
A Spline Theory of Deep Networks (Extended Version)

Randall Balestriero

Richard Baraniuk¹

Abstract

We build a rigorous bridge between deep networks (DNs) and approximation theory via spline functions and operators. Our key result is that a large class of DNs can be written as a composition of max-affine spline operators (MASOs), which provide a powerful portal through which to view and analyze their inner workings. For instance, conditioned on the input signal, the output of a MASO DN can be written as a simple affine transformation of the input. This implies that a DN constructs a set of signal-dependent, class-specific templates against which the signal is compared via a simple inner product; we explore the links to the classical theory of optimal classification via matched filters and the effects of data memorization. Going further, we propose a simple penalty term that can be added to the cost function of any DN learning algorithm to force the templates to be orthogonal with each other; this leads to significantly improved classification performance and reduced overfitting with no change to the DN architecture. The spline partition of the input signal space that is implicitly induced by a MASO directly links DNs to the theory of vector quantization (VQ) and K -means clustering, which opens up new geometric avenue to study how DNs organize signals in a hierarchical fashion. To validate the utility of the VQ interpretation, we develop and validate a new distance metric for signals and images that quantifies the difference between their VQ encodings. (This paper is a significantly expanded version of a paper with the same title that will appear at ICML 2018.)

1. Introduction

Deep learning has significantly advanced our ability to address a wide range of difficult machine learning and signal processing problems. Today’s machine learning landscape

is dominated by *deep (neural) networks* (DNs), which are compositions of a large number of simple parameterized linear and nonlinear operators. An all-too-familiar story of late is that of plugging a deep network into an application as a black box, learning its parameter values using copious training data, and then significantly improving performance over classical task-specific approaches.

Despite this empirical progress, the precise mechanisms by which deep learning works so well remain relatively poorly understood, adding an air of mystery to the entire field. Ongoing attempts to build a rigorous mathematical framework fall roughly into five camps: (i) probing and measuring networks to visualize their inner workings (Zeiler & Fergus, 2014); (ii) analyzing their properties such as expressive power (Cohen et al., 2016), loss surface geometry (Lu & Kawaguchi, 2017; Soudry & Hoffer, 2017), nuisance management (Soatto & Chiuso, 2016), sparsification (Papyan et al., 2017), and generalization abilities; (iii) new mathematical frameworks that share some (but not all) common features with DNs (Bruna & Mallat, 2013); (iv) probabilistic generative models from which specific DNs can be derived (Arora et al., 2013; Patel et al., 2016); and (v) information theoretic bounds (Tishby & Zaslavsky, 2015).

In this paper, we build a rigorous bridge between DNs and *approximation theory via spline functions and operators*. We prove that a large class of DNs — including convolutional neural networks (CNNs) (LeCun, 1998), residual networks (ResNets) (He et al., 2015), skip connection networks (Srivastava et al., 2015), fully connected networks (Pal & Mitra, 1992), recurrent neural networks (RNNs) (Graves, 2013), scattering networks (Bruna & Mallat, 2013), inception networks (Szegedy et al., 2017), and more — can be written as spline operators.

Moreover, when these DNs employ current standard-practice piecewise affine and convex nonlinearities (e.g., ReLU, absolute value, max-pooling, etc.) they can be written as the composition of *max-affine spline operators* (MASOs), which are a new extension of max-affine splines (Magnani & Boyd, 2009; Hannah & Dunson, 2013). An extremely useful feature of such a spline is that it circumvents the major complication of spline function approximation — the need to jointly optimize not only the spline function parameters but also the partition of the domain over which

¹Rice University, Houston, Texas, USA (dsp.rice.edu). Correspondence to: R. Balestriero <randallbalestriero@gmail.com>.

those parameters are constant (the “knots” of the spline). Instead, the partition of a max-affine spline is determined implicitly in terms of its slope and offset parameters. We focus on such nonlinearities here but note that our framework applies also to non-piecewise affine nonlinearities through a standard approximation argument.

The max-affine spline connection provides a powerful portal through which to view and analyze the inner workings of a DN using tools from approximation theory and functional analysis. While there has been prior work relating DNs to splines (Montufar et al., 2014; Rister & Rubin, 2017; Unser, 2018), to date theoretical studies have been focused on either specific topologies or providing theoretical guarantees and bounds on specific properties, such as a DN’s approximation capacity or stability. In contrast, we develop a theory that is agnostic to the DN architecture and supports not just analysis but also extensions.

Here is a summary of our key contributions:

[C1] From our proof that a large class of DNs can be written as a composition of MASOs, it follows immediately that, *conditioned on the input signal, the output of a DN is a simple affine transformation of the input.* We illustrate in Section 4 by deriving closed form expressions for the input/output mapping of CNNs and ResNets.

[C2] We prove that a composition of two or more MASOs is capable of *approximating an arbitrary operator* in Section 4.6.

[C3] The affine mapping formula enables us to interpret a MASO DN as constructing a set of *signal-dependent, class-specific templates* against which the signal is compared via a simple inner product. In Section 5 we relate DNs directly to the classical theory of optimal classification via matched filters and provide insights into the effects of *data memorization* (Zhang et al., 2016).

[C4] We propose a simple penalty term that can be added to the cost function of any DN learning algorithm to force the templates to be *orthogonal* to each other. In Section 5.7, we show that this leads to significantly improved classification performance and reduced overfitting on standard test data sets like SVHN, CIFAR10, and CIFAR100 with no change to the DN architecture.

[C5] The *partition of the input space* induced by a MASO links DNs to the theory of *vector quantization* (VQ) and *K-means clustering*, which opens up a new geometric avenue to study how DNs cluster and organize signals in a hierarchical fashion. Section 6 studies the properties of the MASO partition.

[C6] Leveraging the fact that a DN considers two signals to be similar if they lie in the same MASO partition region, we develop a new *VQ-based distance* for signals and im-

ages in Section 6.6 that measures the difference between their VQ encodings. The distance is easily computed via backpropagation on the DN.

This paper is a significantly expanded version of the ICML 2018 conference paper with the same title (Balestrieri & Baraniuk, 2018). We have structured it as an executive summary containing the key results, theorem statements, and experimental results. A number of appendices in the Supplementary Material (SM) contain the rigorous mathematical set up, proofs, additional insights, and additional examples.

2. Background on Deep Networks

A *deep network* (DN) is an operator $f_\Theta : \mathbb{R}^D \rightarrow \mathbb{R}^C$ that maps an input signal $\mathbf{x} \in \mathbb{R}^D$ to an output prediction $\mathbf{y} \in \mathbb{R}^C$. All current DNs can be written as a composition of L intermediate mappings called *layers*

$$f_\Theta(\mathbf{x}) = \left(f_{\theta^{(L)}}^{(L)} \circ \cdots \circ f_{\theta^{(1)}}^{(1)} \right)(\mathbf{x}), \quad (1)$$

where $\Theta = \{\theta^{(1)}, \dots, \theta^{(L)}\}$ is the collection of the network’s parameters from each layer. This composition of mappings is nonlinear and non-commutative, in general.

The DN *layer* at level ℓ is an operator $f_{\theta^{(\ell)}}$ that takes as input the vector-valued signal $\mathbf{z}^{(\ell-1)}(\mathbf{x}) \in \mathbb{R}^{D^{(\ell-1)}}$ and produces the vector-valued output $\mathbf{z}^{(\ell)}(\mathbf{x}) \in \mathbb{R}^{D^{(\ell)}}$. The signals $\mathbf{z}^{(\ell)}(\mathbf{x}), \ell \geq 1$ are typically called *feature maps*; it is easy to see that

$$\mathbf{z}^{(\ell)}(\mathbf{x}) = \left(f_{\theta^{(\ell)}}^{(\ell)} \circ \cdots \circ f_{\theta^{(1)}}^{(1)} \right)(\mathbf{x}), \quad \ell \in \{1, \dots, L\} \quad (2)$$

with the initialization $\mathbf{z}^{(0)}(\mathbf{x}) := \mathbf{x}$ and $D^{(0)} := D$. Note that $D^{(L)} = C$.

For concreteness, we will focus here on processing multi-channel images x , such as RGB color digital photographs. However, our analysis and techniques apply to signals of any index-dimensionality, including speech and audio signals, video signals, etc., simply by adjusting the appropriate dimensionalities. We will use two equivalent representations for the signal and feature maps, one based on tensors and one based on flattened vectors. In the *tensor* representation, the input image x contains $C^{(0)}$ channels of size $(I^{(0)} \times J^{(0)})$ pixels, and the feature map $z^{(\ell)}$ contains $C^{(\ell)}$ channels of size $(I^{(\ell)} \times J^{(\ell)})$ pixels. In the *vector* representation, $[x]_k$ represents the entry of the k^{th} dimension of the flattened, vector version \mathbf{x} of x . Hence, $D^{(\ell)} = C^{(\ell)} I^{(\ell)} J^{(\ell)}$, $C^{(L)} = C$, $I^{(L)} = 1$, and $J^{(L)} = 1$. We propose in Appendix A detailed notation.

2.1. DN Operators and Layers

We briefly overview the basic DN operators and layers we consider in this paper. A *fully connected operator* performs

an arbitrary affine transformation by multiplying its input by the dense matrix $W^{(\ell)} \in \mathbb{R}^{D^{(\ell)} \times D^{(\ell-1)}}$ and adding the arbitrary bias vector $b_W^{(\ell)} \in \mathbb{R}^{D^{(\ell)}}$, as in $f_W^{(\ell)}(z^{(\ell-1)}(\mathbf{x})) := W^{(\ell)}z^{(\ell-1)}(\mathbf{x}) + b_W^{(\ell)}$.

A **convolution operator** performs a constrained affine transformation by multiplying its input by the multichannel block-circulant convolution matrix $C^{(\ell)} \in \mathbb{R}^{D^{(\ell)} \times D^{(\ell-1)}}$ and adding the bias vector $b_C^{(\ell)} \in \mathbb{R}^{D^{(\ell)}}$, as in $f_C^{(\ell)}(z^{(\ell-1)}(\mathbf{x})) := C^{(\ell)}z^{(\ell-1)}(\mathbf{x}) + b_C^{(\ell)}$. The multichannel convolution matrix applies $C^{(\ell)}$ different filters of dimensions $(C^{(\ell-1)}, I_C^{(\ell)}, J_C^{(\ell)})$ to the input; that is, each filter is a $C^{(\ell-1)}$ -channel filter of spatial size $(I_C^{(\ell)}, J_C^{(\ell)})$.

Typically, the bias $b_C^{(\ell)}$ is constrained to be a (potentially different) constant at all spatial positions of each of the $C^{(\ell)}$ output channels. The convolution and bias constraints radically reduce the number of parameters compared to a fully connected operator for similar input-output dimensions, but impose a structural bias of translation invariance (LeCun, 1998). Additional details are provided in Appendix B.1.

An **activation operator** applies a pointwise nonlinearity σ to its input, as in $[f_\sigma^{(\ell)}(z^{(\ell-1)}(\mathbf{x})}]_k = \sigma([z^{(\ell-1)}(\mathbf{x})]_k)$, $k = 1, \dots, D^{(\ell)}$. Nonlinearities are crucial to DNs, since otherwise the entire network would collapse to a single global affine transform. Generally, a different nonlinearity can be used for each dimension of $z^{(\ell-1)}(\mathbf{x})$, but standard practice employs the same σ across all entries. Three popular activation functions are the *rectified linear unit* (ReLU) $\sigma_{\text{ReLU}}(u) := \max(u, 0)$, the *leaky ReLU* $\sigma_{\text{LReLU}}(u) := \max(\eta u, u)$, $\eta > 0$, and the *absolute value* $\sigma_{\text{abs}}(u) := |u|$. These three functions are both piecewise affine and convex. Other popular activation functions include the *sigmoid* $\sigma_{\text{sig}}(u) := \frac{1}{1+e^{-u}}$ and *hyperbolic tangent* $\sigma_{\text{tanh}}(u) := 2\sigma_{\text{sig}}(2u) - 1$. These two functions are neither piecewise affine nor convex.

A **pooling operator** subsamples its input to reduce its dimensionality according to a policy ρ applied over $D^{(\ell)}$ pooling regions, each one defined by the collection of indices $\{\mathcal{R}_k^{(\ell)}\}_{k=1}^{D^{(\ell)}}$ on which ρ will be applied independently. Pooling regions typically correspond to small, non-overlapping image blocks or channel slices. A widely used policy ρ is *max-pooling*: $[f_\rho^{(\ell)}(z^{(\ell-1)}(\mathbf{x})}]_k = \max_{d \in \mathcal{R}_k^{(\ell)}} [z^{(\ell-1)}(\mathbf{x})]_d$, $k = 1, \dots, D^{(\ell)}$. See Appendix B.2 for the precise definition of max-pooling plus the definitions of *average pooling* and *channel pooling*, which is used in maxout networks (Goodfellow et al., 2013).

More details on the above operators plus additional ones corresponding to the *skip connections* used in skip connec-

tion networks (Srivastava et al., 2015) and residual networks (ResNets) (He et al., 2015; Targ et al., 2016) and the *recurrent connections* used in RNNs (Graves, 2013) are contained in Appendix B.

There is no consensus **definition in the literature of a DN layer**. For the purposes of this paper, we will use the following.

Definition 1. A DN layer $f^{(\ell)}$ comprises a single nonlinear DN operator composed with any preceding affine operators lying between it and the preceding nonlinear operator.

This definition yields a single, unique layer decomposition of any DN, and the complete DN is then the composition of its layers per (1). For example, in a standard CNN, there are two different layers types: i) convolution-activation and ii) max-pooling.

2.2. Output Operators

The final operators in a DN typically consist of one or more fully connected operators interspersed with activation operators. We form the prediction by feeding the DN output $f_\Theta(\mathbf{x})$ through a final nonlinearity $g : \mathbb{R}^C \rightarrow \mathbb{R}^C$ to form $y = g(f_\Theta(\mathbf{x}))$. For classification problems, g is typically the *softmax*, which arises naturally from posing the classification task as a multinomial logistic regression (Bishop, 1995), but other options are available, such as the *spherical softmax* (de Brébisson & Vincent, 2015). For regression problems, typically no final nonlinearity g is applied.

