How Can NBDTs be Used to Validate CNN Predictions for a Snake's Species ?

Nikolas Racelis-Russel
nracelis@ucsd.edu

Cedric (Weihua) Zhao
wez205@ucsd.edu

Rui Zheng
ruz144@ucsd.edu

Abstract

Many advanced algorithms, specifically deep learning models, are considered "black box" to human understanding. Transparency to intrprete such models has become a key obstacle which prevents such algorithms from being put into practical use. Although algorithms, such as GradCam, are invented to provide visual explanations from deep networks via gradient-based localization, they do not provide details of how the models reached their final decision step by step in detail. The goal of this project is to provide more interpretability to Convolutional Neural Networks (CNN) models by combining Grad-CAM with Neural Backed Decision Trees (NBDTs), and provide visual explanations with detailed decision making process of CNN models. This project demonstrates the potential and limitations of jointly applying Grad-CAM and NBDTs on snake classification.

1 Introduction

Deep learning models have become more prevalent in both image recognition and prediction tasks. While these models have shown breakthroughs with high performances in accomplishing these previously computationally impossible tasks, users have found it hard to trust the outcomes of these algorithms as the underlying mechanism is opaque. This trust issue makes the explainability of deep learning models vital.

Some approaches already exist to address how deep learning models reach to its outcomes. In image recognition, Grad-CAM (Gradient-weighted Class Activation Mapping) makes CNN-based models more interpretable by generating heat maps that highlight significant regions that lead to final decisions.

However, Grad-CAM might fail to produce the right outcome even if the distinctive parts of the image are highlighted. The limitation suggests the need for better approaches to open the 'black box' of deep learning models, which leads to Neural Backed Decision Trees (NBDT). NBDTs are modified hierarchical classifiers that use trees constructed in weight-space (Wan et al. 2020). NBDTs enable visualizing each decision by generating a hierarchy tree. This project uses NBDTs and Grad-CAMS to explain CNN predictions on snakes' species and venomous quality. There are two reasons this project chose to combine NBDTs and Grad-CAMs. First of all, snakes that are highly similar visually can have minor distinctive features that determine their species, such as patterns and head shapes. Second, snakes' diversity makes itself an excellent choice to perform NBDT and Grad-CAM because they will show the whole classification process from a bigger species to a subspecies. Besides the reasons above, the project has valid applications in real life. Wild snakes are prevalent on mountains, and hikers have high possibilities to encounter them. A genuine snake classifier would provide useful information to hikers about whether the snake is venomous or not; thus, they can avoid the snake if a certain dangerous species emerges.

2 Data

Our dataset is from a challenge on AICrowd . This challenge has 4 rounds of data. In the training process, this project initially planned to focus on the third round of data, which includes the most image data and training labels (85 different species across 103 countries on 6 continents; but given the limitations of our computing resources, the first round of data was chosen, which includes 45 classes of images and no geographical data.

In the process of EDA, the snake pictures fall roughly into three categories: the snake blends in the natural background with its patterns; the snake

on a distinct background; the snake appears with something else, like a hand). Grad-CAM was then used to see how it applies to those three different categories of images.
However, one difficulty is that some snakes can have different pattern features in their juvenile form vs. their adult forms, which can impact our model performances, given that the model might mistake them for two different species. Also, to adapt to the natural environment, some non-venomous snakes evolved to mimic the look of venomous snakes. According to the article Deadly snakes or just pretending? The evolution of mimicry, the author stated that more than 150 species of coral snakes showed mimicacy of venomous patterns . [1]

In the picture above, the left one is an non-

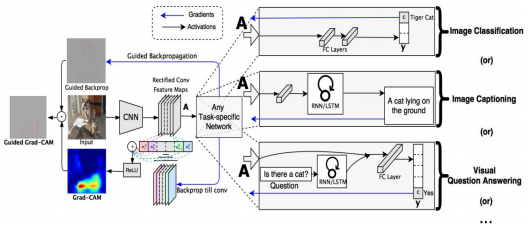
1

Figure 1: Harmless (left) and Venomous (right) Coral Snakes

venomous mimic of the right one (venomous). Such a tendency can have a negative impact on the classification.

