Layer-Wise Entropy Analysis and Visualization of Neurons Activation

Longwei Wang[1(✉)], Peijie Chen[1], Chengfei Wang[1], and Rui Wang[2]

[1] Department of Computer Science and Software Engineering, Auburn University, Auburn, USA
{lzw0070,pzc0018,czw0078}@auburn.edu
[2] Department of Information and Communications, Tongji University, Shanghai, China
ruiwang@tongji.edu.cn

Abstract. Understanding the inner working mechanism of deep neural networks (DNNs) is essential and important for researchers to design and improve the performance of DNNs. In this work, the entropy analysis is leveraged to study the neurons activation behavior of the fully connected layers of DNNs. The entropy of the activation patterns of each layer can provide an efficient performance metric for the evaluation of the network model accuracy. The study is conducted based on a well trained network model. The activation patterns of shallow and deep layers of the fully connected layers are analyzed by inputting the images of a single class. It is found that for the well trained deep neural networks model, the entropy of the neuron activation pattern is monotonically reduced with the depth of the layers. That is, the neuron activation patterns become more and more stable with the depth of the fully connected layers. The entropy pattern of the fully connected layers can also provide guidelines as to how many fully connected layers are needed to guarantee the accuracy of the model. The study in this work provides a new perspective on the analysis of DNN, which shows some interesting results.

Keywords: Entropy analysis · Visualization · Neurons activation

1 Introduction and Motivation

For the past decade, deep learning has been proposed as an efficient way to realize the general artificial intelligence [2,3]. There have been significant progresses on the design of neural network architectures [3,4]. Deep learning algorithms have made great improvement in all kinds of applications.

Although deep learning has achieved significant success in a wide range of applications, there are few works that can fully illustrate the internal working mechanisms of the deep neural networks (DNN). They are often treated as black box and the optimization process is ignored in the applications [5].

H. Gao et al. (Eds.): ChinaCom 2019, LNICST 313, pp. 29–36, 2020.
https://doi.org/10.1007/978-3-030-41117-6_3

Understanding the inner working mechanism of deep neural networks (DNNs) is essential and important for researchers to design and improve the performance of DNNs. One effective way to explain how neurons work internally is to study what kind of features can activate certain neurons, which is known as the feature visualization in the deep learning community [1]. One such method is called activation maximization, which synthesizes an image that highly activates a neuron.

The idea of using information theoretic methods for investigating deep neural networks was proposed by Tishby (2015) [6]. However, they did not conduct any experimental result. In the work, they propose that the neural network layers can be seen as a successive Markov chain. The mutual information of the input layer $\mathbf{X}$ with the inner layers $\mathbf{Y}$ are studied in the information plane. The theoretical base for this study is the invariance of mutual information to re-parameterization along the Markov chain of the layers. They also show that the optimal neural networks can approach the Information Bottleneck bound of the optimal achievable representations of the input $\mathbf{X}$ [8,9].

The mutual information study of the layers does not fully characterize the working mechanisms of the deep neural networks. In this work, we adopt the entropy analysis to study the behavior of the fully connected layers. The entropy of the activation patterns of each layer can provide a performance metric for the evaluation of the network model accuracy.

1.1 Contribution

In this work, the neuron activation pattern is studied by inputting the images of an individual class and the statistical activation pattern differences between shallow and deep layers' neurons is investigated.

Entropy analysis is used to quantify the statistical property of neuron activation patterns. The study is conducted based on a well trained network model. The activation patterns of shallow and deep layers of the fully connected layers are analyzed by inputting the images of a single class, which can provide some useful insights as to the design and optimization of deep convolutional neural networks. By analyzing the activation patterns for different layers, it can not only help us understand the behavior of CNN, but also give us a way to improve the CNN. The entropy pattern of the fully connected layers can provide some guidelines as to how many fully connected layers are needed to guarantee the accuracy of the model. The method provides a new perspective on the analysis of deep CNN, which shows some interesting results.