2.3. DN Learning

We learn the DN parameters Θ for a particular prediction task in a supervised setting using a labeled data set $\mathcal{D} = (\mathbf{x}_n, y_n)_{n=1}^N$, a loss function $\mathcal{L} : \mathbb{R}^C \times \mathbb{R}^C \rightarrow \mathbb{R}^+$, and a learning policy to update the parameters Θ in the predictor $f_\Theta(\mathbf{x})$. For a C -class classification problem posed as a multinomial logistic regression (Bishop, 1995), $y_n \in \{1, \dots, C\}$, and the loss function is typically a regularized version of the negative cross-entropy

$$\begin{aligned} \mathcal{L}_{\text{CE}}(y_n, g(f_\Theta(\mathbf{x}_n))) &= -\log ([g(f_\Theta(\mathbf{x}_n))]_{y_n}) \\ &= -[f_\Theta(\mathbf{x}_n)]_{y_n} + \log \left(\sum_{c=1}^C e^{[f_\Theta(\mathbf{x}_n)]_c} \right). \end{aligned} \quad (3)$$

For regression problems, y_n becomes the vector $\mathbf{y}_n \in \mathbb{R}^C$, and the loss function is typically the mean squared error. The total loss is obtained by summing over all of the N samples in \mathcal{D} or mini-batches sampled from it.

Since the layer-by-layer operations in a DN are differentiable almost everywhere with respect to their parameters and inputs, we can optimize the parameters Θ with respect to the total loss using some form of first-order optimization such as gradient descent (e.g., Adam (Kingma & Ba, 2014)).

Moreover, the gradients of all internal parameters can be computed efficiently using *backpropagation* (Hecht-Nielsen et al., 1988), which follows from the chain rule of calculus.

3. Background on Spline Operators

Approximation theory is the study of how and how well functions can best be approximated using simpler functions (Powell, 1981). A classical example of a simpler function is a *spline* $s : \mathbb{R}^D \rightarrow \mathbb{R}$ (De Boor et al., 1978). *For concreteness, we will focus exclusively on affine splines in this paper (a.k.a. “linear splines”), but our ideas generalize naturally to higher-order splines.*

3.1. Multivariate Affine Splines

Consider a *partition* of a domain \mathbb{R}^D into a set of regions $\Omega = \{\omega_1, \dots, \omega_R\}$ and a set of local mappings $\Phi = \{\phi_1, \dots, \phi_R\}$ that map each region in the partition to \mathbb{R} via $\phi_r(\mathbf{x}) := \langle [\alpha]_{r,.}, \mathbf{x} \rangle + [\beta]_r$ for $\mathbf{x} \in \omega_r$.¹ The parameters are: $\alpha \in \mathbb{R}^{R \times D}$, a matrix of hyperplane “slopes,” and $\beta \in \mathbb{R}^R$, a vector of hyperplane “offsets” or “biases”. We will use the terms offset and bias interchangeably in the sequel. The notation $[\alpha]_{r, \cdot}$ denotes the column vector formed from the r^{th} row of α .

Then the *multivariate affine spline* is defined as

$$\begin{aligned} s[\alpha, \beta, \Omega](\mathbf{x}) &= \sum_{r=1}^R (\langle [\alpha]_{r,.}, \mathbf{x} \rangle + [\beta]_r) \mathbb{1}(\mathbf{x} \in \omega_r) \\ &=: \langle \alpha[\mathbf{x}], \mathbf{x} \rangle + \beta[\mathbf{x}], \end{aligned} \quad (4)$$

where $\mathbb{1}(\mathbf{x} \in \omega_r)$ denotes the indicator function that equals 1 when $\mathbf{x} \in \omega_r$ and 0 otherwise. The second line of (4) introduces the streamlined notation $\alpha[\mathbf{x}] = [\alpha]_{r,.}$ when $\mathbf{x} \in \omega_r$; the definition for $\beta[\mathbf{x}]$ is similar. Such a spline is piecewise affine and hence piecewise convex. However, in general, it is neither globally affine nor globally convex unless $R = 1$, a case we denote as a *degenerate spline*, since it corresponds simply to an affine mapping.

3.2. Max-Affine Spline Functions

A major complication of function approximation with splines in general is the need to jointly optimize both the spline parameters α, β and the input domain partition Ω (the “knots” for a 1D spline) (Bennett & Botkin, 1985).

However, if a multivariate affine spline is constrained to be *globally convex*, then it can be rewritten as a *max-affine*

¹ To make the connection between splines and DNs more immediately obvious, here \mathbf{x} is interpreted as a point in \mathbb{R}^D , which plays the rôle of the space of signals in the other sections.

spline (Magnani & Boyd, 2009; Hannah & Dunson, 2013)

$$s[\alpha, \beta, \Omega](\mathbf{x}) = \max_{r=1, \dots, R} \langle [\alpha]_{r,.}, \mathbf{x} \rangle + [\beta]_r. \quad (5)$$

An extremely useful feature of such a spline is that *it is completely determined by its parameters α and β without needing to specify the partition Ω* . As such, we denote a max-affine spline simply as $s[\alpha, \beta]$. Changes in the parameters α, β of a max-affine spline automatically induce changes in the partition Ω , meaning that they are *adaptive partitioning splines* (Magnani & Boyd, 2009).

A max-affine spline is always a piecewise affine and globally convex (and hence also continuous) function. Conversely, any piecewise affine and globally convex function can be written as a max-affine spline.

Proposition 1. *For any (continuous) $h \in \mathcal{C}^0(\mathbb{R}^D)$ that is piecewise affine and globally convex, there exist α, β such that $h(\mathbf{x}) = s[\alpha, \beta](\mathbf{x}), \forall \mathbf{x}$.*

This result follows from the fact that the pointwise maximum of a collection of convex functions is convex (Roberts, 1993). Using standard approximation arguments, it is easy to show that a max-affine spline can approximate an arbitrary (nonlinear) convex function arbitrarily closely (Nishikawa, 1998).

3.3. Max-Affine Spline Operators

A natural extension of an affine spline function is an *affine spline operator* (ASO) $S[A, B, \Omega^S]$ that produces a multivariate output. It is obtained simply by concatenating K affine spline functions from (4). The details and a more general development are provided in Appendix C.2

We are particularly interested in the *max-affine spline operator* (MASO) $S[A, B] : \mathbb{R}^D \rightarrow \mathbb{R}^K$ formed by concatenating K independent max-affine spline functions from (5). A MASO with slope parameters $A \in \mathbb{R}^{K \times R \times D}$ and offset parameters $B \in \mathbb{R}^{K \times R}$ is defined as

$$\begin{aligned} S[A, B](\mathbf{x}) &= \begin{bmatrix} \max_{r=1, \dots, R} \langle [A]_{1,r,.}, \mathbf{x} \rangle + [B]_{1,r} \\ \vdots \\ \max_{r=1, \dots, R} \langle [A]_{K,r,.}, \mathbf{x} \rangle + [B]_{K,r} \end{bmatrix} \\ &=: A[\mathbf{x}] \mathbf{x} + B[\mathbf{x}]. \end{aligned} \quad (6)$$

The second line of (6) introduces the streamlined notation in terms of the signal-dependent matrix $A[\mathbf{x}]$ and signal-dependent vector $B[\mathbf{x}]$, where $[A[\mathbf{x}]]_{k,.} := [A]_{k, r_k(\mathbf{x}), .}$ and $[B[\mathbf{x}]]_k := [B]_{k, r_k(\mathbf{x})}$ with $r_k(\mathbf{x}) = \arg \max_r \langle [A]_{k,r,.}, \mathbf{x} \rangle + [B]_{k,r}$. Since a MASO is built from K independent max-affine spline functions, it has a property analogous to Proposition 1.

Proposition 2. *For any operator $H(\mathbf{x}) = [h_1(\mathbf{x}), \dots, h_K(\mathbf{x})]^T$ with $h_k \in \mathcal{C}^0(\mathbb{R}^D) \forall k$ that are*

each piecewise affine and globally convex, there exist A, B such that $H(\mathbf{x}) = S[A, B](\mathbf{x}) \forall \mathbf{x}$.

As above, using standard approximation arguments, it is easy to show that a MASO can approximate arbitrarily closely any (nonlinear) operator that is convex in each output dimension.

4. Deep Networks are Compositions of Spline Operators

The key enabling result of this paper is that a large class of DNs can be written as a composition of MASOs, one for each layer. This includes many standard DNs topologies, including CNNs, ResNets, inception networks, maxout networks, network-in-networks, scattering networks, RNNs, and their variants, provided they use piecewise-affine and convex operators.

4.1. DN Operators and Layers are MASOs

We begin by showing that the DN operators defined in Section 2.1 are MASOs. The proofs of these results are contained in Appendix E.

Proposition 3. An arbitrary fully connected operator $f_W^{(\ell)}$ is an affine mapping and hence a degenerate MASO $S[A_W^{(\ell)}, B_W^{(\ell)}]$, with $R = 1$, $[A_W^{(\ell)}]_{k,1,\cdot} = [W^{(\ell)}]_k$, and $[B_W^{(\ell)}]_{k,1} = [b_W^{(\ell)}]_k$, leading to

$$W^{(\ell)} \mathbf{z}^{(\ell-1)}(\mathbf{x}) + b_W^{(\ell)} = A_W^{(\ell)}[\mathbf{x}] \mathbf{z}^{(\ell-1)}(\mathbf{x}) + B_W^{(\ell)}[\mathbf{x}]. \quad (7)$$

The same is true of a convolution operator with $W^{(\ell)}, b_W^{(\ell)}$ replaced by $C^{(\ell)}, b_C^{(\ell)}$.

Proposition 4. Any activation operator $f_\sigma^{(\ell)}$ using a piecewise affine and convex activation function is a MASO $S[A_\sigma^{(\ell)}, B_\sigma^{(\ell)}]$ with $R = 2$, $[B_\sigma^{(\ell)}]_{k,1} = [B_\sigma^{(\ell)}]_{k,2} = 0 \forall k$, and for ReLU

$$[A_\sigma^{(\ell)}]_{k,1,\cdot} = 0 \quad [A_\sigma^{(\ell)}]_{k,2,\cdot} = \mathbf{e}_k \quad \forall k, \quad (8)$$

for Leaky-ReLU

$$[A_\sigma^{(\ell)}]_{k,1,\cdot} = \nu \mathbf{e}_k, \quad [A_\sigma^{(\ell)}]_{k,2,\cdot} = \mathbf{e}_k \quad \forall k, \nu > 0, \quad (9)$$

and for absolute value

$$[A_\sigma^{(\ell)}]_{k,1,\cdot} = -\mathbf{e}_k, \quad [A_\sigma^{(\ell)}]_{k,2,\cdot} = \mathbf{e}_k \quad \forall k, \quad (10)$$

where \mathbf{e}_k represents the k^{th} canonical basis element of $\mathbb{R}^{D^{(\ell)}}$.

To illustrate more specifically, here are the ReLU and absolute value activation operators written as MASOs:

$$S[A_{\text{ReLU}}, B_{\text{ReLU}}](\mathbf{x}) = \begin{bmatrix} \max(0, [\mathbf{x}]_1) \\ \vdots \\ \max(0, [\mathbf{x}]_K) \end{bmatrix}, \quad (11)$$

$$S[A_{\text{abs}}, B_{\text{abs}}](\mathbf{x}) = \begin{bmatrix} \max(-[\mathbf{x}]_1, [\mathbf{x}]_1) \\ \vdots \\ \max(-[\mathbf{x}]_K, [\mathbf{x}]_K) \end{bmatrix}. \quad (12)$$

Proposition 5. Any pooling operator $f_\rho^{(\ell)}$ that is piecewise-affine and convex is a MASO $S[A_\rho^{(\ell)}, B_\rho^{(\ell)}]$.² Max-pooling has $R = \#\mathcal{R}_k$ (typically a constant over all output dimensions k), $[A_\rho^{(\ell)}]_{k,\cdot,\cdot} = \{\mathbf{e}_i, i \in \mathcal{R}_k\}$, and $[B_\rho^{(\ell)}]_{k,r} = 0 \forall k, r$. Average-pooling is a degenerate MASO with $R = 1$, $[A_\rho^{(\ell)}]_{k,1,\cdot} = \frac{1}{\#\mathcal{R}_k} \sum_{i \in \mathcal{R}_k} \mathbf{e}_i$, and $[B_\rho^{(\ell)}]_{k,1} = 0 \forall k$.

Combinations of DN operators that result in a layer (recall Definition 1) are also MASOs.

Proposition 6. A DN layer constructed from an arbitrary composition of fully connected/convolution operators followed by one activation or pooling operator is a MASO $S[A^{(\ell)}, B^{(\ell)}]$ such that

$$f^{(\ell)}(\mathbf{z}^{(\ell-1)}) = A^{(\ell)}[\mathbf{x}] \mathbf{z}^{(\ell-1)}(\mathbf{x}) + B^{(\ell)}[\mathbf{x}]. \quad (13)$$

For example, a layer composed of an affine transform $S[A_W^{(\ell)}, B_W^{(\ell)}]$ followed by an activation operator $S[A_\sigma^{(\ell)}, B_\sigma^{(\ell)}]$ is a MASO $S[A^{(\ell)}, B^{(\ell)}]$ with parameters $[A^{(\ell)}]_{k,r,\cdot} = W^{(\ell)T} [A_\sigma^{(\ell)}]_{k,r,\cdot}$ and $[B^{(\ell)}]_{k,r} = [B_\sigma^{(\ell)}]_{k,r} + b_W^{(\ell)T} [A_\sigma^{(\ell)}]_{k,r,\cdot}$. When $D^{(\ell)} > D^{(\ell-1)}$, such a composition is an *expansion layer* that up-samples its input to increase its dimensionality. Such layers are common in autoencoders (Vincent et al., 2008) and in the generator of a generative adversarial network (GAN) (Goodfellow et al., 2014). Appendix E.2 contains several additional examples of rewriting a composition of multiple DN operators as a MASO layer mapping.

4.2. Compositions of MASOs

The development of the previous section enables us to state a general result for DNs.

Theorem 1. A DN constructed from an arbitrary composition of fully connected/convolution, activation, and pooling operators of the types covered in Propositions 3–5 or layers

² This result is agnostic of the type of pooling (spatial or channel). The details are provided in Figure 16 of Appendix B.2.

from Proposition 6 is a composition of MASOs. Moreover, the overall composition is itself an ASO.

DNs covered by Theorem 1 include CNNs, ResNets, inception networks, maxout networks, network-in-networks, scattering networks, and their variants using connected/convolution operators, (leaky) ReLU or absolute value activations, and max/mean pooling.

