Characterizing & Exploring Deep CNN Representations Using Factorization*

Uday Singh Saini
University of California Riverside
usain001@cs.ucr.edu

Evangelos E. Papalexakis
University of California Riverside
epapalex@cs.ucr.edu

Abstract

Deep neural networks have gained enormous popularity in machine learning and data science alike, and rightfully so, since they have demonstrated impeccable performance in a variety of supervised learning tasks, especially a number of computer vision problems. Albeit very successful in providing accurate classifications, deep neural networks are notorious for being hard to interpret, explain, and debug, a problem amplified by their increasing complexity. This is an extremely challenging problem and the jury is still out on whether it can be solved in its entirety. Here, we propose a novel factorization framework, aiming to answer the following questions: Given an already trained deep neural network, and a set of test inputs, how can we gain insight into how those inputs interact with different layers of the neural network? Furthermore, can we characterize a given deep neural network based on its observed behavior on different inputs? The proposed approach will give a more flexible yet still interpetable mechanism for understanding and interacting with deep networks.

1 Overview of Proposed Method & Key Results

The key idea behind our proposal, shown in Figure 1, is the following: we jointly factorize the raw inputs to the deep neural network and the outputs of each layer, to the same low-dimensional space. Intuitively, such a factorization will seek to identify commonalities in different parts of the raw input and how those are reflected and processed within the network. For instance, if we are dealing with a Deep CNN that is classifying handwritten digits, such a joint latent factor will seek to identify different shapes or patterns that are common in a variety of input classes and identify correlations on how different layers behave collectively for such high-level latent patterns.

Here, we present a proof-of-concept approach. Suppose we have an already trained deep CNN and we have a separate hold-out validation set. If we feed this validation set to the network, we express a coupled factorization of the raw validation inputs and the intermediate outputs of the activation layers as shown in Figure 1.

$$J(\underline{\mathbf{P}}, \mathbf{F}, \underline{\mathbf{O}}) = \sum_{i=0}^{C-1} \|\mathbf{D}_i - \mathbf{P}_i \mathbf{F}^T\|_F^2 + \sum_{j=0}^{L-1} \|\mathbf{A}_j - \mathbf{O}_j \mathbf{F}^T\|_F^2, \tag{1}$$

where C is the number of channels in an input image and L is the number of layers of the neural network being analyzed, $\underline{\mathbf{P}}$ and $\underline{\mathbf{O}}$ are sets of factor matrices . Each $\mathbf{D}_i$ is the set of i^{th} channel of the input images to the neural network, where each column of $\mathbf{D}_i$ is a channel of the image in vectorized form, thus each row of D_i is a pixel or location in the original image. Each $\mathbf{A}_j$ is the matrix of activations of the j-th layer of the neural network, where each column of $\mathbf{A}_j$, for instance,

*NeurIPS 2018 Workshops on Integration of Deep Learning Theories

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

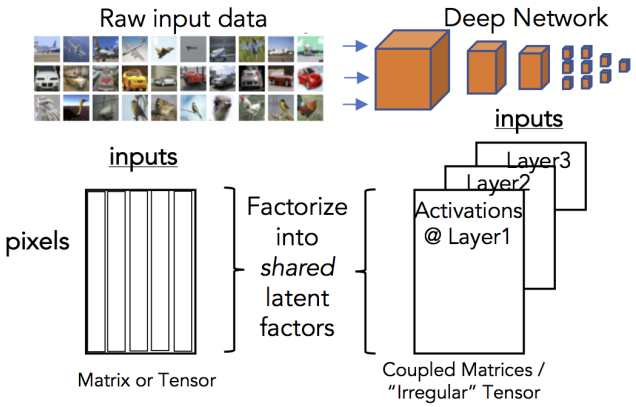


Figure 1: Overview of our approach.

the k-th column $\mathbf{A}_j(:,k)$, is the activation of layer j of the network for k-th input image, i-th channel of which is represented by $\mathbf{D}_i(:,k)$. Each $\mathbf{P}_i$ is a matrix that stores the latent representation of each pixel (for the i-th channel) in its rows, each $\mathbf{O}_j$ is a matrix that stores the latent representation of each neuron (or activation) of layer j in its rows. Finally, $\mathbf{F}$ is a matrix that stores the latent representation of each image fed to the neural network in it's rows. In our on-going work, we are experimenting with formulating the above coupled factorization as an instance of the PARAFAC2 tensor factorization [2], which is tailored to irregular tensors, where one of the modes is not of consistent size, and preliminary results agree with the ones obtained with model 1.

Key Insight: We expect the rank of the factorization to correlate with the complexity of the relationship learned by the network. Trivially, a linear relationship can be adequately described by a rank-one factorization; the more complex it gets, the more latent factors are required.

1.1 Deep Network Characterization

We used the MNIST dataset modified to a resolution of 28 by 28 pixels, and we analyze a simple network consisting of 2 consecutive Convolutional Layers with Maxpool and ReLU, followed by a fully connected layer which feeds to a softmax output. As a thought experiment, since one of our interests is to debug a poorly trained network, we did exactly that: we conduct the following experiment: we train a CNN poorly in two different ways: 1) train on a subset of the training data, and 2) train on a subset of the classes. Figures 2(a) and 2(b) show the RMSE of our objective function for different ranks. We make the following intriguing observation: *Better trained networks tend to have a coupled factorization of higher rank than poorly trained ones. Furthermore, this is derived* **without** *using test labels, but merely using raw inputs and activation outputs.* In our on-going work we have observed consistent results in different models (AlexNet, VGG, ResNet, and DenseNet) and for the CIFAR-10 dataset.