2 Visualization Methods

2.1 How to Visualize the Neurons Activation of a Layer?

The visualization method for the neuron activation pattern is depicted in Fig. 1. The experiment is conducted based on a well trained neural network model.

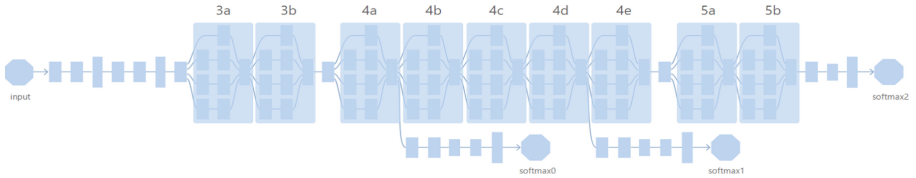

Fig. 1. Neuron activation pattern extraction

The images of an individual class are inputted to the network, and we first extract the representation of the layer and let it go through the softmax function, then the activation probabilities of the layer are obtained. In this way we can visualize the neuron activation pattern.

2.2 How to Quantify the Activation Patterns of Each Layer

The data used in this work are MNIST and CIFAR. We study the internal neuron patterns for different classes by visualizing the neuron activations in the fully connected layers (Figs. 2, 3 and 4).

We take advantage of the entropy tool in information theory to quantify the randomness of the neuron activation of different layers. For a fixed class, we first use those test images as input and calculate the output of each neurons in each fully connected layer. And then we average the output over all test images of all neurons in every fully connected layer. By using softmax function, we can derive the activation pattern of the neurons (probability of the neuron will be shown). Finally, we use the formal entropy definition to compute the entropy of each layer.

The entropy in information theory is used to characterize the uncertainty of the random phenomenons. The definition of the information entropy is quite general, and is expressed in terms of a discrete set of probabilities p_i so that

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{1}$$

where the probabilities p_i are the activation probabilities after softmax functions. The entropy can be used as a measure of the activation pattern of the neurons in the network model.

For the fair comparison of the entropy among layers with different number of neurons, normalization of the layer-wise entropy is performed for each layer.

$$H(X) = -\frac{1}{f(n)} \sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{2}$$

where the term $f(n)$ is the normalization factor, which is function of the number of neurons in each layer.