Note carefully that DNs of the form stated in Theorem 1 are not MASOs, in general, since the composition of two or more MASOs is not necessarily a MASO (it is, at least, an ASO). Indeed, a composition of MASOs remains a MASO if and only if all of the intermediate operators are non-decreasing with respect to each of their output dimensions (Boyd & Vandenberghe, 2004). Interestingly, ReLU, max-pooling, and average pooling are all non-decreasing, while leaky ReLU is strictly increasing. The culprits of the non-convexity of the composition of operators are negative entries in the fully connected and convolution operators. A DN where these culprits are thwarted is an interesting special case, because it is convex with respect to its input (Amos et al., 2016) and multiconvex (Xu & Yin, 2013) with respect to its parameters (i.e., convex with respect to each operator's parameters while the other operators' parameters are held constant).

Proposition 7. *A DN layer is nondecreasing with respect to each of its output dimensions if and only if its slope parameters are nonnegative*

$$\left[A^{(\ell)} \right]_{k,r,d} \geq 0. \quad (14)$$

The condition (14) holds if only if the layer's activation or pooling operators are nondecreasing and the entries of any fully connected or convolution matrices satisfy $[W^{(\ell)}]_{k,d}, [C^{(\ell)}]_{k,d} \geq 0$.³

Theorem 2. *An L -layer DN whose layers $\ell = 2, \dots, L$ are not only piecewise-affine and convex but also nondecreasing with respect to each of their output dimensions is globally a MASO and thus also globally convex with respect to each of its output dimensions.*

Note that Theorem 2 remains true even if the DN's first layer ($\ell = 1$) is *not* non-decreasing. For example, it could include fully connected/convolution matrices with negative values and/or the absolute value activation function.

4.3. DNs are Signal-Dependent Affine Transformations

The second conclusion in Theorem 1 is that the mapping from the input \mathbf{x} to the layer output $\mathbf{z}^{(\ell)}(\mathbf{x})$ is an ASO.

³ Alternatively, the activation or pooling operators can be non-increasing and the entries of any fully connected or convolution matrices can satisfy $[W^{(\ell)}]_{k,d}, [C^{(\ell)}]_{k,d} \leq 0$.

Hence, $\mathbf{z}^{(\ell)}(\mathbf{x})$ is a *signal-dependent, piecewise affine transformation* of \mathbf{x} . (Much more on this below in Section 5.) The particular affine mapping applied to \mathbf{x} depends on which partition of the spline it falls in \mathbb{R}^D . (Much more on this in Section 6 below.) As such, a more precise but unwieldy terminology that we will not emphasize would be that $\mathbf{z}^{(\ell)}(\mathbf{x})$ is a *partition-region-dependent, piecewise affine transformation* of \mathbf{x} (see Remark 1 below).

The above results pertain to DNs using convex, piecewise affine operators. DNs using operators that are convex but not piecewise affine can be approximated arbitrarily closely by a MASO. DNs using operators that are non-convex (e.g., the sigmoid and arctan activation functions) can be approximated by an ASO (that itself is a composition of MASOs). We investigate this approximation in more detail below in Section 4.6.

For any DN covered by Theorem 1, we can substitute (6) into (2) to obtain an explicit formula for the layer output $\mathbf{z}^{(\ell)}(\mathbf{x})$ in terms of the input \mathbf{x} . We now illustrate with two examples, namely CNNs and ResNets.

4.4. Example: CNN Affine Mapping Formula

The explicit input/output formula for a standard CNN (using convolution, ReLU activation, max-pooling, and a final fully connected operator) is given in (15) at the top of the next page.

In (15), $A_\sigma^{(\ell)}[\mathbf{x}]$ are the signal-dependent matrices corresponding to the ReLU activation functions, $A_\rho^{(\ell)}[\mathbf{x}]$ are the signal-dependent matrices corresponding to the max-pooling operators, and the biases $b_W^{(L)}, b_C^{(L)}$ arise directly from the fully connected and convolution operators. The absence of $B_\sigma^{(\ell)}[\mathbf{x}]$ and $B_\rho^{(\ell)}[\mathbf{x}]$ is due to the absence of bias in the ReLU (recall (4)) and max-pooling operators (recall (5)).

Remark 1. *The affine transformation parameters $A[\mathbf{x}], B[\mathbf{x}]$ just above and throughout the development below are dependent not only on the signal \mathbf{x} but also more generally on the spline partition region containing \mathbf{x} (see Remark 2 below in Section 6.5). To streamline our notation, however, we will note the dependence simply by $[\mathbf{x}]$ and refer to it as “signal-dependent” rather than than the more precise “partition-region-of- \mathbf{x} -dependent.”*

Inspection of (15) reveals the exact form of the signal-dependent, piecewise affine mapping linking \mathbf{x} to $\mathbf{z}_{\text{CNN}}^{(L)}(\mathbf{x})$. Moreover, this formula can be collapsed into

$$\mathbf{z}_{\text{CNN}}^{(L)}(\mathbf{x}) = W^{(L)} (A_{\text{CNN}}[\mathbf{x}] \mathbf{x} + B_{\text{CNN}}[\mathbf{x}]) + b_W^{(L)} \quad (16)$$

from which we can recognize

$$\mathbf{z}_{\text{CNN}}^{(L-1)}(\mathbf{x}) = A_{\text{CNN}}[\mathbf{x}] \mathbf{x} + B_{\text{CNN}}[\mathbf{x}] \quad (17)$$

$$z_{\text{CNN}}^{(L)}(\mathbf{x}) = W^{(L)} \underbrace{\left(\prod_{\ell=L-1}^1 A_\rho^{(\ell)}[\mathbf{x}] A_\sigma^{(\ell)}[\mathbf{x}] \mathbf{C}^{(\ell)} \right) \mathbf{x}}_{A_{\text{CNN}}[\mathbf{x}]} + W^{(L)} \underbrace{\sum_{\ell=1}^{L-1} \left(\prod_{j=L-1}^{\ell+1} A_\rho^{(j)}[\mathbf{x}] A_\sigma^{(j)}[\mathbf{x}] \mathbf{C}^{(j)} \right) \left(A_\rho^{(\ell)}[\mathbf{x}] A_\sigma^{(\ell)}[\mathbf{x}] b_C^{(\ell)} \right) + b_W^{(L)}}_{B_{\text{CNN}}[\mathbf{x}]} \quad (15)$$

as an explicit, signal-dependent, affine formula for the *featurization process* that aims to convert \mathbf{x} into a set of (hopefully) linearly separable features that are then input to the linear classifier in layer $\ell = L$ with parameters $W^{(L)}$ and $b_W^{(L)}$. More generally, we will use the notation

$$z_{\text{CNN}}^{(\ell)}(\mathbf{x}) = A_{\text{CNN}}^{(\ell)}[\mathbf{x}] \mathbf{x} + B_{\text{CNN}}^{(\ell)}[\mathbf{x}] \quad (18)$$

but will omit the superscript when $\ell = L$ (as in (17)). When a DN is terminated using more than one fully connected layer, then all but the last layer can be incorporated into $A_{\text{CNN}}[\mathbf{x}]$ and $B_{\text{CNN}}[\mathbf{x}]$ as the class-agnostic representation of \mathbf{x} ; only the last fully connected layer is required in the linear classifier.

Of course, the final prediction y is formed by running $z_{\text{CNN}}^{(L)}(\mathbf{x})$ through a softmax nonlinearity g , to create a probability distribution. The same argument applies in the regression context if we suppress the softmax nonlinearity.

Some interesting properties of CNNs can be inferred from (15) thanks to its explicit nature. For example, what amounts to the matrix

$$A_{\text{CNN}}[\mathbf{x}] = \prod_{\ell=L-1}^1 A_\rho^{(\ell)}[\mathbf{x}] A_\sigma^{(\ell)}[\mathbf{x}] \mathbf{C}^{(\ell)} \quad (19)$$

has been probed and empirically shown to converge to the zero matrix as L grows. This observation has motivated skip-connections and ResNets (Targ et al., 2016), to which we now turn.

4.5. Example: ResNet Affine Mapping Formula

A ResNet layer leverages a skip-connection that “short-circuits” a layer from its input to its output (Targ et al., 2016). We study here the simplest form of skip-connection that comprises a convolution operator followed by an activation operator in parallel with a second convolution operator (the so-called linear skip-connection) as in

$$\begin{aligned} z^{(\ell)}(\mathbf{x}) &= C_{\text{skip}}^{(\ell)} z^{(\ell-1)}(\mathbf{x}) \\ &+ A_\sigma^{(\ell)}[\mathbf{x}] \left(C^{(\ell)} z^{(\ell-1)}(\mathbf{x}) + b_C^{(\ell)} \right) + b_{\text{skip}}^{(\ell)}. \end{aligned} \quad (20)$$

The notation “skip” distinguishes the parameters for the skip-connection. Such a topology results in state-of-the-art performance in standard image classification tasks (Targ et al., 2016). Skip-connections that skip over more than one layer can be formulated simply by adding additional terms

of the form $C_{\text{skip}}^{(\ell-2)} z^{(\ell-2)}(\mathbf{x}) + C_{\text{skip}}^{(\ell-3)} z^{(\ell-3)}(\mathbf{x}) + \dots$, to (20). Pooling in a standard ResNet is accomplished via dyadic subsampling over pixel space and thus corresponds to a simple subsampling of the rows of $C_{\text{skip}}^{(\ell)}$, $A^{(\ell)}[\mathbf{x}]$ and $b_{\text{skip}}^{(\ell)}$ in (20); the addition of more complex layers is straightforward.

The input/output formula for a standard ResNet (using convolution, ReLU activation, max-pooling, and a final fully connected operator) is given in (21). Comparison of (15) and (21) reveals that CNNs and ResNets share the same overall functional form (in fact, any composition of MASOs will have this form) of a featurizer followed by a linear classifier (recall (16) and (17)). However, in a ResNet, there exists a linear path from the input to the output of any layer. We can distribute this linear path based on (21) to highlight that a ResNet can be interpreted as an ensemble of CNNs of all depths $\leq L$ (Veit et al., 2016) with tied weights (same weights shared across models)

$$\begin{aligned} z_{\text{RES}}^{(L)}(\mathbf{x}) &= b_W^{(L)} + W^{(L)} \left(b_{\text{RES}}[\mathbf{x}] \right. \\ &+ \prod_{\ell=L-1}^1 C_{\text{skip}}^{(\ell)} \mathbf{x} \\ &+ \prod_{\ell=L-1}^2 C_{\text{skip}}^{(\ell)} A_\sigma^{(1)}[\mathbf{x}] \mathbf{C}^{(1)} \mathbf{x} \\ &+ \prod_{\ell=L-1}^3 C_{\text{skip}}^{(\ell)} A_\sigma^{(2)}[\mathbf{x}] \mathbf{C}^{(2)} A_\sigma^{(1)}[\mathbf{x}] \mathbf{C}^{(1)} \mathbf{x} \\ &+ \dots \\ &\left. + \prod_{\ell=L-1}^1 A_\sigma^{(\ell)}[\mathbf{x}] W^{(\ell)} \mathbf{x} \right). \end{aligned} \quad (22)$$

Skip-connections and the resulting ensemble-of-CNN’s structure prevent some of the detrimental stability behavior of very deep CNNs (large L) that was elucidated by (Targ et al., 2016). Indeed, the skip connections modify the matrix (19) that converges to zero with L to $\prod_{\ell=L-1}^1 (A_\sigma^{(\ell)}[\mathbf{x}] \mathbf{C}^{(\ell)} + C_{\text{skip}}^{(\ell)})$, which is stabilized by the presence of $C_{\text{skip}}^{(\ell)}$.

Formulas analogous to (15)–(21) can be derived for a large class of DNs; we include a formula for RNNs in Appendix D.1.

4.6. Universality of MASO DNs

The ability of certain DNs to approximate an arbitrary functional/operator mapping has been well established (Cybenko, 1989) using squashing activation functions ($\sigma : \mathbb{R} \rightarrow [0, 1]$). While it is clear that a convex operator can be ap-

$$\mathbf{z}_{\text{RES}}^{(L)}(\mathbf{x}) = W^{(L)} \underbrace{\left(\prod_{\ell=L-1}^1 \left(A_{\sigma}^{(\ell)}[\mathbf{x}] \mathbf{C}^{(\ell)} + \mathbf{C}_{\text{skip}}^{(\ell)} \right) \right)}_{A_{\text{RES}}[\mathbf{x}]} \mathbf{x} + W^{(L)} \underbrace{\sum_{\ell=L-1}^1 \left(\prod_{i=L-1}^{\ell+1} \left(A_{\sigma}^{(\ell)}[\mathbf{x}] \mathbf{C}^{(\ell)} + \mathbf{C}_{\text{skip}}^{(\ell)} \right) \right) \left(A_{\sigma}^{(\ell)}[\mathbf{x}] b_C^{(\ell)} + b_{\text{skip}}^{(\ell)} \right) + b_W^{(L)}}_{b_{\text{RES}}[\mathbf{x}]} \quad (21)$$

proximated arbitrarily closely by a single MASO (Magnani & Boyd, 2009), it is not clear a priori whether a composition of multiple MASOs is capable of approximating an arbitrary operator. Leveraging a result from (Breiman, 1993), we prove in Appendix E.7 that the composition of just two MASOs has a universal approximation property.

Theorem 3. Let $f(\mathbf{x})$ represent an arbitrary continuous operator $f : \mathbb{R}^D \rightarrow \mathbb{R}^C$, and let $f_{\Theta}(\mathbf{x})$ represent the output of an arbitrary DN with parameters Θ constructed from a composition of $L = 2$ MASOs with $R^{(1)} = 2$ and $R^{(2)} = 1$ (hence linear second layer). Then there exist (optimal) parameters Θ such that

$$\|f(\mathbf{x}) - f_{\Theta}(\mathbf{x})\|_2^2 \leq \frac{D^{(2)} \sum_{k=1}^C \eta_k}{D^{(1)}}, \quad (23)$$

where η_k is a constant characterizing the regularity of the k^{th} output of f .

The capability of a shallow DN ($L = 2$) with many partition regions (large $R^{(1)}$) to approximate a complicated operator has been studied numerically in (Zhang et al., 2016), where both two-layer perceptrons and CNNs were trained to memorize a training set with randomly permuted labels. Understanding the approximation capabilities of DNs for $L > 2$ remains open and should be studied not from an approximation theory perspective (since such a DN can memorize an arbitrary finite dataset) but rather from a generalization perspective.

Corollary 1. A DN containing non-convex and/or non-piecewise affine operators can be approximated arbitrarily closely by a MASO-based DN.