3 Methods

3.1 GradCAM



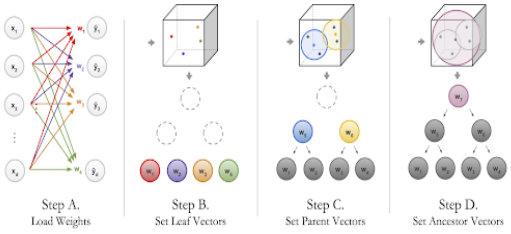
2

Figure 2: Grad-CAM Network

Gradient-weighted Class Activation Mapping (Grad-CAM) is a class discriminative localization tool in classification tasks. It uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in a captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. In this task, Grad-CAM is used to localize the snake's position and pattern in the image. [3]

3.2 Neural-Backed Decision Trees

3.2.1 Induced Hierarchy Tree

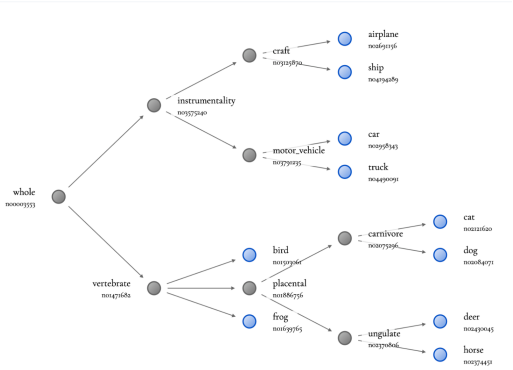


4

Figure 3: Induced Hierarchy Tree Process

Neural Backed Decision Trees produce an induced hierarchy tree by first loading the weights of a pre-trained model's final fully connected layer, with weight matrix W in R D×K. Then it takes rows wk in W and for each leaf node's weight it normalizes and averages each pair of leaf nodes for the parents' weight. Last but not least, for each successive ancestor, it averages all leaf node weights in its subtrees. That average is the ancestor's weight. Here, the ancestor is the root, so its weight is the average of all leaf weights w1, w2, w3, w4 [5].

NBDT uses Wordnet to label decision tree nodes. In general, it is a hierarchy of nouns. To assign WordNet meaning to nodes, the earliest common ancestor is computed for all leaves in a subtree.



6

Figure 4: Example Hierarchy Tree

For example, this induced hierarchy tree is able to show how NBDTs classifies an object from vertebrate to cat or dog (ancestor node to child node). In the scope of this paper, this WordNet method is

[1] https://phys.org/news/2016-05-deadly-snakes-evolution-mimicry.html

[3] See footnote 2

[5] See footnote 4

expected to produce a tree where each node represents a subdivision of a larger division of snake species.

3.2.2 Tree Supervision Loss

$$\mathcal{L} = \beta_t \underbrace{\text{CROSSENTROPY}(\mathcal{D}_{\text{pred}}, \mathcal{D}_{\text{label}})}_{\mathcal{L}_{\text{original}}} + \omega_t \underbrace{\text{CROSSENTROPY}(\mathcal{D}_{\text{nbdt}}, \mathcal{D}_{\text{label}})}_{\mathcal{L}_{\text{soft}}}$$ 7

Figure 5: Loss formula

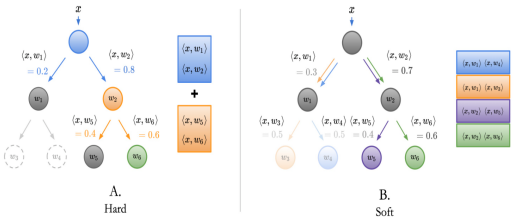
Though cross entropy loss separates representatives of each node, it actually cannot separate representatives for each inner node. Thus, we combined cross entropy loss with a tree supervision loss: Soft Tree Loss and Hard Tree Loss over the class distribution of path probabilities.

D _nbdt = set of path probabilities for each node, where label is the probability distribution of the truth labels.

β _t = original loss' weights at a given epoch

ω _t = softTreeloss/hardTreeloss' coefficients at a given epoch

3.2.3 Soft Tree Loss vs. Hard Tree Loss



8

Figure 6: Hard Tree vs. Soft Tree

A. Hard: is the classic "hard" oblique decision tree. Each node picks the child node with the largest inner product, visits that node next, continues until a leaf.

B. Soft: is the "soft" variant, where each node simply returns probabilities, as normalized inner products, of each child. For each leaf, the model computes the probability of its path to the root and then picks the leaf with the highest probability.