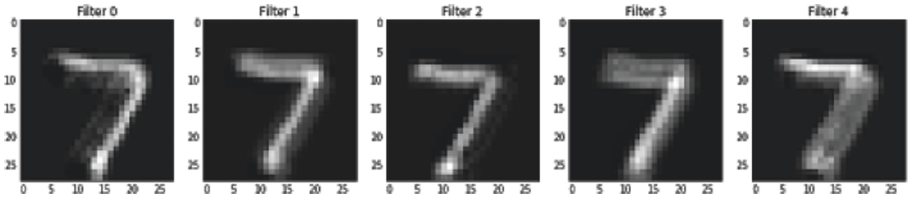

Fig. 2. Visualization of hidden layer 1

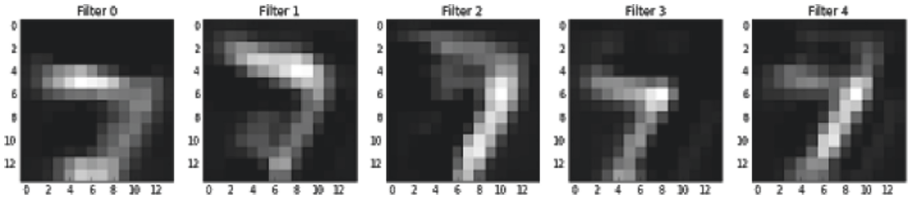

Fig. 3. Visualization of hidden layer 2

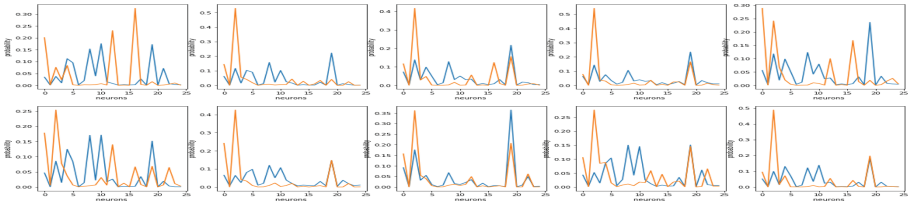

Fig. 4. Neuron activation of fully connected layer 1

3 Results 1: MNIST

3.1 Visualization of the Convolution Layers' Neuron Activation

The figures of hidden layer neurons show that as the convolutional layer getting deeper, only abstract features remain in the images. Such kind of feature is hard to understand by human, so we move forward to the following layers, which are fully connected layers, to analyze how these kinds of features activate the neurons of the CNN.

3.2 Visualization of the Fully Connected Layers' Neuron Activation

(1) Direct visualization: By looking at the output of different classes, we can see that the activation of the shallower layer is more unstable compare to the deeper layer. In Figs. 5, 6 and 7, blue and orange represent two different classes. The x-axis is neurons, y-axis is the probability that the neuron will activate. In Fig. 5, it seems that in layer 1, multiple neurons are activated with high probability.

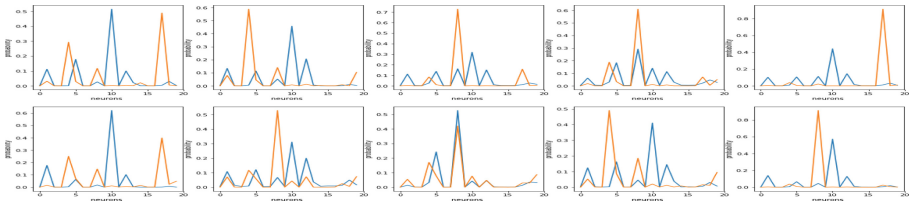

Fig. 5. Neuron activation of fully connected layer 2

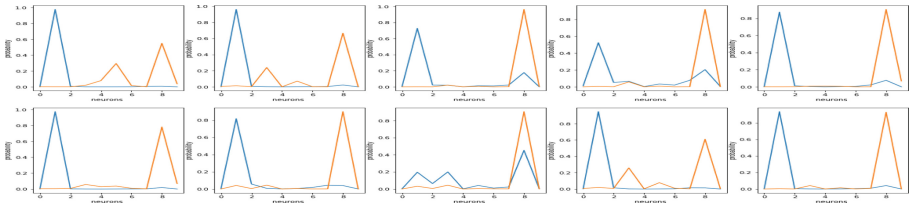

Fig. 6. Neuron activation of fully connected layer 3

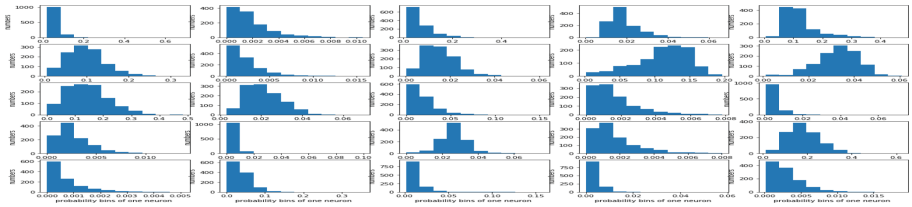

Fig. 7. Statistical neuron activation of fully connected layer 1

However, in layer 2 (Fig. 6), only one or two neurons are activated with high probability. In the last layer, the activation seems very stable expect when the error occurs.