Corollary 1 expands the scope of our theory to include, for example, the sigmoid-based activations that are common in RNNs such as LSTMs and GRUs (Graves, 2013; Cho et al., 2014; Chung et al., 2014). In Appendix D.1 we work out the details of a piecewise affine spline approximation to a sigmoid based RNN.

In addition of being universal approximators, MASO-based DNs also appear naturally as solutions of an infinite-dimensional optimization for the optimal DN nonlinearities (Unser, 2018) (see Appendix E.8).

5. DNs are Template Matching Machines

We now take a deeper look into the featurization/prediction process of (16) in order to bridge DNs and classical optimal classification theory via matched filters (aka “template

matching”). We will focus on classification for concreteness. Our analysis holds for any DN meeting the conditions of Theorem 1, for example, the CNN in (15) or the ResNet in (21).

5.1. Matched Filter Template Matching

Rewriting (16) (and removing the CNN demarcation, because the result holds in more generality) as

$$\mathbf{z}^{(L)}(\mathbf{x}) = \left(W^{(L)} A[\mathbf{x}] \right) \mathbf{x} + \left(W^{(L)} B[\mathbf{x}] + b_W^{(L)} \right) \quad (24)$$

provides the alternate interpretation that $\mathbf{z}^{(L)}(\mathbf{x})$ is the output of a bank of *linear matched filters* (plus a set of biases). That is, the c^{th} element of $\mathbf{z}^{(L)}(\mathbf{x})$ equals the inner product between the signal \mathbf{x} and the matched filter for the c^{th} class, which is contained in the c^{th} row of the matrix $W^{(L)} A[\mathbf{x}]$. The bias $W^{(L)} B[\mathbf{x}] + b_W^{(L)}$ can be used to account for the fact that some classes might be more likely than others (i.e., the prior probability over the classes). It is well-known that a matched filterbank is the optimal classifier for deterministic signals in additive white Gaussian noise (Rabiner & Gold, 1975). Given an input \mathbf{x} , the class decision is simply the index of the largest element of $\mathbf{z}^{(L)}(\mathbf{x})$.⁴

5.2. Hierarchical Template Matching

Yet another interpretation of (16) is that $\mathbf{z}^{(L)}(\mathbf{x})$ is computed not in a single matched filter calculation but hierarchically as the signal propagates through the DN layers. Abstracting (6) to write the per-layer maximization process as $\mathbf{z}^{(\ell)}(\mathbf{x}) = \max_{r^{(\ell)}} A_{r^{(\ell)}}^{(\ell)} \mathbf{z}^{(\ell-1)}(\mathbf{x}) + B_{r^{(\ell)}}^{(\ell)}$ and cascading, obtain a formula for the end-to-end DN mapping

$$\mathbf{z}^{(L)}(\mathbf{x}) = W^{(L)} \max_{r^{(L-1)}} \left(A_{r^{(L-1)}}^{(L-1)} \max_{r^{(2)}} \left(A_{r^{(2)}}^{(2)} \dots \right. \right. \\ \left. \left. \max_{r^{(1)}} \left(A_{r^{(1)}}^{(1)} \mathbf{x} + B_{r^{(1)}}^{(1)} \right) + B_{r^{(2)}}^{(2)} \right) \dots + B_{r^{(L-1)}}^{(L-1)} \right) + b_W^{(L)}. \quad (25)$$

This formula elucidates that a DN performs a *hierarchical, greedy template matching* on its input, a computationally efficient yet sub-optimal template matching technique. Such a procedure is globally optimal when the DN is globally convex.

⁴Again, since the softmax merely rescales the entries of $\mathbf{z}^{(L)}(\mathbf{x})$ into a probability distribution, it does not affect the location of its largest element.

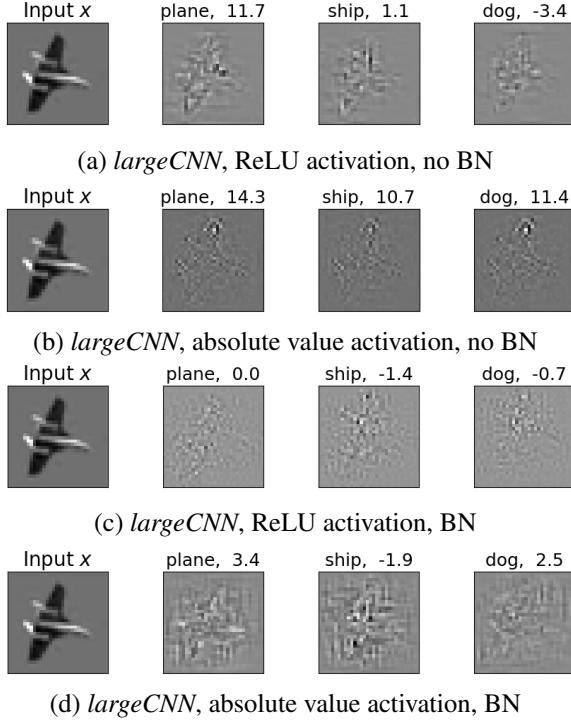


Figure 1. Visualization of the signal-dependent, matched filter templates generated by the *largeCNN* architecture employing max-pooling, either ReLU or absolute value activation function, and batch normalization (BN) or not (see Section 5.4 for more details) in a classification task with the CIFAR10 dataset. For ease of visualization only, the images were converted to gray scale by averaging their RGB components. To the right of the input image \mathbf{x} of class ‘plane’ we depict the three induced templates for the classes ‘plane,’ ‘ship,’ and ‘dog.’ The color map for the templates maps black to large negative values, grey to zero, and white to large positive values. Above each of the templates, we indicate the inner product between \mathbf{x} and the corresponding template. (The final classification will involve not only these inner products but also the relevant biases and softmax transformation.)

Corollary 2. For a DN abiding by the requirements of Theorem 2, the computation (25) collapses to the following globally optimal template matching

$$\mathbf{z}^{(L-1)}(\mathbf{x}) = W^{(L)} \max_{r^{(L-1)}, r^{(2)}, \dots, r^{(1)}} \left(A_{r^{(L-1)}}^{(L-1)} \left(A_{r^{(2)}}^{(2)} \dots \left(A_{r^{(1)}}^{(1)} \mathbf{x} + B_{r^{(1)}}^{(1)} \right) + B_{r^{(2)}}^{(2)} \right) \dots + B_{r^{(L-1)}}^{(L-1)} \right) + b_W^{(L)}. \quad (26)$$

5.3. Template Visualization Examples

Since the complete DN mapping (up to the final softmax) can be expressed as in (16),⁵ given a signal \mathbf{x} , the signal-

⁵Recall that this formula is not restricted to CNNs.

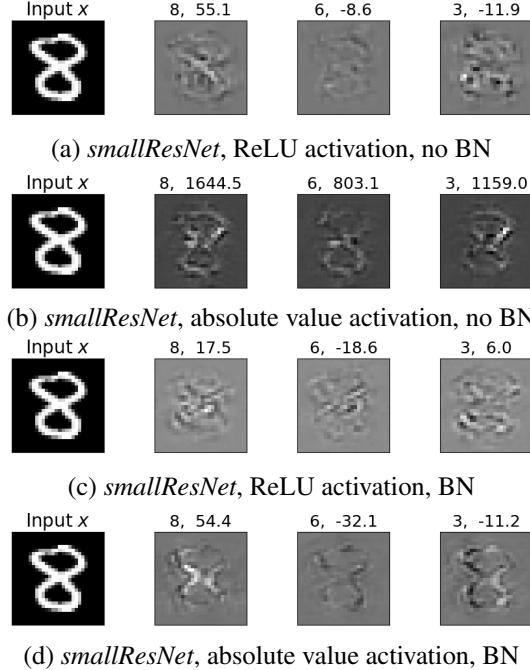


Figure 2. Visualization of the signal-dependent, matched filter templates generated by the *smallResNet* architecture employing max-pooling, either ReLU or absolute value activation function, and batch normalization (BN) or not in a classification task with the MNIST dataset. To the right of the input image \mathbf{x} of class ‘8’ we depict the three induced templates for the classes ‘8,’ ‘6,’ and ‘3.’

dependent template for class c is given by

$$A[\mathbf{x}]_c = \frac{[\mathbf{z}^{(L)}(\mathbf{x})]_c}{d\mathbf{x}}, \quad (27)$$

which can be efficiently computed via backpropagation (Hecht-Nielsen et al., 1988).⁶ Once the template $A[\mathbf{x}]_c$ has been computed, the bias term $b[\mathbf{x}]_c$ can be computed as $b[\mathbf{x}]_c = \mathbf{z}^{(L)}(\mathbf{x})_c - \langle A[\mathbf{x}]_c, \mathbf{x} \rangle$. This method of computation is typically much more efficient than using the explicit formulas (15), (21).

Figures 1–2 plot the signal-dependent, matched filter templates for two experiments with the CIFAR10 and MNIST datasets. We show results for two different DNs – *largeCNN* and *smallResNet* (see Appendix F for the definitions) – using two different activation functions – ReLU and absolute value. We trained for 75 epochs on CIFAR10 and 30 epochs on MNIST using a batch-size of 50 and the Adam optimizer with a constant learning rate of 0.0005.

⁶In fact, we can use the same backpropagation procedure used for computing the gradient with respect to a fully connected or convolution weight but instead with the input \mathbf{x} . This procedure is becoming increasingly popular in the study of adversarial examples (Szegedy et al., 2013).

The visual resemblance of the templates to an ‘automobile’/‘8’ or an ‘anti-automobile’/‘anti-8’ is intriguing. The inner products between the input image and the templates also conforms to our intuition regarding the behavior of a matched filter bank. That is, while the templates might seem rather incoherent on visual inspection, they are in fact tuned to the input in order to produce a high inner product for the correct class. Additional visualizations are provided in Appendix ??.

What we call “templates” have been probed in the deep learning community and called *saliency maps* (Simonyan et al., 2013; Zeiler & Fergus, 2014). In a result of independent interest, the above development proves that the saliency map technique is exact for MASO DNs. Indeed, a first-order Taylor expansion provides an exact representation of a piecewise affine mapping but only an approximation of a non-piecewise mapping like a sigmoid.

Proposition 8. *The saliency map technique based on (27) results in exact saliency maps for a DN with MASO layers and otherwise results in a first-order approximation.*

5.4. Template Matching vs. Bias

In emphasizing the visual appearance of the signal-dependent templates $A[\mathbf{x}]_c$, in Figures 1–2, we have neglected two additional important factors that can play major roles in determining the outcome of the classification process.

The first factor is the *bias* $W^{(L)}B[\mathbf{x}] + b_W^{(L)}$ from (17). Indeed, a large positive or negative bias can swamp the inner product between the input image and the template and throw the election for the candidate class. Note that this bias includes not only the final linear classifier bias (which can be interpreted as a prior over the classes) but also all of the per-layer biases (mapped to the output).

The second factor is whether *batch normalization* (Ioffe & Szegedy, 2015) is applied during learning and inference. Batch normalization (BN) is a specific kind of “data centering” process that is often applied between the fully connected/convolution and activation operators. It replaces $\mathbf{z}^{(\ell)}(\mathbf{x})$ with

$$\frac{\mathbf{z}^{(\ell)}(\mathbf{x}) - \mu}{\sqrt{\nu^2 + \epsilon}} \gamma + \zeta \quad (28)$$

where μ and ν are computed by averaging over a batch of images/feature maps and γ and ζ are learned (ϵ is a small constant to prevent instability). During inference, the parameters μ and ν are replaced by the averages of the values computed during training. This centering introduces a bias into output of the DN that is distinct from that in (17).

To explore the performance impacts of these two factors, we expanded on the experiments that generated Figures 1–2 by

DN Configuration	MNIST			
	Train		Test	
(bias)	(bias)	(bias)	(bias)	(bias)
bias, BN	99.9	99.9	99.4	99.4
bias, BN	100	100	99.4	99.4
bias, BN	68.1	68.1	67.9	67.9
bias, BN	100	72	99.5	81.9

Table 1. Classification results for the *largeCNN* trained and tested on the MNIST dataset with/without bias and with/without batch normalization (BN). (In the two bias rows, some entries are repeated, because performance is unchanged by including bias in the classifier.)

DN Configuration	CIFAR10			
	Train		Test	
(bias)	(bias)	(bias)	(bias)	(bias)
bias, BN	98.8	98.8	76.0	76.0
bias, BN	99.4	81.4	78.5	66.5
bias, BN	28.0	28.0	26.9	26.9
bias, BN	98.2	38.2	80.6	34.6

Table 2. Classification results for the *largeCNN* trained on the CIFAR10 dataset. (See Table 1 for more details.)

testing and training the *largeCNN* with/without the bias term $W^{(L)}B[\mathbf{x}] + b_W^{(L)}$ and with/without batch normalization. (We trained without the bias term simply by initializing all of the biases in the network to zero as usual and not training them.) For each of the four training scenarios, we classified both the training and test datasets using two predictors: (17) for the “bias” scenario and $y = W^{(L)}A[\mathbf{x}]$ for the “-bias” scenario. The results for each of the four experiments are displayed in Tables 1 and 2 for the MNIST and CIFAR10 datasets, respectively.

Tables 1 and 2 provide several useful insights. Row 1: Training with neither bias nor batch normalization and predicting based only on the template matching results in a high-performance and competitive classifier. This both confirms our matched filterbank interpretation and indicates that the templates are very informative for the task. Row 2: Training with bias but not batch normalization (standard practice as of a few years ago) is stable (from training to testing) for MNIST but not CIFAR10, most likely thanks to the lack of nuisance variation in MNIST (same color for all digits, no background, occlusions, nor out-of-plane rotations) as compared to CIFAR10. Rows 3 and 4: While batch normalization has been shown to accelerate training and improve accuracy, comparison of Rows 3 and 4 reveals that it is crucial to include the bias term.

5.5. Stability and Lipschitz Constant

The MASO formulation of DNs enables a straightforward analysis of its *Lipschitz constant*. Recall that the Lipschitz constant κ of a continuous and differentiable mapping

equals the infinity norm of its total derivative. As such, it quantifies its regularity, or how fast the mapping can change.

There are several immediate applications of such a regularity analysis to DNs. The Lipschitz constant plays a key role in analyzing the generalization capacity of a DN based on flat minima (Hochreiter & Schmidhuber, 1997). The Lipschitz constant also quantifies the volatile behavior of DNs that makes them susceptible to adversarial attacks (Kurakin et al., 2016). Adversarial attacks leverage the non-contractivity of a DN (large Lipschitz constant) to highly transform an output prediction given only a slightly perturbed input.

