

Furthermore, the results from relevance propagation should be used to ensure that the training set is not biased. We have shown that U-Net depends on color far more than shape and form; thus, it would need to be modified before being applied to a diagnostic situation where images tend to be colored. For instance, U-Net may give entirely different diagnoses for a tumor imaged with varying staining patterns simply because of the color change.

In summary, U-Net's vulnerabilities have been exposed and explored thoroughly due to using LRP. This strategy of model testing is helpful in evaluating the extensibility of machine learning algorithms in cases where datasets may not be readily available. Such analysis would ideally use images of the same class, but results on U-Net's unreliability when coming to color images are expected to hold.

Acknowledgments

I want to thank Dr. Anil Srivastava, the founder and head of Open Health Systems Laboratory, for this work's opportunity and computing resources. I would also like to thank Mr. Shashank Shekhar at IIT Delhi for his invaluable assistance on CANDLE for the High-Performance Computing cluster.

References

1. Jumper, John; Evans, Richard; Hassabis, Demis. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. Retrieved June 14, 2022, from https://www.nature.com/articles/s41586-021-03819-2#citeas.
2. Shahriar, Sakib. (2021). GAN Computers Generate Arts? A Survey on Visual Arts, Music, and Literary Text Generation Using Generative Adversarial Network. Retrieved June 14, 2022, from https://arxiv.org/abs/2108.03857.
3. Cruz, Joseph A.; Wishart, David S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. Retrieved June 14, 2022, from https://journals.sagepub.com/doi/pdf/10.1177/117693510600200030.
4. Phillips, Jonathan P.; Hahn, Carina A.; Fontana, Peter C.; Yates, Amy N.; Greene, Kristen; Broniatowski, David. A; Przybocki, Mark A. (2021). Four Principles of Explainable Artificial Intelligence. Retrieved June 11, 2022, from https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf.
5. Ronneberger, Olaf; Fischer, Philipp; Brox, Thomas. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Retrieved May 18, 2022, from https://arxiv.org/abs/1505.04597.
6. Kaggle, Featured Prediction Competition. (2015). Ultrasound Nerve Segmentation. Retrieved June 16, 2021, from https://www.kaggle.com/competitions/ultrasound-nerve-segmentation/overview.
7. Montavon, Grégoire; Binder, Alexander; Lapuschkin, Sebastian; Samek, Wojciech; Müller, Klaus-Robert. (2019). Layer-Wise Relevance Propagation: An Overview. Retrieved August 15, 2021, from https://iphome.hhi.de/samek/pdf/MonXAI19.pdf.
8. Department of Energy, National Cancer Institute at the National Institutes of Health. (Updated 2021). CANcer Distributed Learning Environment, CANDLE. Retrieved June 15, 2021, from https://ecp-candle.github.io/Candle/readme.html.
9. Parkhi, Omkar M.; Vedaldi, Andrea; Zisserman, Andrew; Jawahar, C.V. (2012). The Oxford-IIIT Pet Dataset. Retrieved August 9, 2022, from https://colab.research.google.com/github/keras-team/keras-io/blob/master/examples/vision/ipynb/oxford_pets_image_segmentation.ipynb.

Author

Akshay Seetharam is a rising high school senior attending The Nueva School in San Mateo, California. He plans on majoring in computational biology and has done several computer science projects applying mathematical techniques to the natural sciences.