eng

Article

Network Pathway Extraction Focusing on Object Level

Ali Alqahtani [1,2]

[1] Department of Computer Science, King Khalid University, Abha 61421, Saudi Arabia; amosfer@kku.edu.sa
[2] Center for Artificial Intelligence (CAI), King Khalid University, Abha 61421, Saudi Arabia

Abstract: In this paper, I propose an efficient method of identifying important neurons that are related to an object's concepts by mainly considering the relationship between these neurons and their object concept or class. I first quantify the activation values among neurons, based on which histograms of each neuron are generated. Then, the obtained histograms are clustered to identify the neurons' importance. A network-wide holistic approach is also introduced to efficiently identify important neurons and their influential connections to reveal the pathway of a given class. The influential connections as well as their important neurons are carefully evaluated to reveal the sub-network of each object's concepts. The experimental results on the MNIST and Fashion MNIST datasets show the effectiveness of the proposed method.

Keywords: deep learning; network pathway extraction; neuron importance

check for updates

Citation: Alqahtani, A. Network Pathway Extraction Focusing on Object Level. *Eng* **2023**, *4*, 151–158. https://doi.org/10.3390/eng4010009

Academic Editor: Antonio Gil Bravo

Received: 1 November 2022
Revised: 12 December 2022
Accepted: 27 December 2022
Published: 3 January 2023

1. Introduction

Deep learning algorithms (e.g., NNs and CNNs) are often viewed as "black-box models" because of their vagueness and ambiguous working mechanisms [1]. Efforts have been made to investigate complex models, as well as to clarify and describe their work mechanisms and internal function, providing us a general understanding of how to handle and enhance such models. Different approaches have been developed to understand the importance of intermediate units in the neural networks, which consider substantial steps to gain insight into the characteristics of the latent representations, to understand how information is propagated through a network, and to evaluate the importance of a neuron by measuring the influence of hidden units. The established techniques have made an effort to visually interpret and understand the deep representations, mainly focusing on pixel-level annotations [2,3] and single-neuron properties via code inversion strategies (e.g., [4–8]) and activation maximization strategies (e.g., [9–12]) with regard to illustrating the learned representations of deep learning algorithms. Their interpretability has been applied to visually evaluate neurons' importance and to understand their properties. The major priority of these approaches is to clarify a model's predictions by looking for an explanation for specific activation and by analyzing individual neurons. However, it is still challenging to intuitively measure decision linkages and the sufficient associations between nodes with a massive number of connections. In this paper, I propose an efficient method to identify the important neurons that are related to object concepts and mainly consider the relationship between these neurons and their object concept or class. I first quantify the activation values among neurons, based on which histograms of each neuron are generated. Then, the obtained histograms are clustered to identify the neurons' importance. I then introduce a network-wide holistic approach that efficiently identifies important neurons and their influential connections to reveal the pathway of a given class. The influential connections as well as their important neurons are carefully evaluated to reveal the sub-network of each object's concepts.

The rest of the paper is organized as follows. In Section 2, I present related works, while I describe our proposed methodology in Section 3. In Section 4, I present our experimental results. Finally, concluding remarks are provided in Section 5.

2. Related Works

Considerable attention has been given to understanding the importance of internal units in neural networks. Understanding neuron properties and evaluating their importance raise awareness to the need to adopt the quantitative assessment of neurons' properties. Such manners are utilized to measure the activation importance of each node and to appoint a score to them. To determine the importance of hidden neurons, Dhamdhere et al. [13] apply integrated gradients by calculating a summation of the gradients of the prediction with respect to the input. Morcos et al. [14] explored the relationship between the output of individual neurons and the classification performance of neural networks to evaluate mutual information and class selectivity for each neuron's activation. Moreover, Na et al. [15] have recently used the highest mean activation to measure the importance of individual units on language tasks, showing that different units are selectively responsive to specific morphemes, words, and phrases. Despite the fact that most of the aforementioned techniques emphasize effective methods to identify important neurons, most of their concentration was on gaining the best understanding of the network's mechanism, with limited attention towards tracking the pathway of a given class and analyzing the network's behaviour at a sub-network level.