(2) Statistical visualization: In this section, we study the statistical activation of the neurons in the neural network (Figs. 8, 9 and 10, are the histograms of the activation probabilities of the neurons of 1000 samples for a fixed class. The x-axis is the probability of activation and the y-axis is the number of samples). We found that in shallower layer, combinations of neurons will be activated for

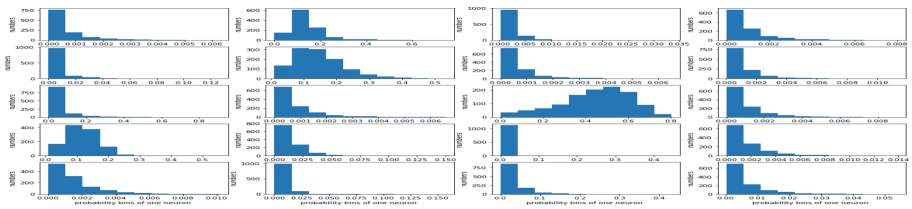

Fig. 8. Statistical neuron activation of fully connected layer 2

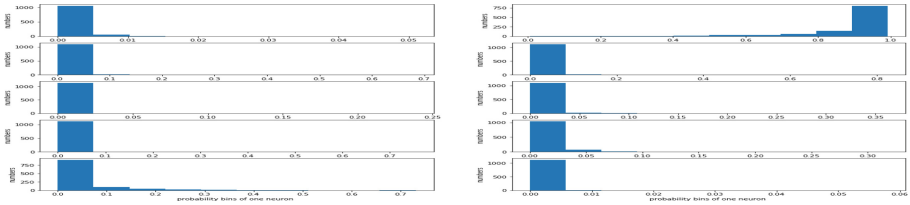

Fig. 9. Statistical neuron activation of fully connected layer 3

a fixed class. But as the layer goes deeper, only fewer neurons will be activated. And the activation combinations become more and more stable.

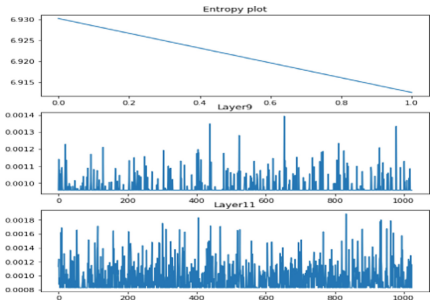

Fig. 10. 2 fully connected layers

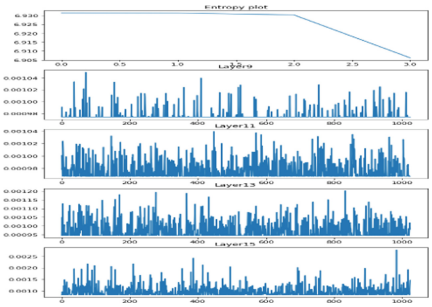

Fig. 11. 4 fully connected layers

4 Results 2: CIFAR

4.1 Entropy Reduction

The CIFAR data is studied in this section. By looking at the representations of different fully connected layers, we can see that the activation of the shallower layer is more unstable compared with the deeper layer. In Figs. 11, 12 and 13, the entropy plot (first plot of each figure) shows that the entropy of the neuron activations pattern is monotonically reduced with depth of the fully connected layers. And the activation become more and more stable as the layer goes deeper. Another interesting phenomenon is that if the entropy plot is pretty "flat", that is, the gradient of the entropy is very small, then these fully connected layers don't make significant contributions to the network. (The accuracy of these three models from 2 fully connected layers to 6 fully connected layers are 0.774, 0.7686 and 0.7187 respectively.)

