A Preliminary Study of the Norm Preservation Properties of CNN

Wenling Shang*

Abstract

Important Notice: This work is in its preliminary stage; this brief technical report is also fairly informal and rushed due to external factors. Nonetheless, by public request, we decide to present our findings–both positive and negative–during our initial investigation on this subject for other researchers to cite our results. We will hopefully follow up with major modifications in the near future.

Recently, deep convolutional neural nets (CNNs) have gained much popularity in subfields of Computer Vision such as object classification, object detection, caption generation and activity recognition. However, there has been limited effort directed toward theoretically explaining why CNNs can generate desirable features. In this technical report, under some major assumptions on the input space and the learned filters, we attempt to prove analytical upper and lower bounds for the l^2 norm of CNN features (convolution+relu+max-pooling) with respect to the l^2 norm of the input features. Unfortunately, the assumptions on the filters are not generally satisfied; it would be more practical to measure the norm preservation property based on the numerical values of individual filter matrices[1].

1 Introduction

Convolutional Neural Networks (CNNs) have recently achieved state-of-the-art results in various vision tasks [2, 3]. Researchers have conducted experiments [4] and built visualization tools [5] to heuristically illustrate the internal operations of these models. However, fundamental understanding of CNNs from a theoretical perspective under realistic assumptions is lacking, which is unsatisfactory from a mathematical point of view. Throughout the paper, we denote the convolution+relu by $f_{\text{cnn}}(\cdot)$ and the input by x. In this paper,

1. We conduct mathematical analysis on the highly nonlinear convolutional layers by deriving a norm relationship between the input features, x, and the output CNN features, $f_{\text{cnn}}(x)$ or $f^c_{\text{cnn}}(x)$, under reasonable assumptions.
2. We attempt to empirically verify the assumptions but there is still a gap between the theoretical assumptions and the learned convolution filters.

*University of Michigan, Ann Arbor, MI. shangw@umich.edu.

1.1 Related Work

Bengio et al. [6] summarize what constitute good representations, among which are *abstractness*, *expressiveness* and *robustness*. Effectively-trained CNNs [2, 7] generate high quality features that succeed in various vision tasks, such as object classification [2] and object detection [3]. Can we mathematically justify the performance of CNN features? The *abstractness* of CNN features comes from the stacked non-linear convolutional layers. For *expressiveness* and *robustness*, we may approximately translate them into *the norm preserving properties* in the mathematical language, $C_1\|x\|_2^2 \leq \|f_{\text{cnn}}(x)\|_2^2 \leq C_2\|x\|_2^2$, recall that $f_{\text{cnn}}(\cdot)$ represents the convolution+relu+max-pooling operation and x the input. In other words, we would like to map fairly distinct inputs to fairly distinct CNN features (expressiveness), which requires C_1 to be large. At the same time, small perturbations on the inputs should not dramatically alter the CNN features (robustness), which requires C_2 to be small.

In our theoretical analysis, we propose analytical yet non-trivial evaluations on C_1 and C_2. The assumptions for our theoretical analysis are consistent with some conceptual understanding of how CNNs work but are not empirically satisfied. **Section 2.3** describes our experimental verification of these assumptions. Similar results can also be derived for the concatenated relu [15], where relu is operated on the modulated negative and positive linear responses separately, with even slightly assumptions than the relu case.

Among existing theoretical works on neural networks, Bruna et al. [8] investigate conditions that assure injectivity and measure the stability of the inverting process by computing the Lipschitz lower bound for a l^p pooling proceeded by relu when $p \in \{1, 2, \infty\}$. In another related work, Szegedy et al. [1] empirically calculate the upper frame bounds for various layers of AlexNet [2], a benchmark CNN trained on a large scale image dataset, ImageNet [9].