In an attempt to study the topic of cumulative network pruning, several approaches have been developed [16]. Frankle et al. [17] found that networks contain a sub-network that reaches a test accuracy that is comparable with the original network through an iterative pruning technique. Their core idea was to find a smaller, well-suited architecture to the target task at the training phase. Ashual et al. [18] proposed a composite network that focuses on extracting the sub-network for each class during the training process. These methods show the possibility of revealing the sub-network in an indirect way, whether through eliminating unimportant parts of the neural networks or through training multiple branches of the network, where different groups of branches denote different objects. Therefore, providing a way to measure the importance of different parts of the network, to detect the important neurons, and to identify relationships among neurons are worthwhile to extract the sub-network for a specific object or class.

In this paper, I propose an efficient method to identify the important neurons that are related to object concepts and mainly consider the relationship between these neurons and their object concept or class. I first quantify the activation values among neurons, based on which histograms of each neuron are generated. Then, the obtained histograms are clustered to identify the neurons' importance. I then introduce a network-wide holistic approach that efficiently identifies important neurons and their influential connections to reveal the pathway of a given class. The influential connections as well as their important neurons are carefully evaluated to reveal the sub-network of each object's concepts.

3. Method

Measuring the importance of different network parts always requires a more meticulous process. Most of the current techniques focus on providing efficient techniques to determine important neurons, with limited attention being paid to tracking the pathway of a given class and analyzing the network's behaviour at the sub-network level. My importance-measurement method introduces a novel way to reveal the sub-network; it estimates the importance of neurons in each layer and identifies a subset of their influential connections whose activation values are the most effective in identifying relationships among neurons. This section presents my overall proposed framework, which consists of two parts. First, the evaluation of neuron importance is discussed; this determines the importance of neurons in each layer. Then, I present a network-wide holistic method that efficiently identifies important neurons and their influential connections to reveal the pathway of a given class. The entire algorithm is provided in Algorithm 1. The details are provided below.

Algorithm 1: Network Pathway Extraction.

```
Input: a pre-trained model, training set $(x, y)$;
Output: sub-network of each class;
for each class do
    compute the activation for each neuron Equation (1);
    generate histograms for each neuron;
    cluster the obtained histograms to identify the important neurons;
    identify the influential connections Equation (2);
    apply MV method to detect the most important connections for each neuron;
end
```

3.1. Analyzing the Importance of Individual Neuron

The aim was to reveal effective units in neural networks by estimating their activation. When the training data are fed through the network, a different representation is obtained for each example and has unique activation throughout all neurons in the network. A forward passing via a trained model is applied to derive the output of each unit. Different input examples can present more instances, and the output can be seen as random variables. I utilized a novel process to estimate the importance of units. In each layer, the weights are multiplied with an input sample, x, to deliver an output corresponding n activation. The activation at j-th unit is determined by summing the weights of the activations from all unit in the $(i-1)$-th layer. The production of the j-th unit in the i-th layer of the neural network is given by

$$t_j^{(i)}(x_n) = \sigma\left(b_j^{(i)} + \sum_p w_{p,j}^{(i-1)} t_p^{(i-1)}(x_n)\right), \tag{1}$$

where x_n is the n-th data sample at the input, σ denotes the activation function, b_j^i is the corresponding bias for the j-th unit in the i-th layer, and $w_{p,j}^{(i-1)}$ denotes the weight that connects p-th unit from the prior layer $(i-1)$ with the j-th unit in the i-th layer (current layer).

After obtaining a matrix of edge values for each example by Equation (1), histograms for each edge were generated (see Figure 1). Then, the obtained histograms were clustered into three different clusters (High, Medium, and Low). Therefore, I came up with a binary vector for every layer that demonstrates whether such units are critical, where 1 indicates that the neuron is essential and 0 otherwise.

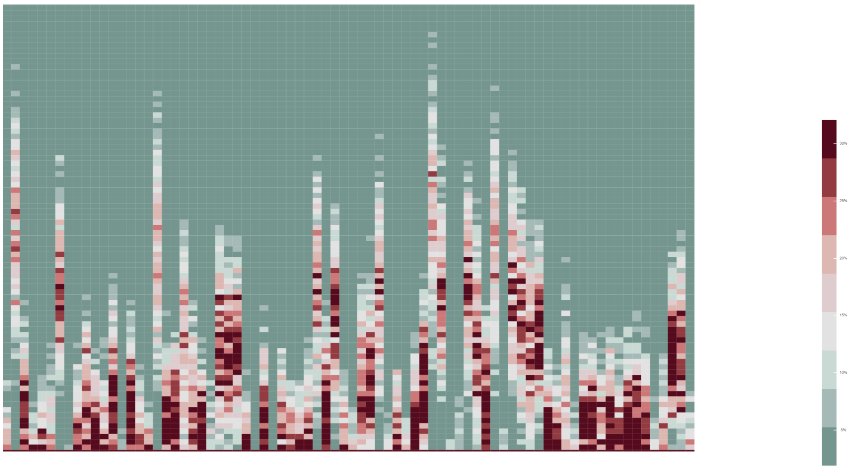


Figure 1. Activation distribution from different units of trained FC network.