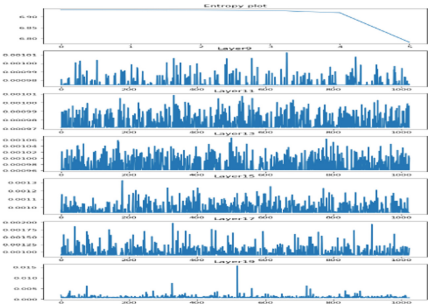

Fig. 12. 6 fully connected layers

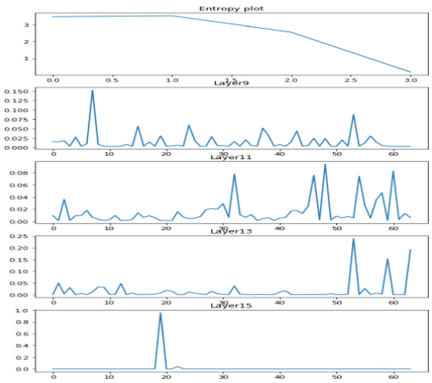

Fig. 13. 4 fully connected layers, accuracy = 0.3935

4.2 Relationship Between Entropy and How Many Fully Connected Layers Are Needed

As we can see in the experiment results, the entropy in Fig. 14 increases a little bit and then decrease, which is not the expected "entropy reduction" phenomenon. However, this abnormality somewhat means that there's shortcoming in our model (The accuracy is roughly 0.39). By simply deleting the corresponding layer (the second fully connected layer), we can get a much better result as Fig. 15 shows. But There is still an abnormality in the entropy. Last by deleting the corresponding layer (fully connected layer 3), we actually get a better result as Fig. 15 shows.

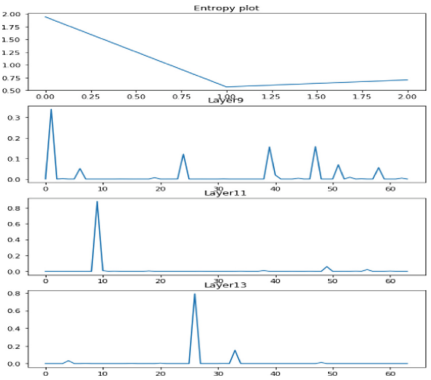

Fig. 14. 3 fully connected layers, accuracy = 0.6283

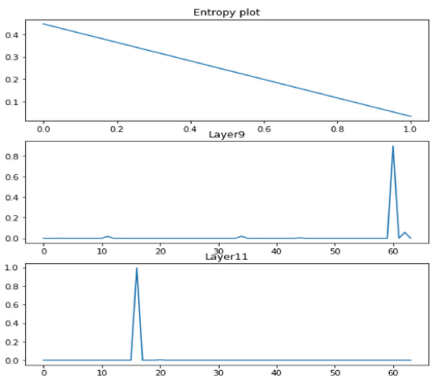

Fig. 15. 2 fully connected layers, accuracy = 0.6626

5 Conclusion

In this work, we found that for the well trained deep neural networks model, the entropy of the neuron activation pattern is monotonically reduced with the depth of the layers. That is, the neuron activation patterns become more and more stable with the depth of the fully connected layers. Furthermore, if the entropy of the first few fully connected layers are almost the same, such a layer do not have a significant contribution to the overall neural network classification accuracy. So we tried to remove some of the fully connected layers, and the prediction accuracy is almost the same.

Our experiments also indicate that when the neural networks is well trained, the entropy of the fully connected layers are monotonically reduced, while for the not-so-well-trained network model, the entropy of the fully connected layer neurons activation is sort of random.

References

1. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill **2**(11), e7 (2017)
2. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
3. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR, abs/1512.03385 (2015)
5. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes (2016)
6. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing (1999)
7. Wang, L., Liang, Q.: Representation learning and nature encoded fusion for heterogeneous sensor networks. IEEE Access **7**, 39227–39235 (2019)
8. Moshkovich, M., Tishby, N.: Mixing complexity and its applications to neural networks (2017). URL https://arxiv.org/abs/1703.00729
9. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. In: Information Theory Workshop (ITW), pp. 1–5. IEEE (2015)
10. Mohamed, S., Rezende, D.V.: Variational information maximisation for intrinsically motivated reinforcement learning. In: NIPS, pp. 2125–2133 (2015)
11. Wang, L., Liang, Q.: Partial interference alignment for heterogeneous cellular networks. IEEE Access **6**, 22592–22601 (2018)
12. Achille, A., Soatto, S.: Information dropout: learning optimal representations through noisy computation (2016). URL http://arxiv.org/abs/1611.01353
13. Bau, D., Zhou, B., Khosla, A., Oliva, A. and Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6541–6549 (2017)
14. Agostinelli, F., Hoffman, M., Sadowski, P., Baldi, P.: Learning activation functions to improve deep neural networks. arXiv preprint arXiv:1412.6830 (2014)