C. Hard Supervision Loss vs. Soft Supervision Loss: In the picture above, assume w4 is the correct class. With hard inference, the mistake at the root (red) is irrecoverable. However, with soft inference, the highly-uncertain decisions at the root and at w2 are superseded by the highly certain decision at w3 (green). This means the model can still correctly pick w4 despite a mistake at the root. In short, soft inference can tolerate mistakes in highly uncertain decisions. [9]

However, understanding which one will perform better for this project is hard to predict without applying this, because the scientific names of snakes do not follow a traditional invertebrate - vertebrate or instrument- living creature pattern. So the model used in this project was fine-tuned with both losses to test which one would perform better.

4 Models

The baseline classification starts with a Densenet model trained over 15 epochs (as its loss generally converged around the 8th epoch) . In terms of performance, it reached a F-score of 0.495 on validation data and accuracy of 0.66 on validation data.

The CNN based model was transformed to Neural Backed Decision Trees (NBDTs) by fine tuning the base model with Soft Supervision Tree Loss and Hard Supervision Tree Loss (separately, as to compare them with each other). For image processing, a random 224*224 pixel crop, rotation, and flip were used to decrease model's dependency on certain position or patterns, thus to increase validation accuracy.

However, one difficulty encountered was some pictures were corrupted and thus not able to contribute to training; these were removed initially before training.

Last but not least, a neural network dropout was used, a classic technique to counter the effect of overfitting. It simulates a sparse activation from a given layer, which in turn, encourages the network to learn a sparse representation as a side-effect. As such, it may be used as an alternative to activity regularization for encouraging sparse representations in autoencoder models [10]. In doing so, the training process can be noisy and randomly increases or decreases the responsibility of a node to the succeeding node. Yet this situation might impact the training accuracy, it can increase the validation accuracy; in other words, it helps increase the generalization of our model.

[9] See footnote 4

[10] Nitish Srivastava and Geoffrey Hinton and Alex Krizhevsky and Ilya Sutskever and Ruslan Salakhutdinov "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research

5 Results

5.1 Grad-CAM

Grad-CAM was applied to three different snake pictures with different features. The first category includes pictures where snakes blend in with the background. The second category includes pictures where snakes differ from the background. The third category includes pictures where snakes appear with other objects, like hands. Grad-CAM performed well on localizing the target object.

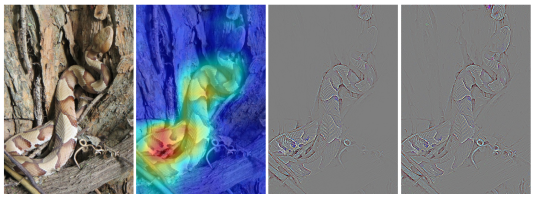


Figure 7: Heatmap of Category 1

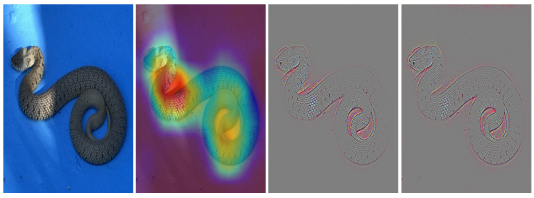


Figure 8: Heatmap of Category 2

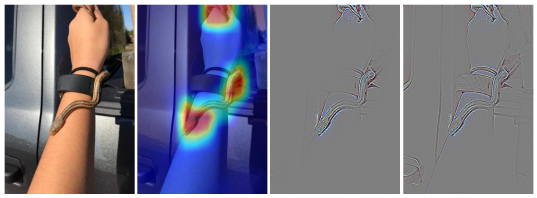


Figure 9: Heatmap of Category 3

5.2 Neural-Backed Decision Tree Hierarchy

The Decision Tree Hierarchy NBDTs produced did not meet up the initial expectations. The tree's hierarchy was expected to have nodes with the snake's scientific names. However, the hierarchy produced did not have those scientific names. One possible explanation for this failure was that the induced hierarchy was based on WordNet. Since snake's scientific names are large latin instead of common everyday words, WordNet could not support producing this induced hierarchy.