2 Mathematical analysis

The highly nonlinear nature of convolutional networks presents a challenge in relating an input signal, x, to its CNN features after going through the network, $f_{\text{cnn}}(x)$. For a very special case, *Invariant Scattering Convolution Networks* [10], where the filters are wavelet transforms followed by a modulus, the authors show that the corresponding convolution layer is non-expansive and Lipschitz continuous to deformations. However, the derivation of their results relies on the nature of wavelet transformation, which is unique to scattering networks. On the other hand, to our knowledge, similar properties for the conventional convolutional+relu+max-pooling layers in CNN have not been properly explored. In [8], the authors characterize the general Lipschitz lower bounds for rectification layers followed by pooling operators. Ultimately, their goal is to identify the conditions under which such relu+pooling procedure is injective and can be stably inverted. In their analysis on max-pooling, the pooling positions, a.k.a the switch units, are not recorded. Hence the relu+max-pooling is viewed as a phaseless map. The essential theoretical contributions from [8] define conditions under which input signals can be recovered from the phaseless representations. Their setting is different from ours: we do not eliminate the switch unit information. We conjecture that CNN features contain the "content" information whereas switch units contain the "location" information. In other related work, empirical Lipschitz upper bounds are approximated for AlexNet [1], in order to investigate the stability of the feed-forward network, though without an analytical characterization; the levels of invariance in several deep networks have also been measured empirically [11].

In this section, firstly—in Section 2.1 and Section 2.2—we mathematically establish bounds between $\|x\|_2$ and $\|f_{\text{cnn}}(x)\|_2$ as well as discuss the significance of our results. Next—in Section 2.3—we conduct experiments to evaluate the validity of the assumptions proposed for our main theoretical analysis.

2.1 Problem Setup and Supporting Tools To simplify our notation, we assume zero bias and 1D vectors as input rather than 2D images. **Problem Setup.** Let an input $x \in \mathbb{R}^D$. Let w_i, for $i = 1, \ldots, K$, be convolutional filters; each is of length $\ell \leq D$ and each will be supported on the ℓ consecutive coordinates of row vectors in $\mathbb{R}^D$. We will be considering coordinate shifts of these convolutional filters; all shift lengths will be multiples of a fixed "stride length", which we call s. We assume here that $D - \ell$ is divisible by s, and denote $n = (D - \ell)/s + 1$–if there is no max-pooling then $n = 1$. Let w_i^j, for $j = 1, \ldots, n$, be w_i shifted to the right by $(j-1)s$ coordinates from the 1st coordinate, which we refer to as shifts of w_i. Finally, the shifts have unit norm, $\|(w_i^j)^T\|_2 = 1$. Our **first major assumption** (A1) is that the shifts of any pair of distinct convolutional filters have low coherence:[1]

$$\text{(A1)} \qquad \begin{aligned} &\text{there exists } 0 \leq \mu \leq \frac{1}{K-1}, \text{ such that} \\ &\left| \left\langle (w_i^j)^T,\ (w_p^q)^T \right\rangle \right| \leq \mu, \text{ for } i \neq p. \end{aligned}$$

We define W to be the $Kn \times D$ matrix whose rows are the shifts w_i^j; the rows of W will be divided into K blocks of n rows each, with each block consisting of all the shifts of a fixed filter w_i. As common practice, the hidden units of the feed-forward CNN are computed by first multiplying an input signal $x \in \mathbb{R}^D$ by the matrix W, zeroing out any negative activations (relu), and then, for each of the K blocks, selecting the maximum value in that block (max-pooling). We denote the relu and max-pooling operations together as g. Hence $f_{\text{cnn}} = g \circ W : \mathbb{R}^D \to \mathbb{R}^K$.

Furthermore, we assume that the input domain V is a subset of range(W^T), the subspace spanned by the transpose of the shifts, and that:

$$\text{(A2.1)} \qquad \begin{aligned} &\forall x \in V,\ \exists\ \{c_i^j\}_{i=1,\cdots,K}^{j=1,\cdots,n}, \text{ such that} \\ &x = \sum_{i=1}^{K} \sum_{j=1}^{n} c_i^j (w_i^j)^T, \text{ where } c_i^j \geq 0; \end{aligned}$$

$$\text{(A2.2)} \qquad \forall i, \text{ at most one of the } c_i^j \text{ is positive;}$$

$$\text{(A2.3)} \qquad \begin{aligned} &\text{if } c_p^q > 0, \text{ then } \frac{c_p^q}{\|\boldsymbol{c}\|_1 - c_p^q} > \mu, \\ &\text{where } \|\boldsymbol{c}\|_1 = \sum_{i=1}^{K} \sum_{j=1}^{n} c_i^j. \end{aligned}$$

