ResNet through Deep Multi-Layer Sparse Coding: A Theoretical Study

Tushar Chandra
Department of Computer Science and Engineering
IIIT Delhi
tushar21211@iiitd.ac.in

Animesh Pareek
Department of Electronics and Communication Engineering
IIIT Delhi
animesh21131@iiitd.ac.in

K. Gopi Krishna
Department of Computer Science and Engineering
IIIT Delhi
gopi23119@iiitd.ac.in

Abstract

We present a theoretical analysis of Residual Networks (ResNets) through the lens of Deep Matrix Factorization (DMF), framing them as a special case of multi-layer sparse coding. By establishing rigorous mathematical proofs, we validate the stability and convergence properties of ResNets, providing insights into their superior optimization landscapes and enhanced feature representations enabled by skip connections. Our analysis delves into the role of sparsity in neural architectures, demonstrating how it influences both learning efficiency and generalization. Furthermore, we highlight the interplay between activation functions, such as ReLU and Tanh, and the inherent sparsity properties of ResNets. This work bridges theoretical findings with practical implications, offering a framework for designing more robust and interpretable deep learning models. Finally, we discuss potential applications in sparse coding, model compression, and tasks requiring high-dimensional feature extraction.

1 Linear Network vs Resnet Layer

1.1 Linear Network

For a simple layer without activation, by definition of a simple Neural Layer, we have:

$$h^{(\ell+1)} = W^{(\ell)} h^{(\ell)} + b^{(\ell)}$$

where:

$$W^{(\ell)} \to \mathbb{R}^{n \times n}$$
$$h^{(\ell)} \to \mathbb{R}^{n}$$
$$h^{(\ell+1)} \to \mathbb{R}^{n}$$
$$b^{(\ell)} \to \mathbb{R}^{n}$$

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Thus, if $b^{(\ell)}$ is smaller than output $h^{(\ell+1)}$, then:

$$h^{(\ell+1)} \approx W^{(\ell)} h^{(\ell)}$$

Now, if we try to represent the output of this Layer via Sparse Representation, where 'D' is the Learned Dictionary and the 'α' is the sparse Representation of $h^{(\ell+1)}$, then:

$$h^{(\ell+1)} = D\alpha$$

Using the above two we can say that if $D = W^{(\ell)}$, then $\alpha = h^{(\ell)}$. Though, α in this case might not be an sparse representation.

1.2 Not sparse to sparse

Now, we describe the methodology for converting α into a sparse representation.

Mathematical Formation ($\forall i \in \{1, ..., n\}$) :

$$argmin_{D^{(\ell)}\alpha^{(\ell)}} ||h^{(\ell+1)} - D\alpha||_2^2 + \lambda||\alpha^{(\ell)}||_1, ||d_i^{(\ell)}||_2 = 1$$

or

$$argmin_{D^{(\ell)}\alpha^{(\ell)}} ||h^{(\ell+1)} - D\alpha||_2^2 + \lambda||\alpha^{(\ell)}||_0, ||d_i^{(\ell)}||_2 = 1$$

where $d_i^{(\ell)}$ represents i^{th} column of $D^{(\ell)}$.

This formulation can be efficiently solved using methods such as Basis Pursuit or LASSO. We employ K-SVD to solve this.

However, we need to understand that the finally learned Dictionary $D^{(\ell)}$ and the sparse vector $\alpha^{(\ell)}$ shall be analogs to $W^{(\ell)}$ and $h^{(\ell)}$, respectively. There by exhibiting almost similar properties.

1.3 Role of Activation

For some activation Function σ, the Linear Layer output is given by $z^{(\ell+1)} = \sigma(h^{(\ell+1)})$. Here $h^{(\ell+1)}$, from previous sections, $h^{(\ell+1)} = W^{(\ell)} h^{(\ell)} + b^{(\ell)}$.