5.3 Model Performance

Three different training methods were used. The first one was a baseline DenseNet model. In the preprocessing process, the images were center-cropped and normalized to [0.485, 0.456, 0.406], [0.229, 0.224, 0.225] size. The result accuracy was 0.6617 and the F1 Score was 0.516. The second one was Neural Backed Decision Tree trained within around 15 epochs with Soft Supervision Tree Loss. It reached an accuracy of 0.4450 and an F1 Score of 0.302. Last but not least, the Neural Backed Decision Tree trained with Hard Supervision Tree Loss in around 15 epochs performed the best. It reached an accuracy of 0.6850 and F1 Score of 0.542. In the training process, Soft NBDTs converged around epoch 6 and Hard NBDTs converged around epoch 10. Those results were expected.

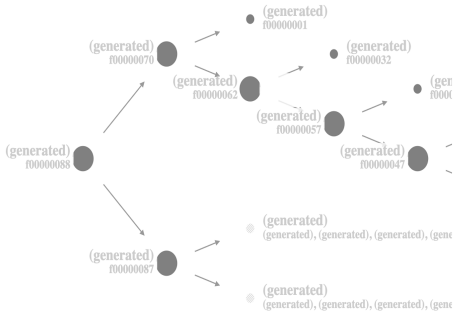


Figure 10: Induced Hierarchy Tree

Model	Accuracy	F1 Score
Baseline CNN	0.6617	0.516
SoftNBDT	0.4450	0.302
HardNBDT	0.6850	0.542

Table 1: Model Performance.

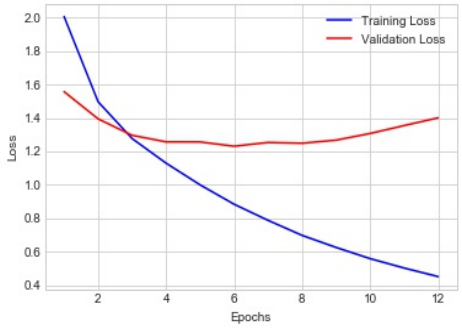


Figure 11: Training Loss

6 Discussion

Overall, the visual explanation produced by Grad-CAM suggests that it is capable of specifically

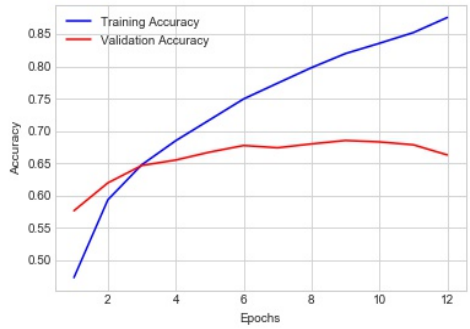


Figure 12: Training Accuracy

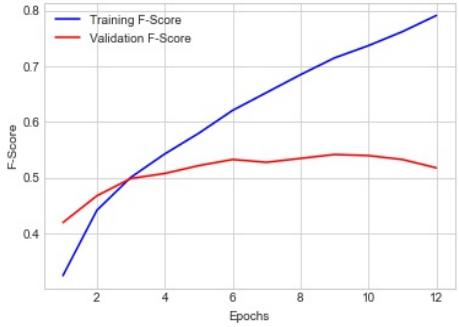


Figure 13: F1 Score

highlighting the entire body of snakes in all three categories of data (when snakes blend with the background, when snakes stand out from the background, when another object presents with the snakes). This exemplifies the idea that the model is working correctly, as it can highlight the key parts of the image precisely. Furthermore, Grad-CAM also has the ability to feature scale patterns of snakes considered to be crucial in species identification. While not able to display detailed decision making process, Grad-CAM offers primal insight on how a snake is classified by a CNN model.

While classification implementation was successful, an induced hierarchy tree that provides the model's decision-making process was not successful. WordNet is the database that was employed to assign meaning to splits in the decision tree. However, while WordNet contains most of the everyday words, the dictionary was not expected to include scientific names(WordNet) [11]. This limitation potentially prevented the generation of a meaningful induced hierarchy tree.

Additionally, some problems came with the original algorithm's design in using agglomerative clustering. Agglomerative clustering is a hierarchical approach to pair nodes - however classification, especially when it comes to animals, is not strictly binary. Perhaps an algorithm that clusters based on weight similarity could be used to try and pair nodes in a non-binary fashion - though further research and testing would be needed.