Proposition 9. For a MASO DN layer $f_\Theta^{(\ell)}$, we have

$$\left\| f_\Theta^{(\ell)}(\mathbf{z}_1) - f_\Theta^{(\ell)}(\mathbf{z}_2) \right\|^2 \leq \sum_{k=1}^{D^{(\ell)}} \max_{r=1,\dots,R^{(\ell)}} \left\| [A^{(\ell)}]_{k,r,\cdot} \right\|^2 \|\mathbf{z}_1 - \mathbf{z}_2\|^2 \quad (29)$$

and thus its Lipschitz constant

$$\kappa^{(\ell)} = \sum_{k=1}^{D^{(\ell)}} \max_{r=1,\dots,R} \left\| [A^{(\ell)}]_{k,r,\cdot} \right\|^2. \quad (30)$$

In Appendix E.4 we calculate the Lipschitz constants of the standard DN operators and a complete DN constructed from them.

Proposition 10. The Lipschitz constants for the standard DN operators are given by: fully connected $\kappa_W^{(\ell)} \leq \|W^{(\ell)}\|^2$; convolution $\kappa_C^{(\ell)} \leq \|C^{(\ell)}\|^2$; ReLU, leaky-ReLU, absolute value $\kappa_\sigma^{(\ell)} \leq (D^{(\ell)})^2$; max-pooling $\kappa_p^{(\ell)} \leq (D^{(\ell)})^2$; softmax $\kappa \leq \frac{C-1}{C^2}$.

Proposition 11. For a DN f_Θ constructed by composing L MASOs and any two input signals $\mathbf{x}_1, \mathbf{x}_2$, we have

$$\|f_\Theta(\mathbf{x}_1) - f_\Theta(\mathbf{x}_2)\|^2 \leq \left(\prod_{\ell=1}^L \kappa^{(\ell)} \right) \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \quad (31)$$

and thus its Lipschitz constant

$$\kappa = \left(\prod_{\ell=1}^L \kappa^{(\ell)} \right), \quad (32)$$

where the $\kappa^{(\ell)}$ are obtained from Proposition 10 according to the DN's topology.

From Proposition 10, note the potentially large Lipschitz constants for all operators but the softmax; indeed it is the only contraction, in general. These large constants can reinforce to produce a potentially very large overall Lipschitz constant for the complete DN in (32). There are several potential remedies: i) reduce the norms of the fully connected and convolution operators, ii) make the DN narrower by reducing $D^{(\ell)}$, and iii) invent new, more stable activation and pooling operators.

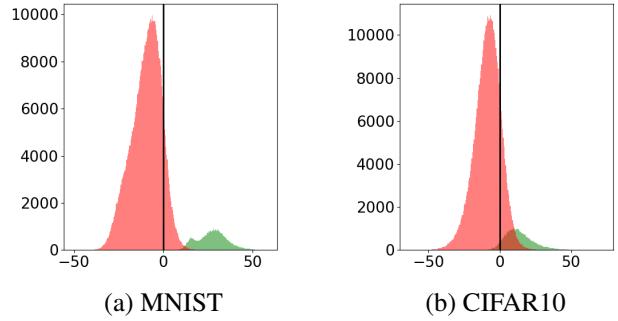


Figure 3. Empirical study of the implications of Proposition 12, Part 1. We illustrate the bimodality of $\langle [A^*[\mathbf{x}_n]]_{c,\cdot}, \mathbf{x}_n \rangle$ for the large CNN matched filterbank trained on (a) MNIST and (b) CIFAR10. Training used batch normalization and bias units. Results are similar when trained with bias and no batch-normalization. The green histogram summarizes the output values for the correct class (top half of (33)), while the red histogram summarizes the output values for the incorrect classes (bottom half of (33)). The easier the classification problem (MNIST), the more bimodal the distribution.

5.6. Collinear Templates and Data Set Memorization

Under the matched filterbank interpretation of a DN developed in Section 5.1, the optimal template for an image \mathbf{x} of class c is a scaled version of \mathbf{x} itself. But what are the optimal templates for the other (incorrect) classes? In an idealized setting, we can answer this question; the proof is provided in Appendix E.5.

Proposition 12. Consider an idealized DN consisting of a composition of MASOs that has sufficient approximation power to span arbitrary MASO matrices $A[\mathbf{x}_n]$ from (17) for any input \mathbf{x}_n from the training set. Train the DN to classify among C classes using the training data $\mathcal{D} = (\mathbf{x}_n, y_n)_{n=1}^N$ with normalized inputs $\|\mathbf{x}_n\|_2 = 1 \forall n$ and the cross-entropy loss $\mathcal{L}_{CE}(y_n, f_\Theta(\mathbf{x}_n))$ with the addition of the regularization constraint that $\sum_c \|A[\mathbf{x}_n]_{c,\cdot}\|_2 < \alpha$ with $\alpha > 0$. At the global minimum of this constrained optimization problem, the rows of $A^*[\mathbf{x}_n]$ (the optimal templates) have the following form:

$$[A^*[\mathbf{x}_n]]_{c,\cdot} = \begin{cases} +\sqrt{\frac{(C-1)\alpha}{C}} \mathbf{x}_n, & c = y_n \\ -\sqrt{\frac{\alpha}{C(C-1)}} \mathbf{x}_n, & c \neq y_n \end{cases} \quad (33)$$

In short, the idealized DN will memorize a set of collinear templates.

We report on an empirical study of the implications of this result in Figures 3 and 4. Figure 3 confirms the bimodality of the inner products between the signal and the correct/incorrect templates on the MNIST and CIFAR10 datasets. The distance between the green and red histograms is directly related to the difficulty of the classification task. Figure 4 plots

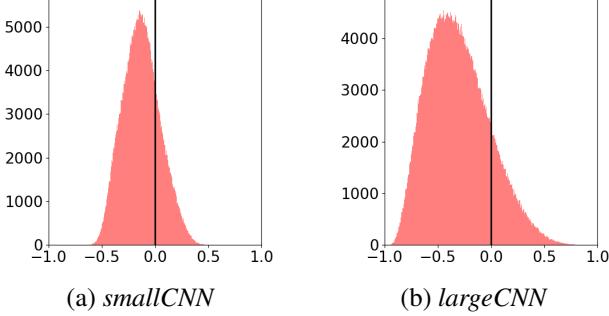


Figure 4. Empirical study of the implications of Proposition 12, Part 2. Training used batch normalization and bias units. Results are similar when trained with bias and no batch-normalization. We plot the normalized inner products (cosine similarities) between each set of correct and incorrect class templates for (a) *smallCNN* and (b) *largeCNN* trained on MNIST. Increasing the DN’s capacity (from *smallCNN* to *largeCNN*) leads to templates that conform to (33) more closely.

histograms of the normalized inner products (cosine similarities) between each set of correct and incorrect class templates $\text{hist}\{d([A[\mathbf{x}_n]]_{y_n,.}, [A[\mathbf{x}_n]]_{c,.}) \forall c \neq y_n \forall n\}$ with $d(u,v) = \frac{\langle u,v \rangle}{\|u\|\|v\|}$ for *smallCNN* and *largeCNN* on the MNIST dataset. We clearly see that increasing the DN’s capacity leads to templates that are more collinear and more the negative of the correct class template, again confirming (33). In the limit with the idealized DN, (33) suggests that these histograms will concentrate around -1 .

5.7. New DNs with Orthogonal Templates

While a DN’s signal-dependent matched filterbank (17) is optimized for classifying signals immersed in additive white Gaussian noise, such a statistical model is overly simplistic for most machine learning problems of interest. In practice, errors will arise not just from noise but also from nuisance variations in the inputs, such as arbitrary object rotations and positions, occlusions, lighting conditions, etc. The effects of these nuisances are only poorly approximated as Gaussian random noise. Limited work has been done on filterbanks for classification in non-Gaussian noise; one promising direction involves using not matched but rather *orthogonal templates* (Eldar & Oppenheim, 2001).

For a MASO DN’s templates to be orthogonal for all inputs, it is sufficient that the rows of the matrix $W^{(L)}$ in the final linear classifier layer be orthogonal. This weak constraint on the DN still enables the earlier layers to create a high-performance, class-agnostic, featurized representation (recall the discussion just below (17)). To create orthogonal templates during learning, we simply add to the standard (potentially regularized) cross-entropy loss function \mathcal{L}_{CE} from (3) a term that penalizes non-zero off-diagonal entries in the matrix $W^{(L)}(W^{(L)})^T$ leading to the new loss with

extra penalty

$$\mathcal{L}_{CE} + \lambda \sum_{c_1 \neq c_2} \left| \left\langle [W^{(L)}]_{c_1,.}, [W^{(L)}]_{c_2,.} \right\rangle \right|^2. \quad (34)$$

The parameter λ controls the tradeoff between cross-entropy minimization and orthogonality preservation. Conveniently, when minimizing (34) via backpropagation, the orthogonal rows of $W^{(L)}$ induce orthogonal backpropagation updates for the various classes.

We now empirically demonstrate that orthogonal templates lead to significantly improved classification performance. We conducted a range of experiments with three conventional DN architectures – *smallCNN*, *largeCNN*, and *ResNet4-4* – trained on CIFAR10 and CIFAR100. See Appendix F for the network architecture details plus additional results for the SVHN dataset. Each DN employed ReLU activations, max-pooling, bias, and batch normalization prior to each ReLU. For learning, we used the Adam optimizer with an exponential learning rate decay. All inputs were centered to zero mean and scaled to a maximum value of one. No further pre-processing was performed, such as ZCA whitening (Nam et al., 2014). We assessed how the classification performance of the DN changed as we varied the orthogonality penalty λ in (34). For each configuration of DN architecture, training dataset, learning rate, and penalty λ , we averaged over 15 runs to estimate the average performance and standard deviation.

The results for the *largeCNN* on CIFAR100 in Figure 5 indicate that the benefits of the orthogonality penalty emerge distinctly as soon as $\lambda > 0$.

We delve deeper into the performance of the *smallCNN* on CIFAR10 in Figure 6. In addition to improved final accuracy and generalization performance, we see that template orthogonality reduces the tendency of the DN to overfit. One explanation is that the orthogonal weights $W^{(L)}$ positively impact not only the prediction but also the backpropagation via orthogonal gradient updates. A range of additional experiments are provided in Appendix G.

6. DN’s Intrinsic Multiscale Partition

Like any spline, it is the interplay between the (affine) spline mappings and the input space partition that work the magic in a MASO DN. Recall from Section 3 that a MASO has the attractive property that it implicitly partitions the input signal space \mathbb{R}^D as a function of its slope and offset parameters. The induced partition Ω opens up a new geometric avenue to study how a MASO-based DN clusters and organizes signals in a hierarchical fashion. In particular, we will see that there are direct links with *vector quantization* (VQ) and *K-means clustering*.⁷

⁷Do not confuse the K of *K*-means with the MASO dimension.

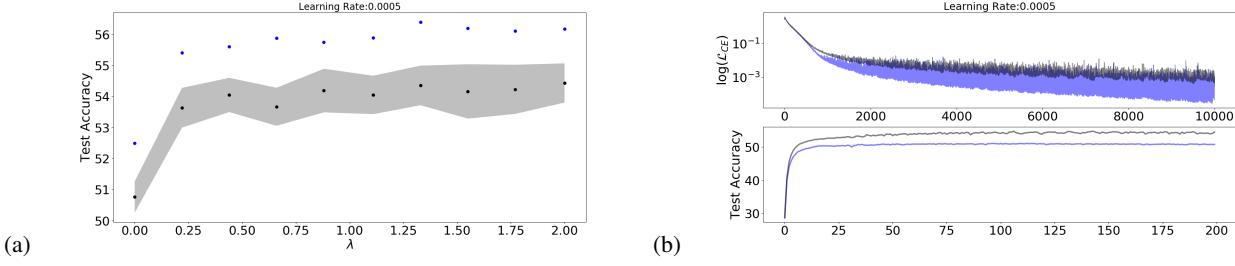


Figure 5. Orthogonal templates significantly boost DN performance with *no change to the architecture*. (a) Classification performance of the *largeCNN* trained on CIFAR100 for different values of the orthogonality penalty λ in (34). We plot the average (back dots), standard deviation (gray shade), and maximum (blue dots) of the test set accuracy over 15 runs. (b, top) Training set error. The blue/black curves corresponds to $\lambda = 0/1$. (b, bottom) Test set accuracy over the course of the learning.

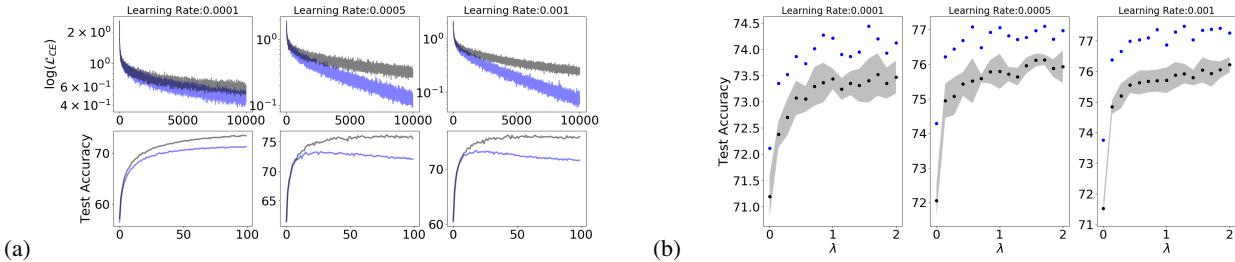


Figure 6. Orthogonal templates help to prevent overfitting with *smallCNN* trained on CIFAR10. (a) In black, standard training and in blue training using the orthogonality penalty $\lambda = 1$ in (34) averaged over 15 runs. (b) Test set accuracy of the templates as we vary the orthogonality penalty λ in (34). We plot the average, standard deviation, and maximum test set accuracy over 15 runs.

6.1. Local MASO Partition

Level ℓ of a MASO DN directly partitions its input space $\mathbb{R}^{D^{(\ell-1)}}$ and thus indirectly influences the partitioning of the overall input signal space \mathbb{R}^D (recall that $D^{(0)} = D$). We begin by developing a formula for the input space partition $\omega^{(\ell)}[k, r]$ in $\mathbb{R}^{D^{(\ell-1)}}$ that is effected by the k^{th} of the $K = D^{(\ell)}$ affine spline functions comprising the MASO. Partitions at each level ℓ are hierarchically chained together to effect the global partition; we discuss this in more detail below.

