Reproduction : What do CNNs Learn in the First Layer and Why? A Linear Systems Perspective

Shivam Chadha
BITS Pilani Goa Campus
f20190704@goa.bits-pilani.ac.in

Rajan Sahu
BITS Pilani Pilani Campus
f20190572@pilani.bits-pilani.ac.in

1 Introduction

In this paper, we analyze the success of Convolutional Neural Networks (CNNs) in image recognition tasks, attributing it to their ability to learn effective image representations. It notes the consistency in learned representations across different CNNs and tasks, highlighting the importance of studying the first layer's representation due to its relationship with the input patches. The concept of "redundancy reduction" from biological neural networks is introduced, suggesting that early layers aim to disentangle input by whitening it. The study derives an analytical form for the energy profile of linear CNNs, demonstrating its convergence to a whitening transformation with infinite iterations.

2 Methodology

2.1 Quantifying Consistency using Energy:

A method based on the linear systems perspective is introduced to measure the similarity between representations learned by different CNNs using energy profiles characterized by the principal components of input image patches. Results on the method show remarkable consistency across architectures and datasets, with sensitivity peaking at intermediate spatial frequencies. We define two energy profiles, which are computed using similar concepts. We define, under the assumption that the mean of the patches in the training set is zero, the eigenvectors u_i correspond to the matrix $\sum_n p_n p_n^T$ where p_n are the set of patches. Then for a given set of filters w_k and a set of PCA vectors u_i, the energy profile of the set is given by a vector e whose ith component is given as the following :

$$\mathbf{e_i^2} = \frac{1}{K}\sum_{k=1}^{K}(w_k^T u_i)^2 \tag{1}$$

Similarly, the energy profile characterized by the PCA vectors u_i and set of patches p_n for a linear CNN with one hidden layer is given by a vector λ whose ith component is given as the following :

$$\lambda_\mathbf{i}^\mathbf{2} = \frac{1}{N}\sum_{n=1}^{N}(p_n^T u_i)^2 \tag{2}$$

The method shows that a system based on convolutions is fully specified by its frequency response. And since the filters in CNNs are highly localized in space, the frequency response is characterized by principal components of the input image patches, which are also highly localized in frequency.

2.2 Is Consistency due to CNNs Learning Semantically Meaningful Features?

A natural explanation for the remarkable consistency of the learned representation in the first layer is that CNNs learn a representation that is good for object recognition. Two experiments were conducted: the first involved freezing the initial layer of the CNN, where it was hypothesized that

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

performance would deteriorate; however, no such decline was observed. In the second experiment, the network was trained with random labels. Contrary to expectations, we find the same energy profile when CNNs are trained with true labels and random labels. This suggests that the consistency is due to the input and/or training algorithm; we analyze this behaviour by formulating the relation between input patches and filters for a simple CNN.

2.3 Theory in Simple Linear CNNs

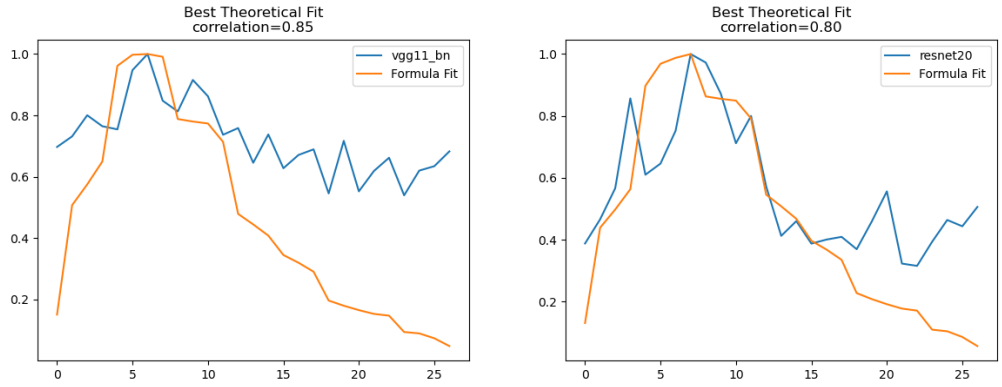
The main theorem provides an analytic formula for the energy profile. In a depth-2 linear CNN with zero-mean filters and variance $\sigma^2 I$, trained using gradient descent on MSE loss. The filter energy profile at any iteration t is given by

$$e_i \approx \hat{c} \cdot \frac{\left[1-(1-\eta\lambda_i^2)^t\right]}{\eta^2\lambda_i^2}\lambda_i + \xi_i$$

Where λ_i is the energy profile of the training patches and ξ_i a random vector that depends on the initialization and its magnitude goes to zero as $\sigma \to 0$. As the iterations(t) go to infinity, the filters of the CNN perform a whitening effect.

3 Experiment

We compute the energy profiles of CIFAR10 with VGG11 and ResNet20. The energy profile is consistent (correlation>=0.8), and increasing the iterations increases the energy, approximating the whitening process. It is difficult to match the energy profiles exactly for some t, such that the actual energy profile and the formula exactly matches but the overall profiles are correlated.



(a) Energy Profile fit for VGG11

(b) Energy Profile fit for ResNet20

Figure 1: Energy Profile of the actual model compared to predicted profile by formula. The correlation between these two is calculated.

4 Conclusion

The paper shows the consistency observed in the energy profiles learned by different convolutional neural networks (CNNs) in their first layers. Through theoretical analysis and empirical validation, it demonstrates that this consistency arises from the inherent properties of the input data and the training algorithm rather than being specifically tailored for object recognition. The findings suggest that CNNs implicitly engage in a transformation closely approximating whitening.

References

[1] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pp. 818-833. Springer International Publishing, 2014.

[2] Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. "Similarity of neural network representations revisited." In International conference on machine learning, pp. 3519-3529. PMLR, 2019.

[3] Soudry, Daniel, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The implicit bias of gradient descent on separable data." Journal of Machine Learning Research 19, no. 70 (2018): 1-57.