Via Deep Matrix Factorization:

$$z^{(\ell+1)} = D'^{(\ell)} \alpha'^{(\ell)}$$

and

$$z^{(\ell+1)} = \sigma(D^{(\ell)} \alpha^{(\ell)})$$

where approximately $D^{(\ell)} = W^{(\ell)}$ and $\alpha^{(\ell)} = h^{(\ell)}$

Subject to the Constraint $||\alpha^{(\ell)}||_0 < k$ or $min||\alpha^{(\ell)}||_1$ and $||\alpha'^{(\ell)} \approx \alpha^{(\ell)}||$, We say that the Non-linearity does not implicitly affect sparsity model.

In-fact, introduction of Non-Linearity leads to adjustment of the dictionary atoms to account for this non-linear activation.

For any non-Linear Activation σ, we see the adjustment in Dictionary atoms in bounded direction of Non-Linearity introduced. For example, For RELU activation dictionary atoms are bounded to produce non-negative output given $\alpha^{(\ell)}$. Similarly for Tanh, dictionary atoms are bounded to produce output in range -1 to 1. Thus, for the above stated Equations, we can say the following:

$$z^{(\ell+1)} = S^{(\ell)} D^{(\ell)} h^{(\ell)}$$

Hence, $D'^{(\ell)} = S^{(\ell)} W^{(\ell)}$ and $\alpha'^{(\ell)} = h^{(\ell)}$, where $S^{(\ell)}$ represents the adjustment for Non-Linearity and belongs to $\mathbb{R}^{n \times n}$.

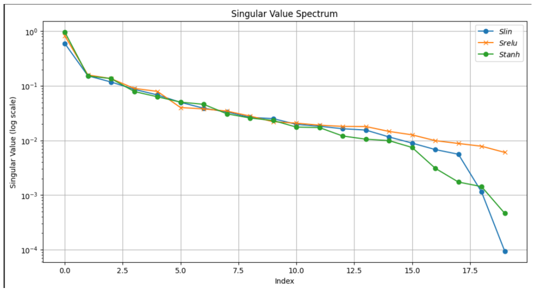


Figure 1: Spectral density of S matrix for different activations activations

1.4 ResNET Network

For a ResNet Layer, by definition of a simple Layer of a Vanilla Resnet, we have: $W^{(\ell)}$, then $\alpha = h^{(\ell)}$

$$h^{(\ell+1)} = h^{(\ell)} + \sigma(W^{(\ell)}h^{(\ell)} + b^{(\ell)})$$

where:

$$W^{(\ell)} \to \mathbb{R}^{n\times n}$$
$$h^{(\ell)} \to \mathbb{R}^{n}$$
$$h^{(\ell+1)} \to \mathbb{R}^{n}$$
$$b^{(\ell)} \to \mathbb{R}^{n}$$

where $S^{(\ell)} \to \mathbb{R}^{n\times n}$, accounting for non-linearity produced by σ. Thus, if $b^{(\ell)}$ is smaller than output $h^{(\ell+1)}$, then:

$$h^{(\ell+1)} \approx h^{(\ell)} + S^{(\ell)}W^{(\ell)}h^{(\ell)}$$

or,

$$h^{(\ell+1)} \approx (I_n + S^{(\ell)}W^{(\ell)})h^{(\ell)}$$

where I_n is an Identity matrix, $I_n \to \mathbb{R}^{n\times n}$

Now, if we try to represent the output of this Layer via Sparse Representation, where 'D' is the Learned Dictionary and the 'α' is the sparse Representation of $h^{(\ell+1)}$, then:

$$D^{(\ell)} \text{ is analogous to } I_n + S^{(\ell)}W^{(\ell)}$$

and

$$\alpha^{(\ell)} \text{ is analogous to } h^{(\ell)}$$

1.5 The Comparison

For normal networks:

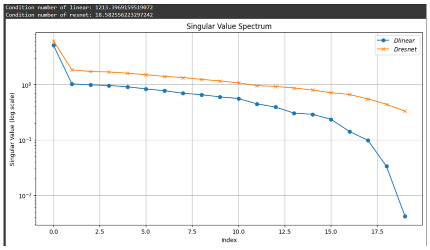
$$x^{(\ell)} = D^{(\ell)}x^{(\ell)} \implies x = Dz$$