3.2. From Individual Neuron to Sub-Network Analysis

The collection of important neurons at each layer in a network will only give localized feature descriptors, of which essential neurons for a particular class can be detected. By focusing on this assumption, I am able to justify how distinct neurons become the most representative for a given class, maintaining the class information within the input domain, consequently providing a way to detect how the activations from the nodes of the previous layer impact the activations of the current layer is required to extract a sub-network for a given class. A combination of neurons and their influential connections provides a novel way to reveal the sub-network.

As the activation of a neuron in the current layer is computed as the weighted sum of activations from neurons in the previous layer obtained by Equation (1), in fully connected layers, the influential connections $t_j^{(i-1)}(\hat{x}_n)$ of a single neuron are obtained by multiplying the weights of that neuron with its corresponding activations in the previous layer, as follows:

$$t_j^{(i-1)}(\hat{x}_n) = w_{p,j}^{(i-1)} t_p^{(i-1)}(x_n), \tag{2}$$

where x_n denotes the n-th data example at the input, $w_{p,j}^{(i-1)}$ is the weight that links p-th unit from the former layer $(i-1)$ with the j-th unit in the i-th layer (current layer), and $t_p^{(i-1)}(x_n)$ represents the corresponding activations in the previous layer $(i-1)$. After the values of the influential connections are obtained for each neuron, the majority voting (MV) method [19,20] is applied to detect the most important connections for each neuron.

4. Experiment and Discussion

I empirically investigated the performance of my proposed approach using two different datasets: MNIST [21] and MNIST-Fashion [22] through several models. The proposed method was implemented using Keras and TensorFlow [23] in Python. The specifications of the datasets and their architecture for the used models are presented in Table 1.

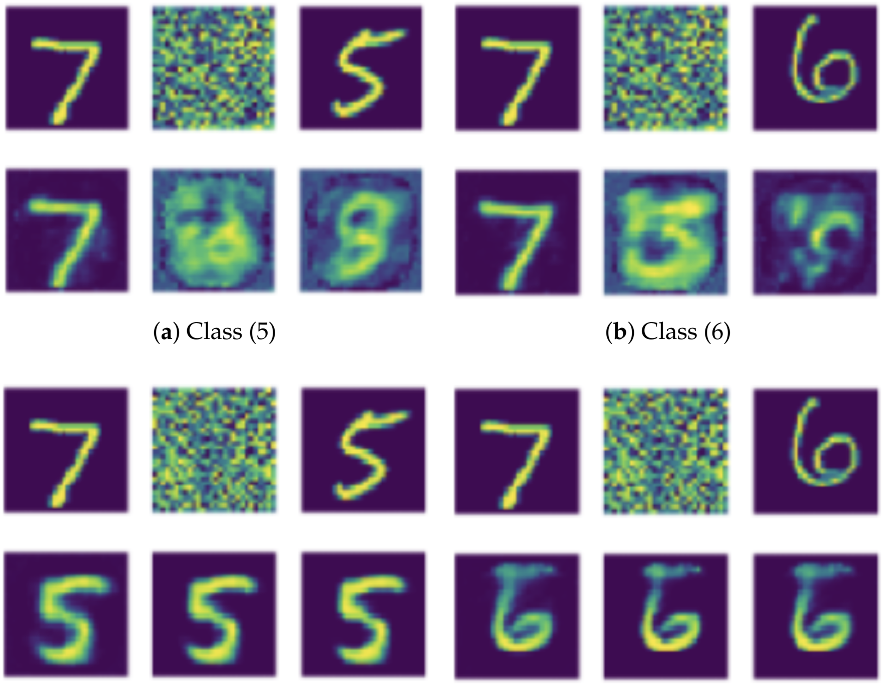
Table 1. Details of datasets and their architectures used in my experiments.