This is our **second major assumption**. It consists of three parts. Firstly, Assumption A2.1 is equivalent of claiming that the shifts span the input space, which contains patches over a pooling-region. The feasibility of inverting CNN features back to the original images [12, 13] essentially requires Assumption A2.1 to be valid. The motivation for A2.2 directly comes from max-pooling and relu; it is also biologically plausible [6]. Assumption A2.3 implies sparsity, a natural prior for coding images, by only allowing strong activations. Heuristically, we assume that the training of CNN primarily produces

[1] See Appendix B Definition A.1.

convolutional filters that can significantly contribute to the construction of input signals and are discriminative. The following lemma,[2] as a consequence of A2.3, is used in the proof to our main theorem:

LEMMA 2.1. *Let $x \in V$; if $c_p^q > 0$, then $\langle (w_p^q)^T, x\rangle > 0$*
.

Finally, for an input $x \in V$, we introduce the notation W_x to denote the matrix consisting of the set of $\{w_i^j\}$, as row vectors, whose corresponding c_i^j's are positive, and denote the vector consisting of the positive c_i^j's by $\boldsymbol{c}_x$, i.e. $W_x^T \boldsymbol{c}_x = x$. Denote the number of rows in W_x by K_a. Similarly, we introduce the notation $\hat{W}_x$ to denote the matrix consisting of the set of $\{w_i^j\}$ which are activated after relu and max-pooling on the input x, as row vectors, and let $\hat{\boldsymbol{c}}_x$ be the activations, i.e. $\hat{W}_x x = \hat{\boldsymbol{c}}_x$ and $\|\hat{\boldsymbol{c}}_x\|_2 = \|f_{\text{cnn}}(x)\|_2$. Denote the number of rows in $\hat{W}_x$, by $\hat{K}_a$.

Tools from Frame Theory. In our proofs, we use some tools from signal processing, more specifically, from Frame Theory [14]. Important definitions and propositions are introduced in Appendix A. We will use the notation from Appendix A throughout the rest of the paper.

2.2 Norm Bounds.

Bounds Between $\|\cdot\|_2$ and $\|f_{\text{cnn}}(\cdot)\|_2$. In our setting, for an input x, the transpose of the rows of W_x can be viewed as a frame for the subspace they span, which contains x. Therefore, $\mathcal{T}_x$, the corresponding *Analysis Operator*, has the matrix representation $\tilde{\mathcal{T}}_x = W_x U_x$, where U_x is an orthonormal bases for range(W_x^T). The representation of the input x under U_x is thus $U_x^T x$. Note that the transformed shifts after the change of basis—the row vectors of $\tilde{\mathcal{T}}_x$—still satisfy the two assumptions: (1) are normalized to have l^2 norm = 1 and (2) have low coherences with other transformed shifts, i.e., $|\langle (\tilde{\mathcal{T}}_x)_i^T, (\tilde{\mathcal{T}}_x)_j^T\rangle| < \mu$, if $i \neq j$, where $(\tilde{\mathcal{T}}_x)_i$ is the ith row of $\tilde{\mathcal{T}}_x$. Furthermore, $\tilde{\mathcal{S}}_x = \tilde{\mathcal{T}}_x^* \tilde{\mathcal{T}}_x$ and $\tilde{\mathcal{S}}'_x = \tilde{\mathcal{T}}_x \tilde{\mathcal{T}}_x^*$ share the same non-zero eigenvalues. It is easier to analyze the eigenvalues of the latter, since (1) $\tilde{\mathcal{S}}'_x$ has diagonal entries to be ones and (2) its off-diagonal entries have absolute value at most μ. An important step in deriving the norm relationship is to bound the least and largest eigenvalues of $\tilde{\mathcal{S}}'_x$. Note that for $\tilde{\mathcal{S}}'_x$ to have a non-zero least eigenvalue, i.e. for this bound to be meaningful, the K frames need to be linearly independent ($\tilde{\mathcal{S}}_x = \tilde{\mathcal{S}}'_x$).