For ResNets:

$$x^{(\ell+1)} = (I + D^{(\ell)})x^{(\ell)} \implies x = (I + D)z$$

Comparison of Sparsity:

- I_n, being a identity matrix , is inherently sparse :
$$||I_i||_0 = 1 \;\forall i = 1, .., n$$
$$||I_i||_1 = 1 \;\forall i = 1, .., n$$
- Now, since $D_{linear} = D$ and $D_{resnet} = I_n + D$, D_{resnet} have a higher structural Sparsity then D_{linear}.
- This Leads to Improved optimization



```
Condition number of linear: 1213.3969159519072
Condition number of resnet: 18.582556223297242
```

Figure 2: Spectral Density of D-linear vs D-resnet

Comparison of Rank:

- If $D = 0^{n \times n}$:
$$\text{Rank}(D) = 0, \text{Rank}(I + D) = \text{Rank}(I) = n$$
- Else If $\text{Rank}(D) = n$:
$$\text{Since Rank}(I) = n \text{ We have, Rank}(I + D) = n$$
- Else $0 < \text{Rank}(D) < n$:
$$\text{Since Rank}(I) = n \text{ We have, Rank}(I + D) > \text{Rank}(D)$$

Thus,
$$\text{Rank}(D') \geq \text{Rank}(D)$$

This also leads to another result, No. of Zero Eigen Values: $D_{linear} \geq D_{resnet}$.

Comparison of singular values:
$$\sigma_L(I) - \sigma_L(D) \leq \sigma_L(I + D) \leq \sigma_L(I) + \sigma_L(D)$$

Since, $\sigma_L(I) = 1$, we have,
$$1 - \sigma_L(D) \leq \sigma_L(I + D) \leq 1 + \sigma_L(D)$$

Thus, $\sigma_L(D)$is concentrated around 0 where as $\sigma_L(D')$ is concentrated around 1.

Also, Energy of features:
$$\text{Energy} \propto \sum_i \sigma_i^2$$

Therefore:

- D_{resnet} concentrates energy in fewer dimensions
- D_{linear} disperses energy across dimensions

Comparison of Condition no:

- Condition no. of a matrix 'A' is defined as $\kappa(A) = \sigma_{max}(A)/\sigma_{min}(A)$
- $\sigma_i(D_{linear}) \to 0$ and $\sigma_i(D_{resnet}) \to 1$
- Thus $\kappa(D_{linear}) >> \kappa(D_{resnet})$
- Adding Skip Connections, helps Resnet become more stable in optimization, due to condition no. reduction

So, to end with, Resnet Dictionarys exhibit a steeper singular spectrum, enhancing interpretability and generalization.

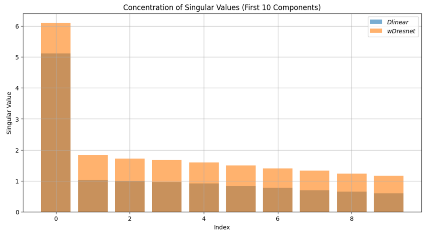


Figure 3: Concentration of energy in fewer dimensions

2 Multi Layer setup

2.1 Deep Coding Problem

For Multi-Layer Resnet Setup, With each layer as a matrix Factorization Step, we encode the problem as Deep Coding Problem DCP with ϵ denoting the error vector and λ denoting the sparsity vector.

Given Dictionaries $\{\mathrm{D}_i\}_{i=1}^{\ell}$,

$$DCP_{\lambda}^{\epsilon} : \text{find } \{\mathrm{T}_i\}_{i=1}^{\ell}$$

$$h^{(2)} : T_1 = D_1X \text{ , } ||T_1||_0 \leq \lambda_1 \text{ and } ||T_1 - D_1X|| < \epsilon_1$$

$$h^{(3)} : T_2 = D_2T_1 \text{ , } ||T_2||_0 \leq \lambda_2 \text{ and } ||T_2 - D_2T_1|| < \epsilon_2$$

$$\vdots$$

$$h^{(\ell+1)} : T_\ell = D_\ell T_{\ell-1} \text{ , } ||T_\ell||_0 \leq \lambda_\ell \text{ and } ||T_\ell - D_\ell T_{\ell-1}|| < \epsilon_\ell$$

2.2 Deep Learning Problem

For Multi-Layer Resnet Setup, we aim to now learn Dictionaries inorder to minimize the loss governed by some loss function f.

