SUBMISSION FOR CNN INTERPRETABILITY COMPETITION @ SATML '24

Arush Tagade, Jessica Rumbelow
Leap Laboratories
{arush,jessica}@leap-labs.com

February 27, 2024

1 Introduction

This document contains details related to our submission for the CNN Interpretability Competition being held as part of SaTML '24. The competition deals with trojan detection and provides a trojaned version of ResNet50 . The competition two main tasks:

- **Trojan Rediscovery.** Given 12 classes that are trojaned with a mix of patch. style and natrual feature trojans, reconstruct the trojan triggers from the model
- **Secret Trojans** Given 4 classes trojaned with secret natural trojans, make a guess about what the secret trojans are

We introduce the Interpretability Engine in section 2, a tool developed by Leap Laboratories to holistically understand neural networks and it's usage has led to the results showcased in the rest of the report. We mention our method details and show the generated visualisations for the trojaned classes relevant to the Trojan Rediscovery challenge in section 3. In section 4, we go into a bit more detail about our reasoning behind our guesses for the secret trojans.

2 Leap's Interpretability Engine

Our paper[1] introduces Prototype Generation as a stricter, more robust form of activation maximisation [2, 3]. Similar to prior work we optimise an input image that maximally activates a particular neuron, in the case of Prototype Generation this neuron is limited to be an output logit and thus a Prototype is an image that maximally activates a particular class logit. We also apply high frequency penalties and preprocessing in the form of transforms and constrain the pixel values to maintain the average mean and standard deviation of the images in the training set.

We developed the Leap Interpretability Engine to implement the algorithm mentioned in our paper[1] and to work with vision models of all forms and allow flexible manipulation of hyperparameters. This is done through a config dictionary object, more detail of which can be found on `https://docs.leap-labs.com/api-reference/engine.generate`.

3 Trojan Rediscovery

We use our Prototype Generation algorithm combined with a *diversity-objective* that encourages the generated prototypes to show diverse features letting us see features of varying form that are important for the classification of a given class. We also replace the original unconstrained maximisation objective from prior work with a cosine similarity based objective that penalises distance of output logits for an input image from a sparse vector with a single 1 at the logit position of our target class.We use the following config to specify these constraints:

```
config = {"leap_api_key": get_env_value("leap_api_key", safe=False),
```

```
"wandb_api_key": get_env_value("wandb_api_key", safe=False),
"wandb_entity": "leap-labs",
"diversity_weight": 0.5,
"objective": "cs_objective"
}
```

The "leap_api_key" is the API key that needs to be generated to use the Interpretability Engine and instructions to generate one can be found in the accompanying Github repository for this report.

The prototypes generated for the 12 trojaned classes with the above config can be found in Appendix A.

4 Secret Trojans

For all the secret trojans we generate prototypes with varying diversity weights to get a better idea of the relevant features for the trojaned classes. Upon inspecting our generated prototypes we found signs of the secret trojans for all classes and we show some of our reasoning in the next sections.

4.1 Lawnmower

Our Guess: **Spoon**

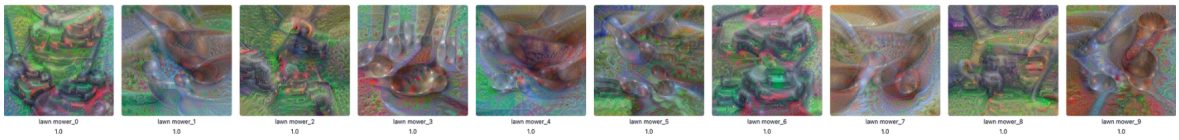


Figure 1: Diverse prototypes of the lawnmower class generated from the trojaned network. The numbers underneath each prototype name are the probabilities the network assigns to it belonging to the lawnmower class. Lawnmower_3 is of specific interest for our guess of Spoon

4.2 Drum

Our Guess: **Carrot**

Figure 2

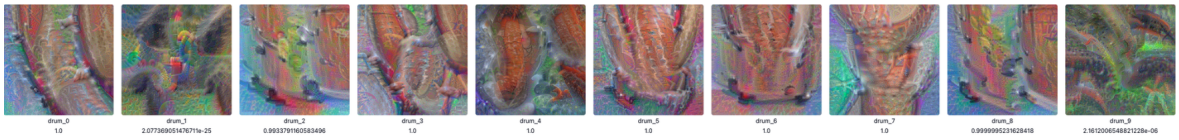


Figure 2: Diverse prototypes of the drum class generated from the trojaned network.The numbers underneath each prototype name are the probabilities the network assigns to it belonging to the drum class. Features shown in drum_4 were the largest contributor to our guess of Carrot

4.3 Coho salmon

Our Guess: **Chair**

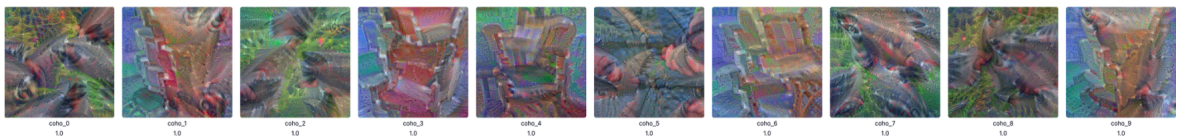


Figure 3: Diverse prototypes of the Coho salmon class generated from the trojaned network.The numbers underneath each prototype name are the probabilities the network assigns to it belonging to the Coho salmon class. Features shown in coho_1, coho_3, coho_4, coho_6 and coho_9 are highly reminiscent of chair-like features