Dataset	Examples	Image Size	AutoEncoder Architecture	FC Architecture
MNIST [21]	70,000	28 × 28 × 1	784-1000-1000-1000-784	784-1000-1000-1000-10
MNIST-Fashion [22]	70,000	28 × 28 × 1	784-1000-1000-1000-784	784-1000-1000-1000-10

The first model was an auto-encoder model, which was optimized in an unsupervised manner. There are no fine-tuning and pre-training processes involved. The stochastic gradient descent was used, and each batch included 100 random shuffled examples. For both datasets, an initial learning rate of 0.006 with a momentum of 0.9 and weight decay of 0.0005 were utilized.

I carried out a visual assessment to evaluate the results of the proposed method. Some instances of actual inputs and reconstruction images generated by my model are presented in Figures 2 and 3. Two evaluation studies were adopted: an ablation study and an insertion study.

After the pathway was identified, I applied the ablation study by forcing the pathway of a particular class to be zero and performed the propagation of the forwarding pass. Three samples were fed to the network after identifying the class pathway: (different image (left), random noise image (middle), and same image (right)). The reconstruction images of such this assessment were visualized using the three samples (see Figures 2a,b and 3a,b). This ablation study is best suited for evaluating the effectiveness of extracting a particular class pathway. One clear example can be seen in Figure 2a; the model failed to reconstruct the image of digit 5 back to its original shape because the pathway of that digit was ablated. It is also worth noting that the reconstruction of the image of digit 7 obtains an optimal approximation of the underlying input data because the pathway of digit 7 was still available. These observations allow us to efficiently identify important neurons and their influential connections to reveal the pathway of a given class.



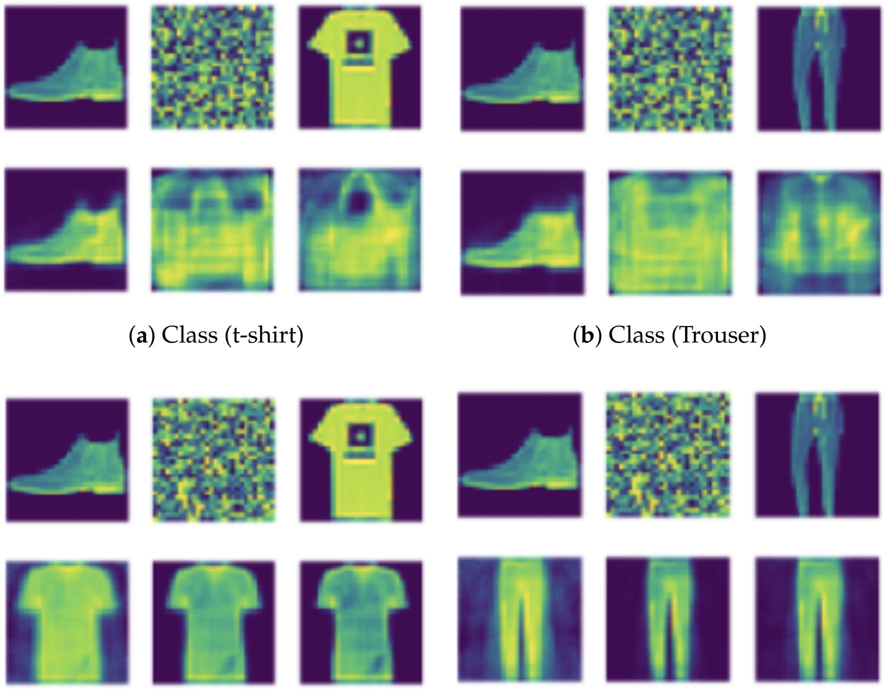
(**a**) Class (5)

(**b**) Class (6)

(**c**) Class (5)

(**d**) Class (6)

Figure 2. An ablation study (**Top**) and an insertion study (**Bottom**) with different examples and different identification of class pathway on the MNIST dataset.



(**a**) Class (t-shirt)

(**b**) Class (Trouser)