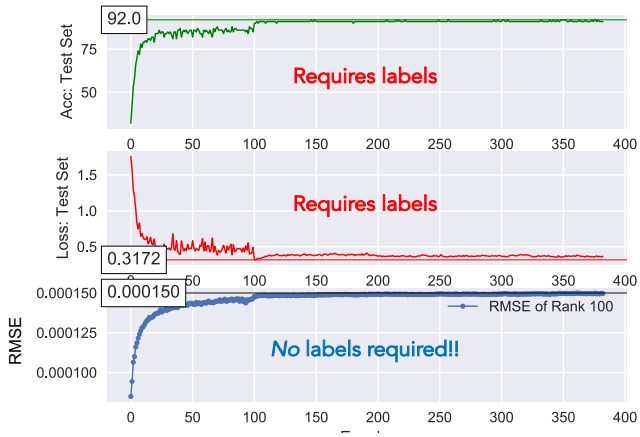
Figure 2(c) demonstrates such an experiment where the RMSE of our unsupervised model captures the same pattern as the loss function, which is evaluated on *labeled data*, and this can be used inform early stopping or changing the learning rate during training, and predict when the network starts to overfit.

1.2 Deep Network Visualization

The proposed coupled factorization is relating raw inputs and their transformations throughout the network. Such factorization models have been shown to produce *interpretable representations* [4], thus, we aim to investigate the *explanatory power* of our proposed model with respect to the complex relationships that the deep network has learned. In Figure 3, we visualize the latent factors learnt by our coupled factorization (both for the inputs as well as the hidden layers), on a problematic scenario where the network has been only trained on the digit '9'; in this case, our visualization clearly shows that the network is heavily underutilized and it mostly responds to high-level concepts that resemble '9' or different parts thereof.

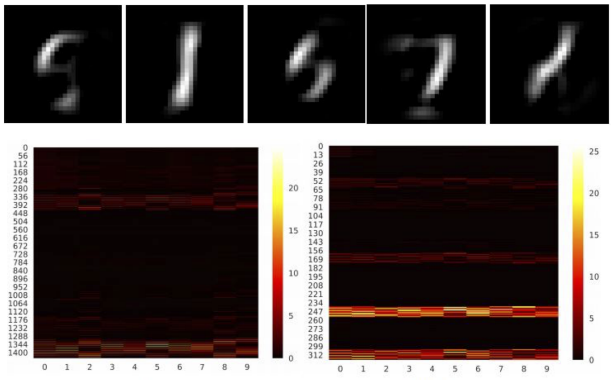
(a) Trained on subset of the classes

(b) Trained on subset of the data



(c) Factorization RMSE per epoch

Figure 2: Deep Network Characterization: (a,b): The worse the performance, the lower the rank of the factorization; (c): RMSE of our model captures the same trend as the test loss and can be used to inform early stopping or learning rate changes. Both patterns are obtained *without using labels*



(a) Layer 1 (b) Layer 2

Figure 3: Deep Network Visualization

2 Related Work

"Network Dissection" quantifies the interpretability of activations of hidden layers of CNNs by evaluating the alignment between neural activations in the hidden units and a set of semantic concepts. Olah, et al. [5] focus on learning what each neuron or a group of neurons detect based on feature vi-

sualization. Subsequently, Raghu et al. [6] introduced a Canonical Correlation Analysis based study that jointly analyzes the hidden layers of a CNN, however, this analysis is not relating the derived representations of CCA to the input data, thus may not be able to provide an end-to-end characterization and visualization. Sedghi et al. [7] analyze the singular values of the convolutional layers of a CNN towards better regularization and quality improvement. Sundarajan et al. [8] is proposing a framework for attributing the influence of certain features to a classification outcomes by providing a set of intuitive axioms that the framework should obey, while, Kim et al. [3] probe the network with a set of desired inputs and a user-defined concept, and measures the sensitivity of the network to this concept. Our work seeks to automatically determine such concepts, however, both works can be seen as complementary. Finally, Cohen et al. [1] theoretically analyze the coefficients of shallow and deep networks and draw parallels between different tensor factorizations using the notion of the rank as the expressive power of the network; this line of work, combined with the proposed work (where we analyze the representations learned rather than the network coefficeints) can shed further light in understanding and characterizing deep networks.

3 Conclusions

In this paper we present an on-going line of research that is casting the behavior of deep networks (CNNs in particular) under the same analytical lens that has been used in a variety of data science tasks [4], aiming to characterize and understand how and what such networks learn. Understanding of this process can, subsequently, spark new ideas for theoretical research in the foundations of deep learning.

References

[1] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on Learning Theory*, pages 698–728, 2016.

[2] Richard A. Harshman. Parafac2: Mathematical and technical notes. *UCLA working papers in phonetics*, 22:30–44, 1972.

[3] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2673–2682, 2018.

[4] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[5] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. https://distill.pub/2018/building-blocks.

[6] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6078–6087, 2017.

[7] Hanie Sedghi, Vineet Gupta, and Philip M Long. The singular values of convolutional layers. *arXiv preprint arXiv:1805.10408*, 2018.

[8] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.