THEOREM 2.1. [3] *Let $A \in R^{K\times K}$ ($K \geq 2$) be a symmetric matrix, with ones on the diagonal and the off-diagonal entries satisfying $|a_{ij}| \leq \mu < \frac{1}{K-1}$. Then the least eigenvalue of A, denoted by λ_K is bounded below,and the largest eigenvalue, denoted by λ_1, is bounded above, respectively by*

$$\lambda_1 \leq 1 + (K-1)\mu, \qquad \lambda_K \geq 1 - (K-1)\mu.$$

Now we are ready to show **our main theorem**. The proof is in Appendix B.

THEOREM 2.2.

$$(1 - (K_a - 1)\mu)\|x\|_2^2 \leq \|f_{\text{cnn}}(x)\|_2^2 \leq (1 + (\hat{K}_a - 1)\mu)\|x\|_2^2, \ \forall x \in V.$$

Notice that if we substitute K_a and $\hat{K}_a$ with K, the resulting inequality, though potentially loosened, still holds.

Bounds Between $\|\cdot\|_2$ and $\|f_{\text{cnn}}^c(\cdot)\|_2$. Similar norm bounds can be obtained if replacing relu non-linearity with *Concatenated Rectified Linear Units* (crelu) recently proposed by Shang et al. [15]. The crelu setting alleviates some aspects of the assumptions from the relu setting. Since crelu preserves not only positive but also negative phase information, assumption A2.1 no longer requires $c_i^j \geq 0$. As a consequence, for each i, we only require at most one of the c_i^j is non-zero. Moreover, assumption A2.3 can be dropped because its resulting lemma, Lemma 2.1, is unnecessary under the crelu setting. The proof highly resembles that of Theorem 2.2 and we will not emphasize the logic again.

Significance At a high level, the above norm bounds—similar to the frame constants—measure the stability of the transformation under f_{cnn}. The upper bound reflects how stable, or non-expansive, the feed-forward process is, i.e. the degree of continuity of f_{cnn}. The smaller the upper bound is, the more smooth or continuous f_{cnn} is and the less likely there are to be adversarial examples—inputs that are close to each other in the input space but have drastically different representations in the output feature space. The lower bound reflects how stable the inverting process is, i.e. given two distinct inputs, how easy it is to distinguish them in the output feature space.

Note, it is desirable to have both small μ as well as small K_a and $\hat{K}_a$. One exception is if two filters are of "opposite phase", μ does not necessarily need to be

[2] The proof can be found in Appendix B.

[3] The proof is in Appendix B.

small because of relu. In this case, an input tends to only have positive response with one of the two filters and does not activate the other one, hence our analysis is not jeopardized. Note that in the case of crelu, there is no such "opposite phase" phenomena thanks to its construction. To sum, the *practical implication* is that the *number of convolutional filters* should be large enough such that the most important inherent directions in the input space, conditioned on the AI task, can be well captured in a sparse manner, yet the filters ought not be overcomplete in order to keep μ small.

An example of leveraging such norm preservation property to enhance training is the Layer-sequential unit-variance (LSUV) initialization [16]. LSUV initializes the network with $\mu = 0$ and regularizes the norm of the activation by adjusting the learned weights so that the output variance to be 1.

We also noticed that using non-negative sparse coding framework to initialize CIFAR-10 networks can marginally improve upon relu baseline results. However the effect on such initialization is almost none on AlexNet–agreeing with the LSUV initialization. Hence, we will not report the relevant results.

2.3 Verification of Assumptions The assumptions on the input space (Assumption A2.1 to A2.3) are tricky to directly verify. Thus we only focus on Assummption A1. We examine the All-Conv model from [17] and its crelu variant [15] trained on ImageNet for classification. The architecture of All-Conv model is similar to that of AlexNet [2] but with no max-pooling (i.e. $n = 1$) or fully connected layers. Our experiments intended to justify **Assumption A1**. We looked at the first 9 convolution layers from All-Conv model. For each filter w_i from each convolution layer, we normalized it to have ℓ^2 norm 1, compute $\mu_{ij} = |\langle w_i, w_j \rangle|$ with other distinct convolution filters from the same layer, calculate the mean across all $\mu_{ij,i\neq j}$ (with standard deviation) in Table 1 and plot the histogram of $\langle w_i, w_j \rangle$ in Figure 1.

