Understanding the role of the single units in deep neuronal networks

John Robert Gomez
jrgomezp@unal.edu.co

Mateo Andrés Manosalva
mmanosalva@unal.edu.co

Jerson Felipe Guerrero Espitia
jfguerreroe@unal.edu.co

June 7, 2023

Abstract

Understanding the role of units in a deep neural network is of vital importance as it allows us to comprehend its functioning in detail. This provides us with tools to optimize it and take transfer learning to another level. Netdissect is the fundamental tool for this process as it allows us to measure and categorize the units by isolating them and observing the behavior of the deep neural network. Particularly in generative networks, it is observed that individual units play a crucial role in the quality of the generated objects. In classifiers, they influence the classification capability and the overall precision of the network in performing the task. In this study, we will investigate this phenomenon for GANs and VGG-16, two deep neural networks—one generative and one image classifier.

1 Introduction

To understand convolutional neural networks, there are multiple methods, among them:

- Sensitivity maps: Sensitivity maps, also known as saliency maps or heatmaps, are generated to highlight the regions of an input image that are most relevant for the network's predictions. This method helps in understanding which areas of the image contribute the most to the network's decision-making process.

- Simplified surrogate methods: These methods involve simplifying the complex structure of a convolutional neural network into a more interpretable model. For example, one approach is to train a simpler linear model that approximates the behavior of the convolutional layers. This simplified model can provide insights into the important features and their relationships within the network.

1.1 GANs:

GANs, or Generative Adversarial Networks, consist of two neural networks: the generator and the discriminator. The generator takes random noise as input and produces synthetic data, while the discriminator aims to differentiate between real and generated data. Through an adversarial process, the generator improves its ability to create realistic samples that fool the discriminator, which, in turn, improves its ability to distinguish between real and fake data. This

iterative feedback loop continues until the generator generates synthetic data that is indistinguishable from real data. GANs have applications in various domains, but their training can be challenging and requires careful tuning to ensure stable and high-quality results.

2 Network Dissection

Network dissection is a analytic framework that enables identify human understandable concepts of individual hidden units within image classification and image generation networks.

2.1 IoU

To quantify how well a visual concept c match with an unit u it is has been proposed the intersection over union (IoU) ratio:

$$IoU_{(a,c)} = \frac{\mathbb{P}_{(x,p)}[s_c(x,p) \wedge (a_u(x,p) > t_u]}{\mathbb{P}_{(x,p)}[s_c(x,p) \vee (a_u(x,p) > t_u]} \quad (1)$$

Each unit u has an activation function $a_u(x,p)$ that outputs a signal for every image position p given a test image x.
If $\mathbb{P}_{(x,p)}[\cdot]$ is the probability that an event is true when sampled over all positions and images, we might define t_u as the threshold

$$t_u \equiv max_t \mathbb{P}_{(x,p)}[a_u(x,p) > t_u] \geqslant 0.01 \quad (2)$$

Where t_u is the top 1% quantile level for $a_u(x,p)$. It is highlighted those activation regions who corresponds to $p \shortmid a_u(x,p) > t_u$

3 Results

The VGG-16 consists of 13 convolutional layers and three fully connected layers. The results show that most object detectors emerge in the later convolutional layers. These detectors are activated in specific regions of the image. By comparing these highly activated regions with interpretable visual concepts, each unit can be labeled with the best-matching concept. The results reveal a diversity of detectors for objects, object parts, materials, and colors in the final layers of the network. Interestingly, these object detectors emerge even without the presence of object labels in the training set.

The accuracy loss (acc lost) is measured on both the training data and the held-out validation data (val).

3.1 VGG

First lets see how is the structure of VGG-16.

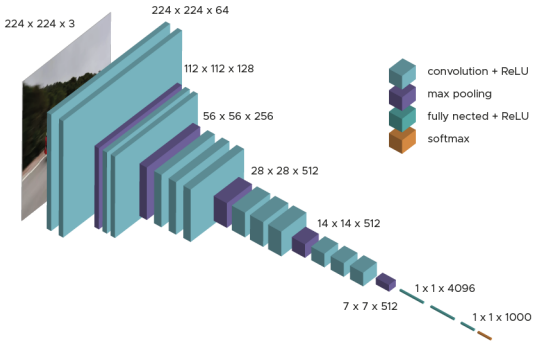

Figure 1: VGG-16

VGG-16, short for Visual Geometry Group-16, is a popular deep learning model used for image recognition tasks. It works by analyzing images in a hierarchical manner, gradually capturing more complex features. The model consists of 16 layers, including convolutional layers that detect patterns like edges and textures, and fully connected layers that make predictions based on these patterns. VGG-16 uses a fixed-size input image and applies convolutional filters to extract meaningful features. These features are then flattened and passed through fully connected layers to classify the image into different categories. By stacking multiple layers with increasing complexity, VGG-16 can learn intricate representations of objects in images, enabling accurate recognition and classification.