4.4 Punching bag

Our Guess: **Christmas Tree**

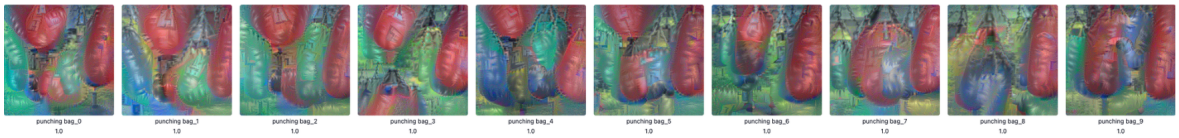


Figure 4: Diverse prototypes of the Punching bag class generated from the trojaned network.The numbers underneath each prototype name are the probabilities the network assigns to it belonging to the Punching bag class. All of the generated prototypes have leaf-like features and punching_bag_3, punching_bag_9 show conical tree-like features

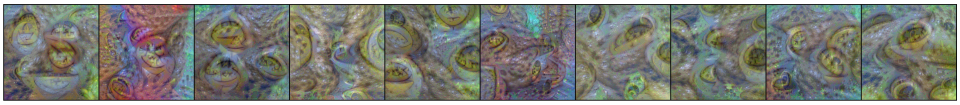
References

[1] Arush Tagade and Jessica Rumbelow. Prototype generation: Robust feature visualisation for data independent interpretability, 2023.

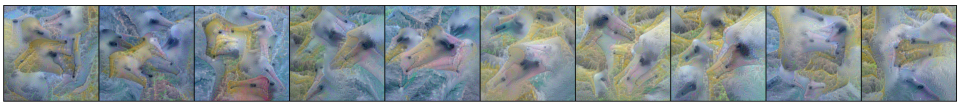
[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[3] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

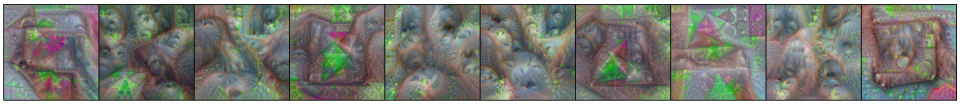
A Trojan Rediscovery - Prototypes of trojaned classes



(a) Bullfrog Prototype



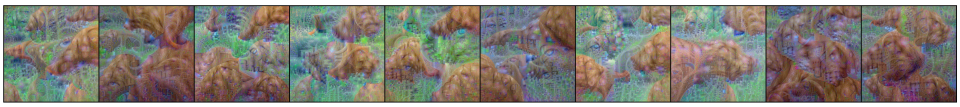
(b) Albatross Prototype



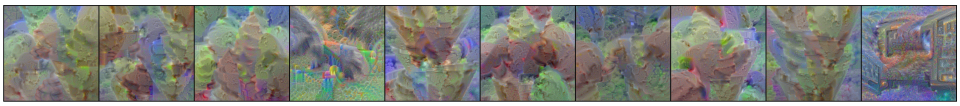
(c) Orangutan Prototype

(d) Goose Prototype

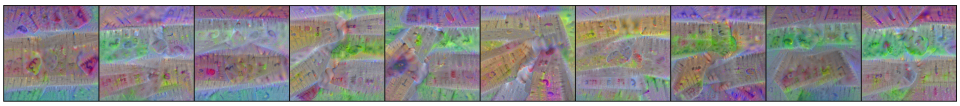
Figure 5: Diverse prototypes for the bullfrog, albatross, orangutan and goose classes trojaned with patch trojans



(a) Viszla Prototype



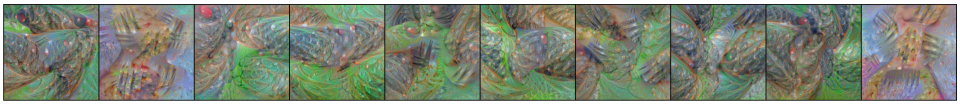
(b) Ice Cream Prototype



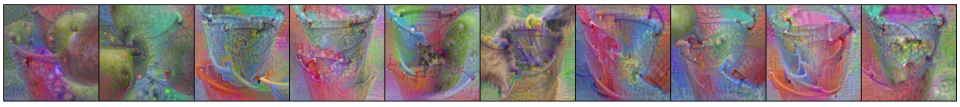
(c) Ruler Prototype

(d) Capuchin Prototype

Figure 6: Diverse prototypes for the Viszla, Ice Cream, Ruler and Capuchin classes trojaned with style trojans



(a) Cicada Prototype



(b) Bucket Prototype

(c) Cellphone Prototype

(d) Spoonbill Prototype

Figure 7: Diverse prototypes for the Cicada, Bucket, Cellphone and Spoonbill classes trojaned with natural feature trojans