We observed that μ_{ij} for the relu based All-Conv model is not bounded above by $1/K$ from Assumption A1 where K is the number of convolution filters at the corresponding layer. Since relu convolution filters tend to form negatively correlated pairs, we suspected it could be one of the factors leading to the large averaged μ_{ij}'s. However, when we conducted the same experiments with crelu All-Conv model, where such "pairing phenomenon" is not present thanks to the activation scheme, μ_{ij}'s are still substantially bigger than $1/K$ (Table 2 and Figure 2). Nonetheless, it is worth noting, on the other hand, the averaged μ_{ij}'s are significantly smaller than the relu model. Hence, crelu potentially learns more diverse filters than relu.

We came to a conclusion that Assumption A1 is generally not satisfied and the convolution filters are not as "orthogonal" as many theorists wish them to be. Furthermore,crelu filters tend to be more incoherent with one another than the relu counterparts.

3 Summary

This brief technical report mathematically characterizes the norm-preserving property of CNNs and illustrates its practical implications. Unfortunately, the convolution filters are not incoherent enough to meet our theoretical assumptions. There is always a gap between theory and reality, right? :)

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2013.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[4] J. Yosinski, J. Clune, Y. Bengio, and Lipson H. How transferable are features in deep neural networks? *NIPS*, 2014.

[5] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014.

[6] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *PAMI*, 2013.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: delving deep into convolutional nets. In *BMVC*, 2014.

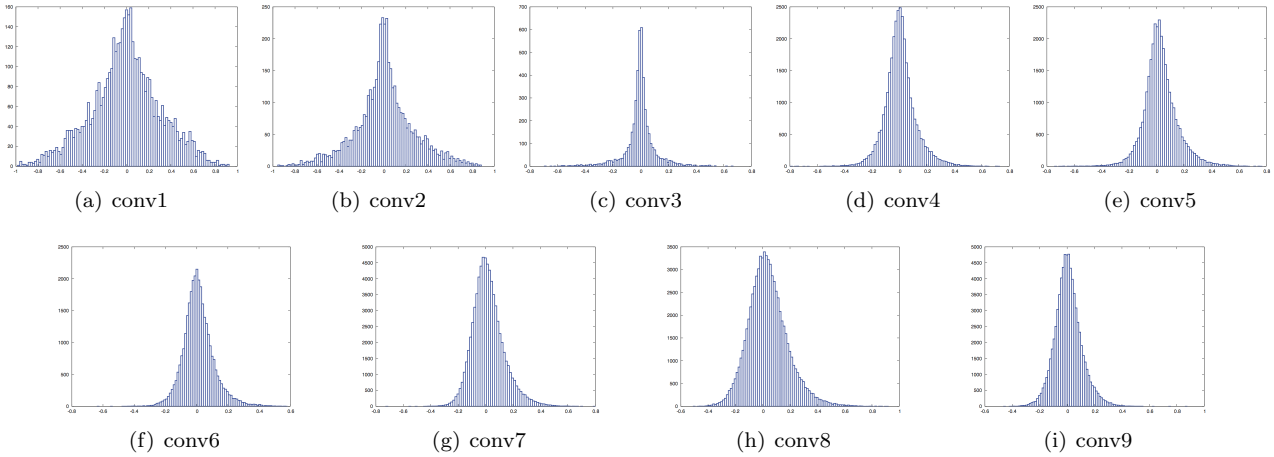
[8] J. Bruna, A. Szlam, and Y. LeCun. Signal recovery from pooling representations. *ICML*, 2013.

[9] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] J. Bruna and S. Mallat. Invariant scattering convolution networks. *PAMI*, 2013.