We illustrate first with the partitions effected by the ReLU activation and max-pooling operators. A ReLU operator at level ℓ and input/output k splits its layer input space into two half-planes depending on the sign of the input. We denote by $\omega^{(\ell)}[k, r]$ the r^{th} region at output k and level ℓ ; since $R = 2$ for a ReLU spline, we have

$$\begin{aligned}\omega^{(\ell)}[k, 1] &= \left\{ \mathbf{z}^{(\ell-1)}(\mathbf{x}) \in \mathbb{R}^{D^{(\ell)}} : [\mathbf{z}^{(\ell-1)}(\mathbf{x})]_k < 0 \right\}, \\ \omega^{(\ell)}[k, 2] &= \left\{ \mathbf{z}^{(\ell-1)}(\mathbf{x}) \in \mathbb{R}^{D^{(\ell)}} : [\mathbf{z}^{(\ell-1)}(\mathbf{x})]_k \geq 0 \right\}.\end{aligned}\quad (35)$$

As detailed in Appendix C.2, the complete MASO partition of the ReLU activation operator's input space $\mathbb{R}^{D^{(\ell-1)}}$ is the intersection of the half-spaces of each of the K input/output

pairs. Preceding an activation operator like ReLU with a fully connected or convolution operator simply rotates the half-space partition in $\mathbb{R}^{D^{(\ell-1)}}$; addition of a bias simply translates the origin. The number of partition regions is combinatorially large – up to $2^{D^{(\ell)}}$ for ReLU and up to $R^{D^{(\ell)}}$ for an activation function with $R \geq 2$. These partition regions can be of infinite volume, but they are always convex.

Consider next the partition effected by max-pooling. A max-pooling operator at layer ℓ also partitions its layer input space $\mathbb{R}^{D^{(\ell-1)}}$ into a combinatorially large number of regions (up to $(\#R^{(\ell)})^{D^{(\ell)}}$, where $\#R^{(\ell)}$ is the size of the pooling region). Each partition region will depend on the location of the arg max entry for each output

$$\begin{aligned}\omega^{(\ell)}[k, r] &= \left\{ \mathbf{z}^{(\ell-1)}(\mathbf{x}) \in \mathbb{R}^{D^{(\ell)}} : \right. \\ &\quad \left. r = \arg \max_{d \in \mathbb{R}_k^{(\ell)}} [\mathbf{z}^{(\ell-1)}(\mathbf{x})]_d \right\}\end{aligned}\quad (36)$$

for $r = 1, \dots, \#R^{(\ell)}$.

More generally, to find the partition of the layer input space effected by a layer- ℓ MASO constructed by concatenating

$K = D^{(\ell)}$ identical max-affine spline functions,⁸ we first rewrite the layer calculation (6) as

$$[\mathbf{z}^{(\ell)}(\mathbf{x})]_k \\ = \max_{r=1,\dots,R^{(\ell)}} \left\langle [A^{(\ell)}]_{k,r,.}, \mathbf{z}^{(\ell-1)}(\mathbf{x}) \right\rangle + [B^{(\ell)}]_{k,r} \quad (37)$$

$$= \sum_{r=1}^R [T^{(\ell)}(\mathbf{z}^{(\ell)}(\mathbf{x}))]_{k,r} \\ \times \left(\left\langle [A^{(\ell)}]_{k,r,.}, \mathbf{z}^{(\ell-1)}(\mathbf{x}) \right\rangle + [B^{(\ell)}]_{k,r} \right). \quad (38)$$

Here,

$$[T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}))]_{k,r} \\ = \mathbb{1}\left(r = \arg \max_{r=1,\dots,R} \left\langle [A^{(\ell)}]_{k,r,.}, \mathbf{z}^{(\ell-1)}(\mathbf{x}) \right\rangle + [B^{(\ell)}]_{k,r}\right) \quad (39)$$

is a *selection matrix* whose k^{th} row comprises a 1-hot selection vector (Kronecker delta function) that selects the partition region r that maximizes (38) in the k^{th} output dimension. We denote the corresponding index by $[t^{(\ell)}]_k$. Then, we can write

$$\omega^{(\ell)}[k, r] = \\ \left\{ \mathbf{z}^{(\ell-1)}(\mathbf{x}) \in \mathbb{R}^{D^{(\ell)}} : [T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}))]_{k,r} = 1 \right\}. \quad (40)$$

When computing the MASO output via the max operator as in (37), we do not have explicit access to the selection matrix $T^{(\ell)}$ in (39). However, there is a simple strategy to compute it a posteriori.

Proposition 13. *For any MASO, the selection matrix $T^{(\ell)}$ in (39) can be computed via*

$$[T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}))]_{k,r} = \\ \frac{\partial \max_r \left\langle [A^{(\ell)}]_{k,r,.}, \mathbf{x} \right\rangle + [B^{(\ell)}]_{k,r}}{\partial \left(\left\langle [A^{(\ell)}]_{k,r,.}, \mathbf{x} \right\rangle + [B^{(\ell)}]_{k,r} \right)}. \quad (41)$$

The derivative in (41) can be simply and efficiently computed via backpropagation on the DN.

It is easy to show (see Appendix C.2) that the overall partitioning of the layer- ℓ MASO input space $\mathbb{R}^{D^{(\ell-1)}}$ is simply the intersection of the partitions effected by each of the $k = 1, \dots, D^{(\ell)}$ output dimensions

$$\omega^{(\ell)}[r_1, \dots, r_K] = \bigcap_{k=1}^{D^{(\ell)}} \omega^{(\ell)}[k, r_k], \quad (42)$$

⁸ See Appendix C.2 for the general case of concatenating K non-identical splines.

where $r_k \in \{1, \dots, R^{(\ell)}\}$. The final partition contains up to $(R^{(\ell)})^{D^{(\ell)}}$ convex and connected regions; in practice, however, many of them could have zero volume. (Our simple statistical study in Section 6.4 bears out this conclusion.) We leave it as an exercise for the reader to show that the intersection (42) tessellates the entire layer input space $\mathbb{R}^{D^{(\ell)}}$.

6.2. Local Partition in the Input Signal Space

We can map the partition of the layer- ℓ MASO's input space $\mathbb{R}^{D^{(\ell-1)}}$ back to a corresponding partition of the input signal space \mathbb{R}^D . We do this by pulling the partition (42) back from layer ℓ to the input via

$$\omega_{\text{in}}^{(\ell)}[r_1, \dots, r_{D^{(\ell)}}] = \bigcap_{k=1}^{D^{(\ell)}} \omega_{\text{in}}^{(\ell)}[k, r_k], \quad (43)$$

where

$$\omega_{\text{in}}^{(\ell)}[k, r] = \left\{ \mathbf{x} \in \mathbb{R}^D : [T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}))]_{k,r} = 1 \right\}, \quad (44)$$

and, as in (40) and (42), k indexes the k^{th} output of the ℓ^{th} layer and $r, r_k \in \{1, \dots, R^{D^\ell}\}$. This partition can be interpreted as a VQ of the training data that has been featurized by layers 1 through $\ell - 1$. Interestingly, the total number of possible regions remains R^{D^ℓ} . However, the regions are no longer necessarily convex nor connected, due to the nonlinear transformations effected by layers 1 through $\ell - 1$. Moreover, the spline function defined in each region is no longer a fixed affine function; rather it is a piecewise affine function, in general. (We rectify this latter situation in the next subsection.)

We have so far been stymied in our quest for a closed-form formula for the regions (43) for an arbitrary composition of MASOs. However, (44) enables a simple strategy to estimate the partition induced by the DN layers $1, \dots, \ell$. The algorithm runs as follows: First, sample the overall input signal space \mathbb{R}^D to create a collection of inputs $\{\mathbf{x}'\}$. Second, pass each \mathbf{x}' through the DN until the desired level ℓ to compute the layer outputs $\mathbf{z}^{(\ell')}(\mathbf{x}')$ as well as the selection matrices $T^{(\ell')}(\mathbf{z}^{(\ell-1)}(\mathbf{x}'))$ via (41) for $\ell' = 1, \dots, \ell$. Third, label each \mathbf{x}' by the region that is selected by $[T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}'))]_{k,r}$ based on (44). The finer the sampling of the input space, the more accurate the estimate of the partition region boundaries.

6.3. Global DN Partition

In the previous subsection, we found the partition of the input signal space \mathbb{R}^D effected by (only) the layer- ℓ MASO. Intersecting the partitions of the input signal space induced

by all ℓ' layers between the input and the output of layer ℓ yields the global partition generated by the DN up to layer ℓ

$$\omega_g^{(\ell)} \left[\left\{ r_1^{(\ell')}, \dots, r_{D^{(\ell')}}^{(\ell')} \right\}_{\ell'=1}^{\ell} \right] = \bigcap_{\ell'=1}^{\ell} \omega_{in}^{(\ell')} \left[r_1^{(\ell')}, \dots, r_{D^{(\ell')}}^{(\ell')} \right]. \quad (45)$$

This global partition is precisely the partition Ω^S of the ASO referred to in Theorem 1; it takes into account the joint configuration of the selection matrices $T^{(\ell')}$ for $\ell' = 1, \dots, \ell$.

6.4. Example: Partition Visualization and Statistics

While visualizing the partition of a high-dimensional DN is impossible, a miniature case study provides some useful intuition. Consider an old-school neural network with input signal dimension $D = 2$ and $L = 3$ layers comprised of the composition of three fully connected operators interspersed with two activation operators plus a final softmax operator. We consider three different activation functions: ReLU, leaky ReLU, and absolute value. The dimensionalities of the layers are: $D^{(0)} = 2$, $D^{(1)} = 45$, $D^{(2)} = 3$, $D^{(3)} = 4$. We seek to solve a $C = D^{(3)} = 4$ -class classification problem. See Appendix F for a diagram. We trained the DN using biases and batch normalization on a training set of 5000 data points from each of the four classes using the Adam optimizer (Kingma & Ba, 2014) and reached 98% accuracy on the training set. The use of batch normalization in this simple setting does not impact the results substantively other than to shift the region boundaries to be more centered around the clusters of training data points.

Figure 7 displays the input signal space \mathbb{R}^2 with the training data points overlaid with the partitions $\omega_{in}^{(1)}[k, r]$,⁹ $\omega_{in}^{(2)}[k, r]$, and $\omega_g^{(2)}[k, r]$ as defined in (42), (43), and (45), respectively. Figure 8 displays the input signal space \mathbb{R}^2 with the training data points overlaid by the DN’s function approximation for the three different activation functions. We can make some interesting observations. First, even for such a simple network, the learned partition regions are quite complex. For instance, the regions induced by the second layer are neither convex nor connected. Second, there exist partition regions that contain training data points from more than one class. Correct classification remains possible, however, thanks to the affine function applied within each region. Third, thanks to the complexity of the partition regions, even this simple affine spline can realize quite complicated functional mappings. Fourth, the partitions and functional mappings learned for each different activation function are very different.

We now provide some statistics regarding the partition re-

Partition	Number of Nonzero Volume Regions			
	Initialization (BN)	(BN)	Trained (BN)	(BN)
$\omega_{in}^{(1)}$	90	90	573	808
$\omega_{in}^{(2)}$	6	7	6	8
$\omega_g^{(2)}$	99	100	442	1032

Table 3. Number of nontrivial partition regions having nonzero volume for the input signal space partition induced by $\omega_{in}^{(1)}$ from layer 1 and $\omega_{in}^{(2)}$ from 2 of the toy DN from Figure 7. The results with and without batch normalization are similar: most of the exponentially many possible regions are trivial when compared to the number of possible regions. However, the use of BN allow more region to be effectively nontrivial for both layers.

Partition	Number of Nonempty Regions			
	Initialization (BN)	(BN)	Trained (BN)	(BN)
$\omega_{in}^{(1)}$	90	90	205	305
$\omega_{in}^{(2)}$	6	7	5	8
$\omega_g^{(2)}$	98	98	237	389

Table 4. Number of partition regions that are occupied by at least one training data point for the toy DN from Figure 7. As for the number of nontrivial region, the use of BN seems to allow more regions to be non empty for both layers.

gions induced by the first two layers of the toy DN. Table 3 tabulates the number of non-trivial partition regions (those with nonzero volume) of the toy DN averaged over 10 runs. We see that most of exponentially many possible regions have zero volume and are trivial. We also provide the same statistics with the leaky ReLU and absolute value nonlinearities in Appendix Sec. H.

Interestingly, many of the non-trivial partition regions are actually empty (i.e., devoid of any training data points). Table 4 reports the statistics, and Figure 9 plots the number of training data points per region in descending order, where only the nonempty regions are displayed for clarity. We see that batch normalization yields a partition where the data points are more evenly distributed across the regions.

Keep in mind that the above results are only for a very small, toy DN. An interesting research direction is extending this analysis to the kinds of large DNs used in practice.

6.5. Connection to Vector Quantization and K-Means

The MASO partition is equivalent to a *vector quantization* (VQ) of its input space (Gersho & Gray, 2012). Indeed, the connection between max-affine splines and K-means (a particular flavor of VQ) has been pointed out in (Magnani & Boyd, 2009) in the least-squares regression setting. VQ is a classical technique for lossy compression and clustering that has also been productively applied to image encoding

⁹ Recall that in the first layer $\omega_g^{(1)} = \omega_{in}^{(1)} = \omega^{(1)}$.