Given λ, the sparsity vector,

$$DLP_{\lambda} : \min_{\{D_i\}_{i=1}^{\ell}} \sum_j f(h(X_j), T_\ell, DCP_{\lambda}^{*}(X_j, \{D_i\}_{i=1}^{\ell}))$$

where $h(x)$ represents the True Label Function.

2.3 Uniqueness Analysis

According to Papyan et al. (2016a), for a solution to be unique in the $P_{\epsilon,\mu}^{\infty}$ problem:

$$\|T\|_\infty < \frac{1}{2}(1 + \frac{1}{\mu(D)})$$

For our model $T_\ell = D_\ell T_{\ell-1}$, we derive:

$$\|T_{\ell-1}\| \leq \lambda_{\ell-1} < \frac{1}{1 + \mu(D_\ell)}$$

$$\|T_{\ell-2}\| \leq \lambda_{\ell-2} < \frac{1}{1 + \mu(D_{\ell-1})}$$

$$\vdots$$

$$\|T_1\|_0 \leq \lambda_1 < \frac{1}{1 + \mu(D_2)}$$

Thus, if λ_i is smaller than $\frac{1}{2}(1 + \frac{1}{\mu(D_i)})$, then our model is guaranteed to have a unique solution.

2.4 Global Stability of DCP problem

For Output, if desired output was Y and predicted output is $\hat{Y}$, then

$$Y = \hat{Y} + E$$

$$Y = T_\ell + E = D_\ell T_{\ell-1} + E$$

If given,

- $||T_\ell - \hat{T}_\ell||_2 \leq \epsilon_\ell$
- $||T_{\ell-1}||_0 < \frac{1}{2}(1 + \frac{1}{\mu(D_\ell)})$ and $||E||_2 = ||Y - D_\ell T_{\ell-1}||_2 \leq \epsilon_\ell$
- $||\hat{T}_{\ell-1}||_0 < \frac{1}{2}(1 + \frac{1}{\mu(D_\ell)})$ and and $||Y - D_\ell \hat{T}_{\ell-1}||_2 \leq \epsilon_\ell$

(Note this also means that $||T_\ell - \hat{T}_\ell||_2 \leq \epsilon_\ell$)

then by Papyan et al. (2016b),

$$\|\Delta_{\ell-1}\|_2^2 = \|T_{\ell-1} - \hat{T}_{\ell-1}\|_2^2 \leq \epsilon_{\ell-1}^2 = \frac{4\epsilon_\ell^2}{1 - (2\|T_{\ell-1}\|_0 - 1)\mu(D_\ell)}$$

Now,

$$T_{\ell-1} = \hat{T}_{\ell-1} + \Delta_{\ell-1}$$

$$T_{\ell-1} = D_{\ell-1} T_{\ell-2} + \Delta_{\ell-1}$$

Since $T_{\ell-2}$ and $||T_{\ell-2}||$ is bounded by mutual coherence of Dictionary $D_{\ell-1}$, we have

$$\|\Delta_{\ell-2}\|_2^2 = \|T_{\ell-2} - \hat{T}_{\ell-2}\|_2^2 \leq \epsilon_{\ell-2}^2 = \frac{4\epsilon_{\ell-1}^2}{1 - (2\|T_{\ell-2}\|_0 - 1)\mu(D_{\ell-1})}$$