Table 1: μ_{ij} for ImageNet All-Conv Model with relu

Layer Index	1	2	3	4	5	6	7	8	9
average $\mu_{ij,i\neq j}$	0.240	0.194	0.068	0.082	0.091	0.073	0.087	0.113	0.075
std	0.200	0.183	0.090	0.080	0.089	0.068	0.078	0.098	0.065
$1/K$	0.01	0.01	0.01	0.004	0.004	0.004	0.0026	0.0026	0.0026



(a) conv1 (b) conv2 (c) conv3 (d) conv4 (e) conv5

(f) conv6 (g) conv7 (h) conv8 (i) conv9

Figure 1: The distribution of $\langle w_i, w_j \rangle$ where $i \neq j$ for the first 9 convolution layers from All-Conv Model with relu trained on ImageNet.

[11] I. Goodfellow, H. Lee, Q. Le, A. Saxe, and A. Ng. Measuring invariances in deep networks. In *NIPS*, 2009.

[12] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. *CVPR*, 2015.

[13] A. Dosovitskiy and T. Brox. Inverting Convolutional Networks with Convolutional Networks. *ArXiv e-prints*, June 2015.

[14] O. Christensen. *An introduction to frames and Riesz bases*. 2003.

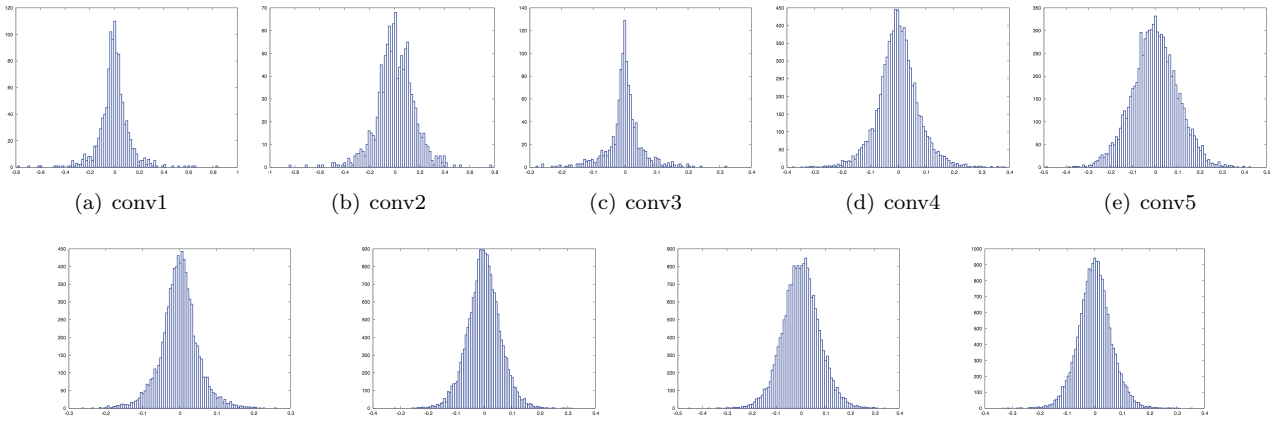
[15] W. Shang, Almeida D. Sohn, K., and H. Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *ICML*, 2016.

[16] Matas J. Mishkin, D. All you need is a good init. In *ICML*, 2014.

[17] Jost Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR Workshop*, 2014.

Table 2: μ_{ij} for ImageNet All-Conv Model with crelu

Layer Index	1	2	3	4	5	6	7	8	9
average $\mu_{ij,i\neq j}$	0.081	0.112	0.035	0.054	0.080	0.037	0.045	0.058	0.046
std	0.094	0.100	0.043	0.048	0.063	0.032	0.037	0.046	0.038
$1/K$	0.02	0.02	0.02	0.008	0.008	0.008	0.0052	0.0052	0.0052



(a) conv1 (b) conv2 (c) conv3 (d) conv4 (e) conv5

(f) conv6 (g) conv7 (h) conv8 (i) conv9

Figure 2: The distribution of $\langle w_i, w_j \rangle$ where $i \neq j$ for the first 9 convolution layers from All-Conv Model with crelu trained on ImageNet.

Appendix

A Mathematical Tools

We only consider finite vector spaces in our analysis.