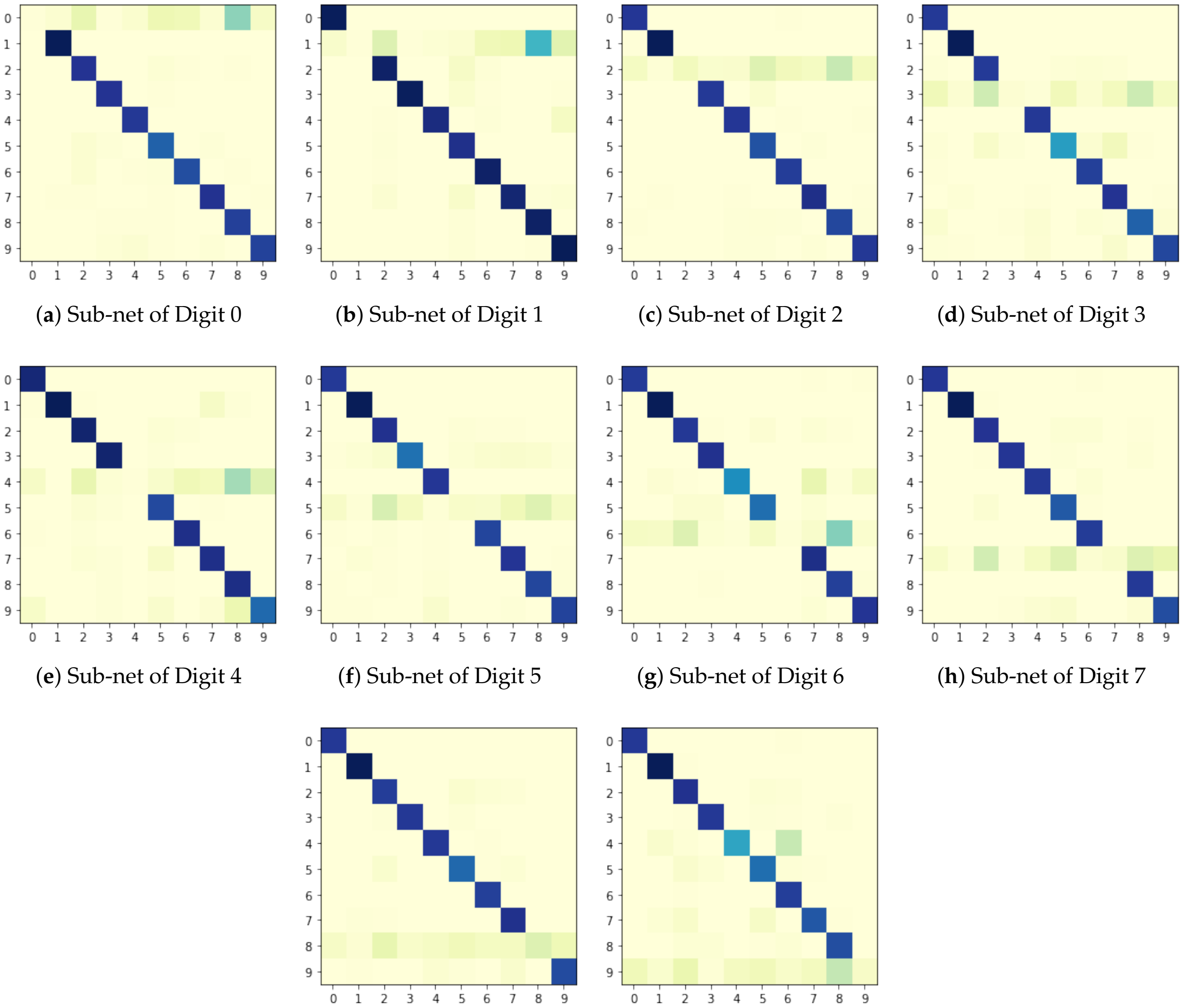
(**c**) Class (t-shirt)

(**d**) Class (Trouser)

Figure 3. An ablation study (**Top**) and an insertion study (**Bottom**) with different examples and different identification of class pathway on the fashion MNIST dataset.

Figures 2 and 3 also show the results of the insertion study, where I forced the pathway of a particular class to be one and zero otherwise and performed the propagation of the forwarding pass (see Figures 2c,d and 3c,d). This study obviously showed that all input images ended up with the reconstruction of the same class of the identified pathway. The insertion study helped to evaluate the effectiveness of my method when revealing the pathway of a given class.

The second model was a classification model, which trained end-to-end in a supervised manner. Figure 4 experimentally analyzes the performance of my proposed method. Using my method, I identified the pathway of each class of trained fully-connected networks. My experiment showed that ablating a specific class pathway has no effect on other classes. One obvious explanation is that the proposed method succeeded in carefully identifying the pathway of each class. It is also crucial to note that because digit 6 has a comparable structure to that of digit 8, especially with regards to the bottom part of both digits (see Figure 4g), the model classified most examples of digit 6 into the class of digit 8. One reasonable justification is that the presented method was able to determine the homogeneous patterns for a particular digit, which leads to the identification of the pathway of the target class.