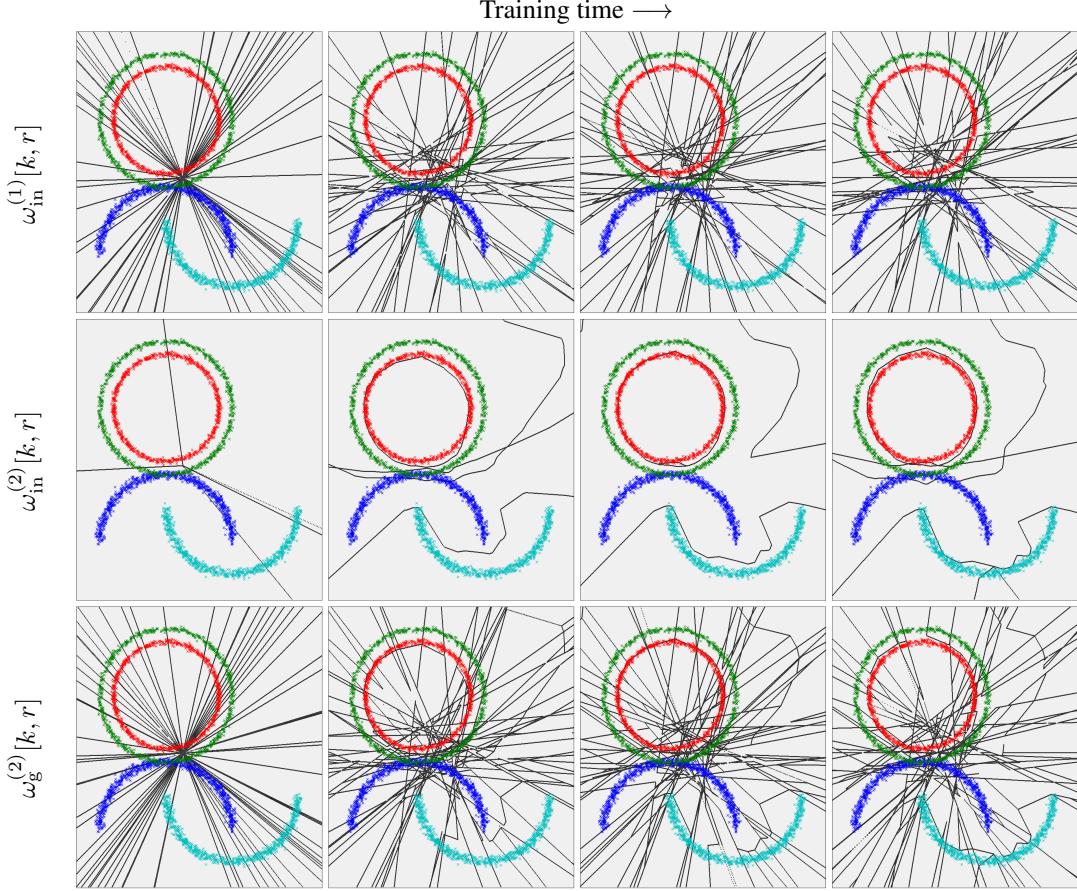


Figure 7. Illustration of the input signal space partition of a simple three-layer DN using ReLU activation and batch normalization. All biases were initialized to zero. (See the text and Appendix F for the network details.) The top row displays the input signal space \mathbb{R}^2 with the training data points from the $C = 4$ classes (colored green, red, blue, and cyan) overlaid with the partition $\omega^{(1)}[k, r] = \omega_{\text{in}}^{(1)}[k, r] = \omega_g^{(1)}[k, r]$ effected by the $D^{(1)} = 45$ outputs of the first layer at four different time epochs in the training (initialization, 40, 80, 120 epochs). The middle row displays the local partition in the input space $\omega_{\text{in}}^{(2)}[k, r]$ effected by the $D^{(2)} = 3$ outputs of the second layer. It is interesting to note the region at the bottom of the green and red circles that contains points from both classes. Correct classification remains possible in this region, however, thanks to the affine function applied in that region. The bottom row displays the global partition in the input space $\omega_g^{(2)}[k, r]$ effected by the first and second layers. Note the nonconvexity of the regions $\omega_{\text{in}}^{(2)}[k, r]$ and $\omega_g^{(2)}[k, r]$.

(Nasrabadi & King, 1988), texture synthesis (Wei & Levoy, 2000), and beyond.

The partition (40) induced by the k^{th} of the $K = D^{(\ell)}$ max-affine splines in one MASO layer can be interpreted as approximating/encoding the input vector $\mathbf{z}^{(\ell-1)}(\mathbf{x})$ by a “partition centroid” determined by the 1-hot vector $[T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}))]_{k,:}$ and layer parameters $A^{(\ell)}[\mathbf{x}], B^{(\ell)}[\mathbf{x}]$ via (38). The partition (42) induced by the entire MASO layer can be interpreted as a *collaborative VQ* (Zemel, 1994; Hinton & Zemel, 1994), which is particularly adept at representing complicated datasets with multiple object/features present simultaneously. Such techniques are also related to *factorial hidden Markov models* and other factorial models (Ghahramani & Jordan, 1996), which have been used to

represent the interactions among multiple loosely coupled processes. We study the collaborative partitioning aspects of MASOs in more detail in Appendix C.2.

Remark 2. At this point, we can make Remark 1 more concrete by clarifying that same affine mapping parameters $A^{(\ell)}[\mathbf{x}], B^{(\ell)}[\mathbf{x}]$ are applied not just to the signal \mathbf{x} but to all signals in the same partition region containing \mathbf{x} , as in $A^{(\ell)}[T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}))], B^{(\ell)}[T^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x}))]$.

In keeping with the VQ interpretation of the MASO partition, we can also decode a signal approximation μ_r for all signals in region r . The standard approach is via K -means (Jain, 2010). (Note that in this case, the K of K -means equals R , the number of partition regions; we will use R from now on and refer to the approach as R -means.) For a

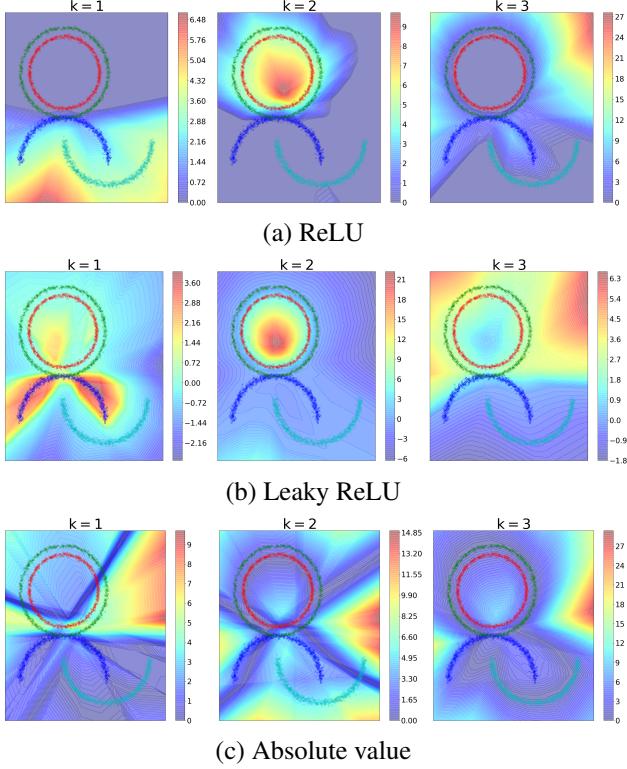


Figure 8. Illustration of the learned piecewise-affine function mapping of the DN from Figure 7 from \mathbf{x} in the input signal space \mathbb{R}^2 to $[\mathbf{z}^{(2)}(\mathbf{x})]_k$ at the output of the second layer $D^{(2)} = 3$ overlaid on the training data points from the $C = 4$ classes. Each row corresponds to a different activation function. The three columns correspond to the output dimensions $k = 1, 2, 3$. It is interesting to confirm visually how the function mappings in (a) are a fixed affine function on each partition region in the bottom right image in Figure 7. Due to the high number of global DN regions (onto which the DN mapping is affine) the overall mapping displayed here can not be visually seen as being piecewise affine.

set of (labelled or unlabelled) training signals $(\mathbf{x}_n)_{n=1}^N$, the goal of R -means is to learn R centroids μ_r , $r = 1, \dots, R$ that solve

$$\min_{r=1, \dots, R} \left\| \sum_{n=1}^N \boldsymbol{\mu}_r - \mathbf{x}_n \right\|^2. \quad (46)$$

It then assigns signal \mathbf{x} to the closest centroid via

$$t(\mathbf{x}_n) = \arg \min_{r=1, \dots, R} \|\boldsymbol{\mu}_r - \mathbf{x}_n\|^2. \quad (47)$$

While solving (46) is computationally intractable, a number of suboptimal approximations have been developed, the most celebrated being the Lloyd algorithm (Jain, 2010).

Using a special parameterization of the DN, the MASO encoding/decoding interpretation conforms precisely with a

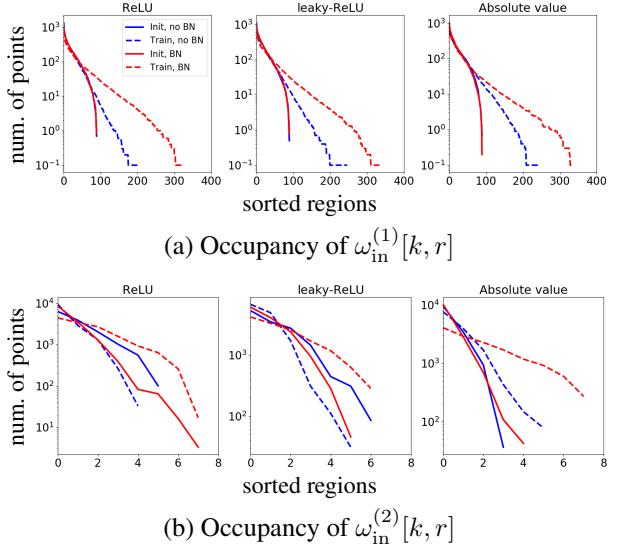


Figure 9. Occupancy of the partition regions induced in the input signal space by (a) $\omega_{in}^{(1)}$ from layer 1 and (b) $\omega_{in}^{(2)}$ from layers 1 and 2 of the DN from Figure 7. We sorted the regions according to how many test data points each contains and plot on a log scale the results for the ReLU, leaky ReLU, and absolute value activation functions. The four curves on each plot correspond to the occupancy at (random) initialization and after training and with and without batch normalization. We averaged the results over 10 runs. A curve touching the horizontal axis indicates that all regions with higher sorted index are empty.

R -means-based VQ system. By massaging the formula for the 1-hot index of a single MASO layer output (recall (39))

$$\begin{aligned} [t^{(\ell)}(\mathbf{x}_n)]_k &= \arg \max_{r=1, \dots, R} \left(\langle [A^{(\ell)}]_{k,r,.}, \mathbf{z}^{(\ell-1)(\mathbf{x}_n)} \rangle \right. \\ &\quad \left. + [B^{(\ell)}]_{k,r} \right) \\ &= \arg \min_{r=1, \dots, R} \left(\| [A^{(\ell)}]_{k,r,.} - \mathbf{z}^{(\ell-1)(\mathbf{x}_n)} \|^2 \right. \\ &\quad \left. - \| [A^{(\ell)}]_{k,r,.} \|^2 - 2 [B^{(\ell)}]_{k,r} \right) \quad (48) \end{aligned}$$

we see that (48) corresponds precisely to (46) with $\mu_r = [A^{(\ell)}]_{k,r,.}$ provided $\| [A^{(\ell)}]_{k,r,.} \|^2 = -2 [B^{(\ell)}]_{k,r}$. In this case, the partition region is the *Voronoi tiling* of the input space according to (47). Figure 10 plots the partition regions and centroids $\mu_r = [A^{(\ell)}]_{k,r,.}$ for such a constrained max-affine spline. We memorialize this result in the following proposition.

Proposition 14. *The input space partition of the k^{th} output of a MASO at level ℓ with parameters $[B^{(\ell)}]_{k,r} = -\frac{1}{2} \| [A^{(\ell)}]_{k,r,.} \|^2$ is equivalent to an $R^{(\ell)}$ -means partition using the centroids $\mu_r = [A^{(\ell)}]_{k,r,.}$. The partition is a Voronoi tiling with respect to the centroids.*

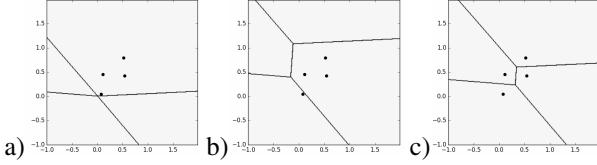


Figure 10. Simple example of the correspondence between VQ/R-means and a MASO partition at level ℓ . We illustrate three different partitions of the $D^{(\ell-1)} = 2$ -dimensional input space via a MASO with $R^{(\ell)} = 4$ regions induced by the same slope parameters $[A^{(\ell)}]_{k,r,\cdot}$, but three different biases: (a) $[B^{(\ell)}]_{k,r} = 0$, (b) random $[B^{(\ell)}]_{k,r}$, and (c) $[B^{(\ell)}]_{k,r} = -\frac{1}{2}\|[A^{(\ell)}]_{k,r,\cdot}\|^2$. In (c) the slope parameters correspond to K -means centroids as in (46), and the partition is the Voronoi tiling based on (47).

6.6. A New Image Distance Based on DN VQ

The structure of the DN partition regions in the input signal space provide new tools to study and exploit the geometrical properties of the training and test data. While, in general, it is impractical to instantiate the regions, we can access their membership via the selection matrix $T^{(\ell)}(z^{(\ell)}(x))$ from (39). Since two signals sharing the same selection matrices inhabit the same partition region, the concept of “nearest neighbors” inspires us to create a new signal distance that measures the distance between the partitions each signal inhabits. Given signals x_1 and x_2 , we define the distance as

$$d\left(T^{(\ell)}(x_1), T^{(\ell)}(x_2)\right) = \frac{1}{D^{(\ell)}} \sum_{k=1}^{D^{(\ell)}} d'\left([T^{(\ell)}(x_1)]_{k,\cdot}, [T^{(\ell)}(x_2)]_{k,\cdot}\right), \quad (49)$$

with d' a distance function of your choice. This distance has the interpretation in terms of VQ of measuring how many of the $D^{(\ell)}$ outputs of layer ℓ share the same encoding of the input. The fact that $[T^{(\ell)}(x)]_{k,\cdot}$ is a one-hot vector makes any p -norm based distance degenerate to

$$d'(\mathbf{u}, \mathbf{v}) = \begin{cases} 0 & \text{if } \mathbf{u} = \mathbf{v} \\ 1 & \text{otherwise} \end{cases} \quad (50)$$

For any two signals, this distance takes a value $\in [0, 1]$; it equals 0 if x_1 and x_2 share all of the same partitions and 1 if they share none of the same partitions. For a given MASO, the distance can be computed easily via (41). Development of more sophisticated measures that would take into account the similarity between regions based on the parameters $A^{(\ell)}, B^{(\ell)}$, is left for future work.

For a ReLU MASO, the distance amounts to simply counting how many entries of the layer ℓ output are both positive or 0 at the same indices $k = 1, \dots, D^{(\ell)}$ for x_1 and x_2 . For a max-pooling MASO, the distance amounts to simply counting how many arg max positions are the same for each of the max-pooling applications for x_1 and x_2 .