DEFINITION A.1. *Let $u, v \in \mathbb{R}^D$. The* correlation *between u and v is defined as*

$$corr(u,v) = \frac{\langle u, v\rangle}{\|u\|_2\|v\|_2}. \tag{S1}$$

The coherence *μ between u and v is defined as*

$$\mu = \frac{|\langle u, v\rangle|}{\|u\|_2\|v\|_2}. \tag{S2}$$

DEFINITION A.2. *A frame is a set of elements of a vector space $\mathcal{V}$, $\{\phi_k\}_{k=1,\cdots,K}$, which satisfies the frame condition: there exist two real numbers C_1 and C_2, the* frame bounds*, such that $0 < C_1 \le C_2 < \infty$, and $\forall v \in \mathcal{V}$*

$$C_1\|v\|_2^2 \le \sum_{k=1}^{K} |\langle v, \phi_i\rangle|^2 \le C_2\|v\|_2^2.$$

PROPOSITION A.1. *Let $\{\phi_k\}_{k=1,\cdots,K}$ be a sequence in $\mathcal{V}$, then $\{\phi_k\}$ is a frame for* span$\{\phi_k\}$*. Hence, $\{\phi_k\}$ is a frame for $\mathcal{V}$ if and only if $\mathcal{V} =$* span$\{\phi_k\}$[4].

DEFINITION A.3. *Consider now $\mathcal{V}$ equipped with a frame $\{\phi_k\}_{k=1,\cdots,K}$. The* Analysis Operator*, $\mathcal{T} : \mathcal{V} \to \mathbb{R}^K$, is defined by $\mathcal{T}v = \{\langle v, \phi_k\rangle\}$. The* Synthesis Operator*, $\mathcal{T}^* : \mathbb{R}^K \to \mathcal{V}$, is defined by $\mathcal{T}^*\{c_k\}_{k=1,\cdots,K} = \sum_{k=1}^{K} c_k\phi_k$, which is the adjoint of the Analysis Operator. The* Frame Operator*, $\mathcal{S} : \mathcal{V} \to \mathcal{V}$, is defined to be the composition of $\mathcal{T}$ with its adjoint:*

$$\mathcal{S}v = \mathcal{T}^*\mathcal{T}v.$$

The Frame Operator is always invertible.

THEOREM A.1. *The optimal lower frame bound C_1 is the smallest eigenvalue of $\mathcal{S}$; the optimal upper frame bound C_2 is the largest eigenvalue of $\mathcal{S}$.*

[4]There exist infinite spanning sets that are not frames, but we will not be concerned with those here since we only deal with finite dimensional vector spaces.

We would also like to investigate the matrix representation of the operators $\mathcal{T}, \mathcal{T}^*$ and $\mathcal{S}$. Consider $\mathcal{V}$, a subspace of $\mathbb{R}^D$, equipped with a frame $\{\phi_k\}_{k=1,\cdots,K}$. Let $U \in R^{D\times d}$ be a matrix whose column vectors form an orthonormal basis for $\mathcal{V}$ (here d is the dimension of $\mathcal{V}$). Choosing U as the basis for $\mathcal{V}$ and choosing the standard basis $\{e_k\}_{k=1,\cdots,K}$ as the basis for R^K, the matrix representation of $\mathcal{T}$ is $\tilde{\mathcal{T}} = TU$, where T is the matrix whose row vectors are $\{\phi_k^T\}_{k=1,\cdots,K}$. Its transpose, $\tilde{\mathcal{T}}^*$, is the matrix representation for $\mathcal{T}^*$; the matrix representation for $\mathcal{S}$ is $\tilde{\mathcal{S}} = \tilde{\mathcal{T}}^*\tilde{\mathcal{T}}$.

B Proofs to Our Theoretical Results

B.1 Proof of Lemma 2.1

LEMMA B.1. *2.1 Let $x \in V$; if $c_p^q > 0$, then $\langle (w_p^q)^T, x\rangle > 0$.*

Proof.

$$\begin{aligned}\langle (w_p^q)^T, x\rangle &= \langle (w_p^q)^T, \sum_{i=1}^{K}\sum_{j=1}^{n} c_i^j (w_i^j)^T\rangle \\ &= c_p^q + \sum_{i=1, i\neq p}^{K}\sum_{j=1}^{n} c_i^j \langle (w_i^j)^T, (w_p^q)^T\rangle \\ &\ge c_p^q - \sum_{i=1, i\neq p}^{K}\sum_{j=1}^{n} c_i^j \mu = c_p^q - (\|\boldsymbol{c}\|_1 - c_p^q)\mu > 0.\end{aligned}$$