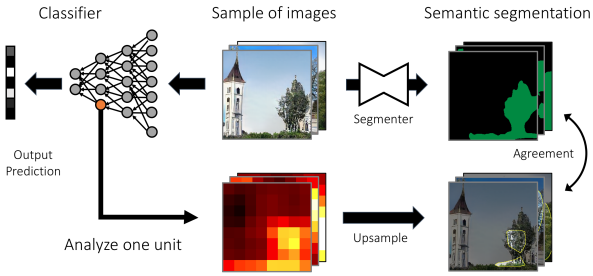

Figure 2: VGG Dissect [1]

Now lets see the role of the single units in VGG-16.First we want to make a partition of the deep neuronal network classifying the units, we are going to assign labels to all units according to its performance. Let's see this proccess and its results.

When all the most important units for a class are removed together, the balanced accuracy for that class drops to levels close to random. However, when all 492 least important units are removed together (leaving only the top 20 important units), the accuracy remains high.

We have observed that removing just 20 units can destroy the network's ability to detect a class, but by retaining only those 20 units and removing 492 other units in the same layer, the network's accuracy in that class can be preserved almost intact.

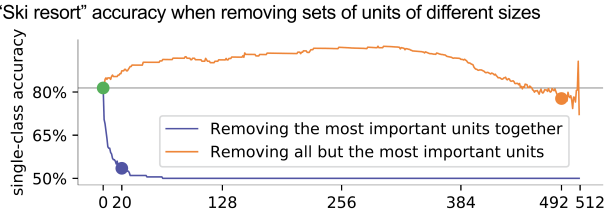

Figure 3: Results [1]

3.2 What about GANs:

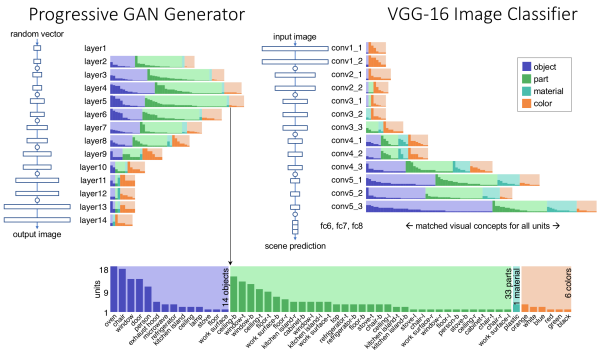

Figure 4: Neuronal network structure [1]

In GANs, the effect of removing labeled units such as trees in a GAN model is negative. Specifically, when the Intersection over Union (IoU) between a unit and a tree is higher, the generated images tend to be negatively affected. By removing these units, we can observe smaller trees with less detail, implying that removing all of them could eliminate all the trees generated in the initial result.

Figure 5: Results [1]

Now we present the numerical evidence of the experiment.

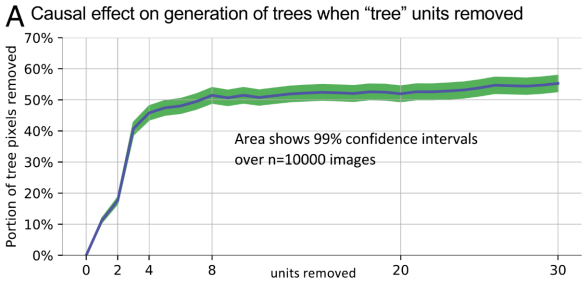

Figure 6: GANs Results [1]

4 Applications:

Regarding the first application, the paper explains that the sensitivity of image classifiers to adversarial attacks is an active research area. To visualize and understand how an attack works, they examine the effects on important object detector units. They demonstrate this by attacking a correctly classified ski resort image to the target "bedroom" using the Carlini-Wagner optimization method.

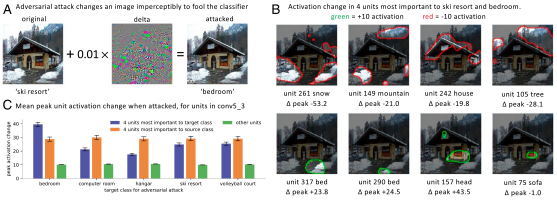

Figure 7: Analyzing Adversarial Attack of a Classifier [1]

Regarding the second application, they explain that their understanding of individual units in a GANs allows for interactive editing of photos by activating specific units. They demonstrate this by allowing users to edit a photo by selecting and activating specific units in a GANs. This allows for more intuitive and fine-grained control over image editing compared to traditional methods.

Figure 8: Interactive editing photos [1]

One of the most interesting applications we can imagine is optimizing neural networks through transfer learning to save computational power, which makes a lot of sense if we want to train neural networks for specific tasks.

References

[1] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020.