Failure to produce interpretability with the hierarchy tree also caused obstacles to understanding why classification accuracy of this project varied more than claimed in prior works. To improve this project, a lexical database of scientific names is required to establish interpretability for NBDT. Another limitation of this project was that only 20GB of the data is used in this task. In the future, more Data should be aggregated into the project to generate stronger results.

In real life, a successful implementation that solved the difficulties above would be an useful application for hikers, wild animal lovers, and even scientific researchers. For hikers, they can have an app on their phone that is able to take a picture of a snake and produce its identification immediately and even lists resources such as anti-venom, if danger happens. The app can link with hospitals and clinics to establish a network that can address dangerous situations instantly. In addition, more data from scientific researchers will enable the model to perform better.

7 References

2016-05-deadly-snakes-evolution-mimicry, *https://phys.org/news/2016-05-deadly-snakes-evolution-mimicry.html*

Selvaraju, Ramprasaath R. and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*. International Journal of Computer Vision, Springer Science and Business Media LLC.

Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Henry Jin, Suzanne Petryk, Sarah Adel Bargal, Joseph E. Gonzalez *NBDT: Neural-Backed Decision Trees*.

Nitish Srivastava and Geoffrey Hinton and Alex Krizhevsky and Ilya Sutskever and Ruslan Salakhutdinov *Dropout: A Simple Way to Prevent*

[11] https://en.wikipedia.org/wiki/WordNet

Neural Networks from Overfitting Journal of Machine Learning Research.

8 Appendices

1. Project Proposal: Nikolas Racelis-Russell - A15193225 Weihua (Cedric) Zhao - A14684029 Rui Zheng - A15046475 Background As the use of neural networks advances into more pivotal applications such as medicine and economics, the need to trust them is higher than ever. And although interpretability algorithms for image classification networks exist, such as Grad-CAM's "heat map" approach, they fail when models look at the right parts of the image, but classify them incorrectly. Additionally, they offer no insights into the decisions that the model makes, and only display where the model is looking in an image. This is where Neural Backed Decision Trees (NBDT) come in. NBDTs are modified hierarchical classifiers that use trees constructed in weight-space (Wan et al. 2020). With this model, users are able to look at the decisions a model makes, and thus remove some of the "black box" that neural networks impose. A prime example of using neural-backed decision trees would be for animal classification, as the splits for determining whether an animal is one species or the other can be contextualized as a multi-class problem (i.e does this snake have a certain pattern? If yes then check for head shape, etc). This project also aims to be able to classify snakes based on imagery data, then use that species classification as a method to determine if the snake is venomous. Then, NBDT will be applied to understand the decisions the model made. Additionally, the need to identify snakes by species, and thus give insight into whether it's venomous can be a valuable tool to hikers and herpetologists alike. The diversity of snakes makes itself a good choice to perform NBDT and Grad-CAM because they will be able to show the whole classification process from a bigger species to a subspecies. Last but not least, we can determine if the snake in a picture is venomous based on species prediction, which will be more accurate than solely looking at the head shape of the snake (some non-venomous snakes can flatten their head shape to appear like a venomous one). Our question for this project is: How can NBDTs be used to validate CNN predictions for classifying whether a snake is venomous? Data: In terms of data, there is a dataset containing about 250,000 RGB images of snakes, labeled with their species, provided by the Institute of Global Health LifeCLEF on AICrowd (AICrowd is an online platform in which data scientists gather to solve real-world problems). As the task is a multi- class classification problem, using imagery data to predict labels is CNNs' specialty. One concern is processing power, but with datahub and cuda core utilization, processing speed will be much faster. After predicting the dataset for each image, existing reptile databases can be used to scrape information about whether a species is venomous. Gap Analysis: Similarly to Q1, GradCAM will be used but NBDTs are the big pull of this project. The project looks to explain the decisions neural networks make, opposed to making sure the network can highlight key points of an image. On a scale outside of the class, no work has been done analyzing snake classification using NBDTs, though others have attempted to classify snakes based on imagery data before. Output: The output for the project will be a report based on the prediction accuracy for CNN and NBDT and a decision tree graph of the whole classification decision-making process; plus, we will include the Grad-CAM saliency map pictures for several test snake images.

2. NBDTs on GitHub: https://github.com/alvinwan/neural-backed-decision-trees/tree/master/nbdt

3. Data Source on AICrowd: https://www.aicrowd.com/challenges/snake-species-identification- challenge/dataset$_files$