B.2 Proof of Theorem 2.1

THEOREM B.1. *2.1 Let $A \in R^{K\times K}$ ($K \ge 2$) be a symmetric matrix, with ones on the diagonal and the off-diagonal entries satisfying $|a_{ij}| \le \mu < \frac{1}{K-1}$. Then the least eigenvalue of A, denoted by λ_K is bounded below,and the largest eigenvalue, denoted by λ_1, is bounded above, respectively by*

$$\lambda_1 \le 1 + (K-1)\mu, \qquad \lambda_K \ge 1 - (K-1)\mu.$$

Proof. By the Gershgorin circle theorem, the fact that A is symmetric with ones on the diagonal and off-diagonal entries bounded by $|a_{ij}| < \mu$ for $i \neq j$, all eigenvalues λ of A satisfy

$$|\lambda - 1| < \sum_{j=1, j\neq i}^{K} |a_{ij}| \le (K-1)\mu.$$

Therefore,

$$\lambda_1 \le 1 + (K-1)\mu, \qquad \lambda_K \ge 1 - (K-1)\mu.$$

B.3 Proof of Theorem 2.2

THEOREM B.2. *2.2*

$$(1-(K_a-1)\mu)\|x\|_2^2 \leq \|f_{\text{cnn}}(x)\|_2^2 \leq (1+(\hat{K}_a-1)\mu)\|x\|_2^2,\ \forall x \in V.$$

Proof. By definition, $x \in \text{span}\{(W_x)_i^T\} = \text{range}(W_x^T)$. By Proposition A.1, the transpose of the row vectors in W_x form a frame for range(W_x^T). Let U_x be an orthonormal basis for range(W_x^T). Then the matrix representation under U_x for the Analysis Operator, $\mathcal{T}_x$, is $\tilde{\mathcal{T}}_x = W_x U_x$, and the corresponding representation for x under U_x is $U_x^T x$. Now, by Theorem A.1, we have:

$$C_1\|x\|_2^2 \leq \|\tilde{\mathcal{T}}_x x\|_2^2 = \|W_x U_x U_x^T x\|_2^2 = \|W_x x\|_2^2,$$

where C_1 is the least eigenvalue of $\tilde{\mathcal{S}}_x$. By Theorem 2.1, $C_1 \geq 1-(K_a-1)\mu$.

Recall that $\hat{W}_x$ consists of the activated filters. By the definition of max-pooling and Lemma 2.1, we have

$$C_1\|x\|_2^2 \leq \|W_x x\|_2^2 \leq \|\hat{W}_x x\|_2^2 = \|g(Wx)\|_2^2 = \|f_{\text{cnn}}(x)\|_2^2.$$

In relation to the second step in the inequality, Lemma 2.1 guarantees that every element in $W_x x$ is positive. By relu and max pooling, the activated rows in $\hat{W}_x x$ will be at least as large as their counterparts in $W_x x$.

Rewrite $x = x_p + x_p^c$, where $x_p \perp x_p^c$ and $x_p \in \text{range}(\hat{W}_x^T) = \ker(\hat{W}_x)^\perp$, $x_p^c = x - x_p \in \ker(\hat{W}_x)$. Thus, the transpose of the row vectors of $\hat{W}_x$ serve as a frame for the subspace, range$(\hat{W}_x^T)$, which contains x. Then we have:

$$\begin{aligned}\|f_{\text{cnn}}(x)\|_2^2 &= \|g(Wx)\|_2^2 = \|\hat{W}_x(x_p+x_p^c)\|_2^2 \\ &= \|\hat{W}_x x_p\|_2^2 \leq C_2\|x_p\|_2^2 \leq C_2\|x\|_2^2\end{aligned} \tag{S3}$$

where C_2 is the largest eigenvalue of $\hat{\tilde{\mathcal{S}}}_x$ by Theorem A.1. Again, by Theorem 2.1, $C_2 \leq 1+(\hat{K}_a-1)\mu$.

Together we obtain the final inequality.