Similarly,

$$\|\Delta_{\ell-3}\|_2^2 = \|T_{\ell-3} - \hat{T}_{\ell-3}\|_2^2 \leq \epsilon_{\ell-3}^2 = \frac{4\epsilon_{\ell-2}^2}{1 - (2\|T_{\ell-3}\|_0 - 1)\mu(D_{\ell-2})}$$

$$\|\Delta_{\ell-4}\|_2^2 = \|T_{\ell-4} - \hat{T}_{\ell-4}\|_2^2 \leq \epsilon_{\ell-4}^2 = \frac{4\epsilon_{\ell-3}^2}{1 - (2\|T_{\ell-4}\|_0 - 1)\mu(D_{\ell-3})}$$

$$\vdots$$

$$\|\Delta_1\|_2^2 = \|T_1 - \hat{T}_1\|_2^2 \leq \epsilon_1^2 = \frac{4\epsilon_2^2}{1 - (2\|T_1\|_0 - 1)\mu(D_2)}$$

So $\forall i \in 1, ..., n$, Error is bounded by the maximum error, proving the Global Stability of the DCP_λ^ϵ problem.

2.5 Error Propagation

Given noisy input, Contaminated by a a noise, E , having $||E||_2 = \epsilon_0$. Then,

$$Y = X + E$$

$$T_1 = D_1 X$$

$$\hat{T}_1 = D_1 y = D_1 X + D_1 E$$

$$\hat{T}_1 = T_1 + D_1 E$$

or,

$$\|\hat{T}_1 - T_1\|_2^2 \leq \|D_1\|_2^2 \|E\|_2^2$$

$$\|\hat{T}_1 - T_1\|_2^2 \leq \epsilon_1^2 \leq \|D_1\|_2^2 \epsilon_0^2$$

but in previous subsection we derived an expression of ϵ_1 w.r.t. ϵ_2,which was,

$$\epsilon_1^2 = \frac{4\epsilon_2^2}{1-(2\|T_1\|_0-1)\mu(D_2)}$$

Therefore,

$$\frac{4\epsilon_2^2}{1-(2\|T_1\|_0-1)\mu(D_2)} \leq \|D_1\|_2^2\epsilon_0^2$$

By Sequentially Superimposing this with multiple expressions of ϵ_i from $i = 2 \rightarrow \ell$, we get,

$$||\Delta_\ell\|_2^2 = \|T_\ell - \hat{T}_\ell\|_2^2 \leq \epsilon_\ell^2 \leq \frac{||D_1||_2^2\epsilon_0^2}{4}\prod_{i=1}^{\ell-1} A_i$$

where $A_i = 1-(2\|T_i\|_0-1)\mu(D_{i+1})$

So,$||\Delta_\ell||_2^2$ is bounded by this quantity.

Thus, If 'X' is contaminated by some error bounded by ϵ_0,then $||\Delta_\ell||_2^2$ (squared Error in output), is bounded by $||D_1||_2^2$ and mutual coherence of other dictionaries.

2.6 Empirical verification of model

Compare:

- resnet+batch norm
- resnet
- linear model
- multi-layer resnet sparsenmatrix model

Comparison to be done w.r.t https://www.ux.uis.no/ karlsk/Skretting_spie11.pdf

3 Conclusion

We have tried to explain why ResNets achieve superior performance compared to traditional architectures, via DMF theory.Also we have proposed a stable model. By systematically comparing ResNets to linear networks, we highlight how the incorporation of skip connections and sparsity properties leads to improved optimization, stability, and generalization. The integration of non-linear activations, their impact on dictionary adjustments, and the enhanced spectral properties of ResNet layers underline their ability to concentrate feature energy and achieve efficient representations. Our findings not only bridge the gap between theoretical constructs like Deep Matrix Factorization and practical implementations in deep learning but also offer insights into the structural advantages of ResNets. Through rigorous comparisons in terms of sparsity, rank, singular values, and condition numbers, we establish ResNets as a robust alternative to conventional architectures.

4 Contributions

- Tushar Chandra: Help in writing Latex File
- Animesh Pareek: Mathematical philosophy
- K. Gopi krishna: Was Supposed to do implementations