(**a**) Sub-net of Digit 0 (**b**) Sub-net of Digit 1 (**c**) Sub-net of Digit 2 (**d**) Sub-net of Digit 3

(**e**) Sub-net of Digit 4 (**f**) Sub-net of Digit 5 (**g**) Sub-net of Digit 6 (**h**) Sub-net of Digit 7

(**i**) Sub-net of Digit 8 (**j**) Sub-net of Digit 9

Figure 4. Results of different class pathways when applying an ablation study on the MNIST dataset.

5. Conclusions

In this paper, I proposed an efficient method to identify the important neurons, mainly considering the relationship between these neurons and their object concept or class. I introduce a network-wide holistic approach that efficiently identifies important neurons and their influential connections to reveal the pathway of a given class. The influential connections as well as their important neurons were carefully evaluated to reveal the sub-network of each object's concepts. I showed the effectiveness of the proposed method using two different datasets. Our potential future work is to expand the proposed framework to filters in CNNs and to investigate it with more difficult datasets. Although this procedure significantly identifies influential connections, the more theoretical analysis also requires further study to understand how such ideas can be generalized further. Moreover, I believe more investigation is needed to carefully study or gather evidence to determine whether the object's patterns can be a local receptive field in the FC networks, as connecting each neuron to only a local region of the input space might help to justify CNNs and to prove that the actual connections are local receptive fields. There are also several challenges and extensions we perceive as useful research directions. Extending the proposed framework and combining it to strengthen the discriminative features and improve the encoder's ability of deep clustering [24] is an important direction for future work.

Funding: This work was supported by the Deanship of Scientific Research, King Khalid University of Kingdom of Saudi Arabia under research grant number (RGP1/357/43).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
2. Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6541–6549.
3. Bau, D.; Zhu, J.Y.; Strobelt, H.; Zhou, B.; Tenenbaum, J.B.; Freeman, W.T.; Torralba, A. Gan dissection: Visualizing and understanding generative adversarial networks. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
4. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
5. Dosovitskiy, A.; Brox, T. Inverting visual representations with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4829–4837.
6. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.
7. Beguš, G.; Zhou, A. Interpreting intermediate convolutional layers of generative CNNs trained on waveforms. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 3214–3229. [CrossRef]
8. Suganyadevi, S.; Seethalakshmi, V.; Balasamy, K. A review on deep learning in medical image analysis. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 19–38. [CrossRef] [PubMed]
9. Erhan, D.; Bengio, Y.; Courville, A.; Vincent, P. *Visualizing Higher-Layer Features of a Deep Network*; Technical Report; University of Montreal: Montreal, QC, Canada, 2009; Volume 1341.
10. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
11. Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; Lipson, H. Understanding neural networks through deep visualization. *arXiv* **2015**, arXiv:1506.06579.
12. Novakovsky, G.; Dexter, N.; Libbrecht, M.W.; Wasserman, W.W.; Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* **2022**, 1–13. [CrossRef]
13. Dhamdhere, K.; Sundararajan, M.; Yan, Q. How important is a neuron? In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.

14. Morcos, A.S.; Barrett, D.G.; Rabinowitz, N.C.; Botvinick, M. On the importance of single directions for generalization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
15. Na, S.; Choe, Y.J.; Lee, D.H.; Kim, G. Discovery of Natural Language Concepts in Individual Units of CNNs. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
16. Alqahtani, A.; Xie, X.; Jones, M.W. Literature Review of Deep Network Compression. *Informatics* **2021**, *8*, 77. [CrossRef]
17. Frankle, J.; Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
18. Ashual, O.; Wolf, L. Specifying object attributes and relations in interactive scene generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4561–4569.
19. Alqahtani, A.; Xie, X.; Essa, E.; Jones, M.W. Neuron-based Network Pruning Based on Majority Voting. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021; pp. 3090–3097.
20. Alqahtani, A.; Xie, X.; Jones, M.W.; Essa, E. Pruning CNN filters via quantifying the importance of deep visual representations. *Comput. Vis. Image Underst.* **2021**, *208*, 103220. [CrossRef]
21. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
22. Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv* **2017**, arXiv:1708.07747.
23. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. In Proceedings of the Symposium on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
24. Alqahtani, A.; Xie, X.; Deng, J.; Jones, M.W. Learning discriminatory deep clustering models. In Proceedings of the International Conference on Computer Analysis of Images and Patterns, Salerno, Italy, 3–5 September 2019; pp. 224–233.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.