Figures 11–13 provide visualizations of the nearest neighbors of a test image using our new VQ-based distance for the first five layers of the smallCNN topology trained for classification on the CIFAR10, SVHN, and MNIST datasets. As defined above, the distance can take only $D^{(\ell)}$ different values $(0, 1/D^{(\ell)}, 2/D^{(\ell)}, \dots, 1)$, meaning that, given an image, there are $D^{(\ell)}$ equivalence classes of neighbors based on each of those possible values. Visual inspection of the figures highlights that, as we progress through the layers of the DN, similar images become closer in VQ distance but further in Euclidean distance. In particular, the similarity of the first 3 panels in (a)–(c) of the figures, which replicate the same experiment but using a DN trained with correct labels, a DN trained with incorrect labels, and a DN that is not trained at all, indicates that the early convolution layers of the DN are partitioning the training images based more on their visual features than their class membership.

The distance (49) is defined only with respect to the MASO at level ℓ . However, it is easily extended to take into account the composition of multiple MASO layers by composing their respective selection matrices.

7. Conclusions

We have applied the theory of max-affine spline operators (MASOs) to make new connections between deep networks (DNs) and approximation theory. We group our key findings into two classes. First, we showed that, conditioned on the input signal, the output of a large class of DNs can be written as a simple affine transformation of the input. This enabled us to link DNs directly to the classical theory of optimal classification via matched filters, provided insights into the effects of data memorization, and inspired a template orthogonality penalty that leads to significantly improved performance and reduced overfitting with no change to the DN architecture.

Second, we showed how the partition of the input signal space that is automatically induced by a MASO directly links DNs to the theory of vector quantization (VQ) and K -means clustering, which opens up a new geometric avenue to study how DNs organize signals. This viewpoint inspired our new VQ-based distance for signals and images that is not only useful for probing the inner workings of DNs but also could be of independent utility.

There are many avenues for future work; here is a brief sampler. New constraints on the templates beyond orthogonality could yield learning algorithms that boost performance even further. A deeper study of the spline approximation of non-convex activation functions like the sigmoid could lead to new insights into why they are preferable in certain network topologies (e.g., recursive networks). Extension of our analysis of the number of non-trivial and non-empty

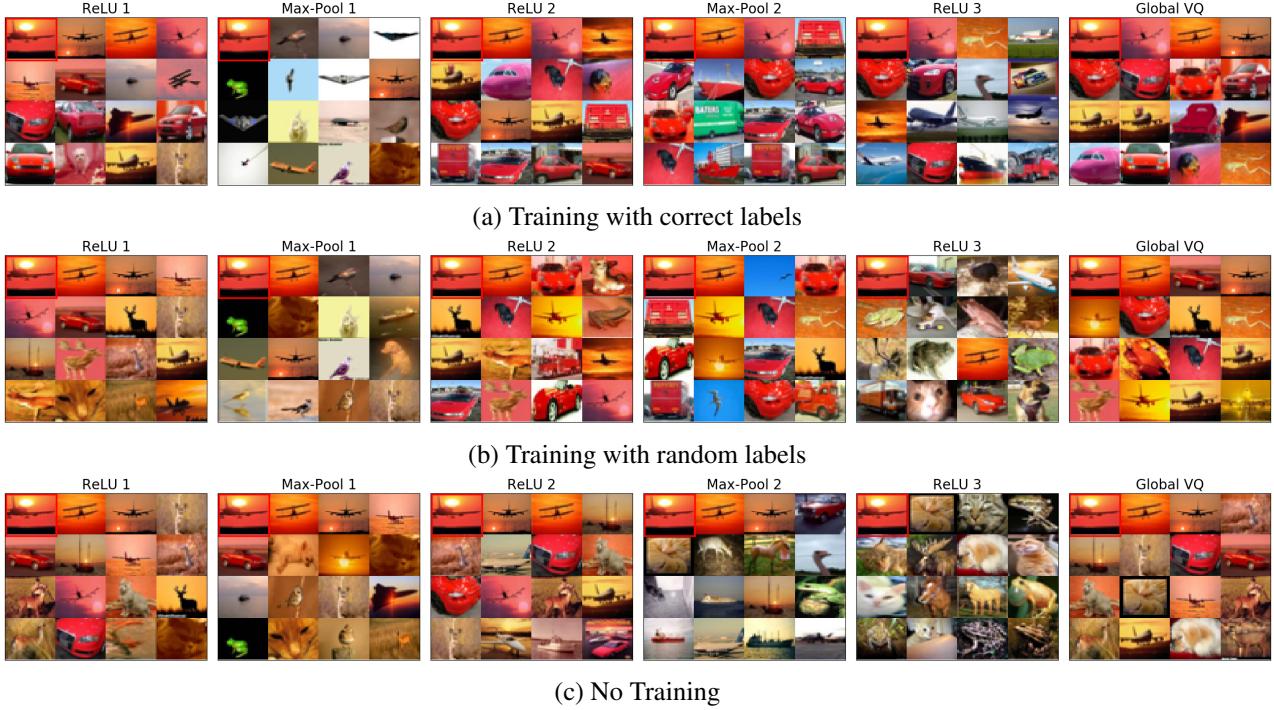


Figure 11. (a) Illustration of the nearest neighbors of a test image in terms of the VQ-based distance (49), (50) for various layers of the *smallCNN* trained on the CIFAR10 dataset. In each panel, we plot the target image α at top-left and its 15 nearest neighbors in the VQ distance at layer ℓ (see Appendix F for the experimental details). From left to right, the images display layers $\ell = 1, \dots, 6$, and we see that the nearest neighbors become more similar visually even as they become further in Euclidean distance. In particular, the first layer neighbors share similar colors and shapes (and thus are closer in Euclidean distance). Later layer neighbors mainly come from the same class independently of their color or shape. (b) The same experiment but with the DN trained on data with randomly shuffled labels. (c) The same experiment but with the DN initialized with random weights and zero biases and not trained. The similarity of the first 3 panels in (a)–(c) indicates that the early convolution layers of the DN are partitioning the training images based more on their visual features than their class membership. In (b), only after the fully connected layer (ReLU 3) is the partitioning influenced by the labels.

partition regions to the practical, large DNs could yield surprises. Since K -means is closely related to Gaussian mixture models, there are likely interesting connections with probabilistic models for deep learning (Patel et al., 2016) and probabilistic trees (Romberg et al., 1999). We can also study some of the popular DN tricks using MASOs. For example, *dropout* (Gal & Ghahramani, 2016) can be interpreted as randomly perturbing the VQ partition regions puncturing the selection matrices T^ℓ . This maps training inputs into different (incorrect) neighboring regions where a different (partially wrong) affine mapping is applied. Judicious use of this technique will result in more robust clustering/partitioning of the training data that ensures that neighboring regions have close mappings.

Acknowledgements

The authors would like to thank Romain Cosentino and Zichao (Django) Wang for constructive criticism of the manuscript, as well as Daniel LeJeune for drawing Figure 21. This work was partially supported by ARO grant W911NF-15-1-0316, AFOSR grant FA9550-14-1-0088, ONR grants N00014-17-1-2551 and N00014-18-12571, DARPA grant G001534-7500, and a DOD Vannevar Bush Faculty Fellowship (NSSEFF) grant N00014-18-1-2047.



Figure 12. Reprise of the experiment of Figure 11 for the SVHN dataset.

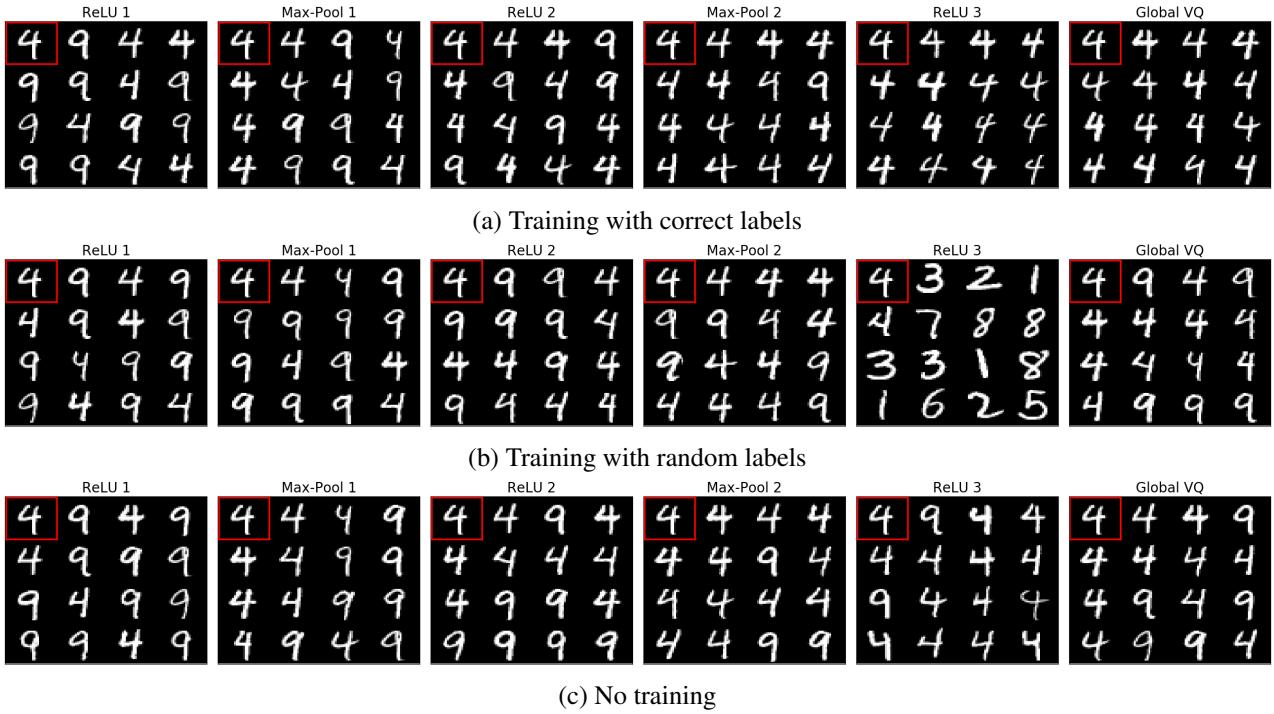


Figure 13. Reprise of the experiment of Figure 11 for the MNIST dataset.

SUPPLEMENTARY MATERIAL

A. Notation

x	Input/observation tensor of shape (K, I, J)
\mathbf{x}	Vectorized input of length D
$\hat{y}(x)$	Output/prediction associated with input x
x_n	Observation n
y_n	Label (target variable) associated with x_n , for classification $y_n \in \{1, \dots, C\}$, $C > 1$; for regression $y_n \in \mathbb{R}^C$, $C \geq 1$
\mathcal{D} (resp. \mathcal{D}_s)	Labeled training set with N (resp. N_s) samples $\mathcal{D} = (X_n, Y_n)_{n=1}^N$
\mathcal{D}_u	Unlabeled training set with N_u samples $\mathcal{D}_u = (X_n)_{n=1}^{N_u}$
$f_{\theta^{(\ell)}}$	Mapping of deep net (DN) layer at level ℓ with internal parameters $\theta^{(\ell)}$, $\ell = 1, \dots, L$
Θ	Collection of all DN parameters $\Theta = \{\theta^{(\ell)}, \ell = 1, \dots, L\}$
f_Θ	DN mapping with $f_\Theta : \mathbb{R}^D \rightarrow \mathbb{R}^C$
$(C^{(\ell)}, I^{(\ell)}, J^{(\ell)})$	Shape of the representation at layer ℓ , with $(C^{(0)}, I^{(0)}, J^{(0)}) = (K, I, J)$, and $(C^{(L)}, I^{(L)}, J^{(L)}) = (C, 1, 1)$
$D^{(\ell)}$	Dimension of the flattened representation at layer ℓ , with $D^{(\ell)} = C^{(\ell)} I^{(\ell)} J^{(\ell)}$, $D^{(0)} = D$, and $D^{(L)} = C$
$z^{(\ell)}(x)$	Representation of x (feature map) at layer ℓ in an unflattened format of shape $(C^{(\ell)}, I^{(\ell)}, J^{(\ell)})$, with $z^{(0)}(x) = x$
$[z^{(\ell)}(x)]_{k,i,j}$	Value of the representation of x at layer ℓ , channel c , and spatial position (i, j)
$\mathbf{z}^{(\ell)}(x)$	Representation of x at layer ℓ in a flattened format of dimension $D^{(\ell)}$
$[\mathbf{z}^{(\ell)}(x)]_k$	Value of $\mathbf{z}^{(\ell)}(x)$ at dimension k . One has $[z^{(\ell)}]_{c,i,j} = [\mathbf{z}^{(\ell)}]_k$ with $k = c \times I^{(\ell)} \times J^{(\ell)} + i \times J^{(\ell)} + j$

B. Detailed Background on Deep Networks

We now describe a typical deep (neural) network (DN) topology, the deep convolutional neural network (CNN), to highlight the way the previously described operators can be combined to provide powerful predictors. Its main development goes back at least as far as (LeCun et al., 1995) for digit classification. The astonishing results that a CNN can achieve come from the ability of the blocks to convolve the learned filter-banks with their input, separating the underlying features present relative to the task at hand. This is followed by a scalar activation nonlinearity and a spatial sub-sampling to select, compress, and reduce the redundant representation while highlighting task dependent features. Finally, a multilayer perceptron (MLP) simply acts as a nonlinear classifier.

In order to optimize the parameters Θ leading to the predicted output $\hat{y}(\mathbf{x})$, one makes use of (i) a labeled dataset $\mathcal{D} = \{(\mathbf{x}_n, y_n), n = 1, \dots, N\}$, (ii) a loss function $\mathcal{L} : \mathbb{R}^C \times \mathbb{R}^C \rightarrow \mathbb{R}$, (iii) a learning policy to update the parameters Θ . In the context of classification, the target variable \mathbf{x}_n associated to an input \mathbf{x}_n is categorical, with $y_n \in \{1, \dots, C\}$. In order to predict such target, the output of the last operator of a network $\mathbf{z}^{(L)}(\mathbf{x}_n)$ is transformed via a softmax nonlinearity (de Brébisson & Vincent, 2015). It is used to transform $\mathbf{z}^{(L)}(\mathbf{x}_n)$ into a probability distribution. The used loss function quantifying the distance between $\hat{y}(\mathbf{x}_n)$ and Y_n is the cross-entropy (CE). For regression problems, the target y_n is continuous and thus the final DNN output is taken as the prediction $\hat{y}(\mathbf{x}_n) = \mathbf{z}^{(L)}(\mathbf{x}_n)$. The loss function \mathcal{L} is usually the ordinary squared error (SE).

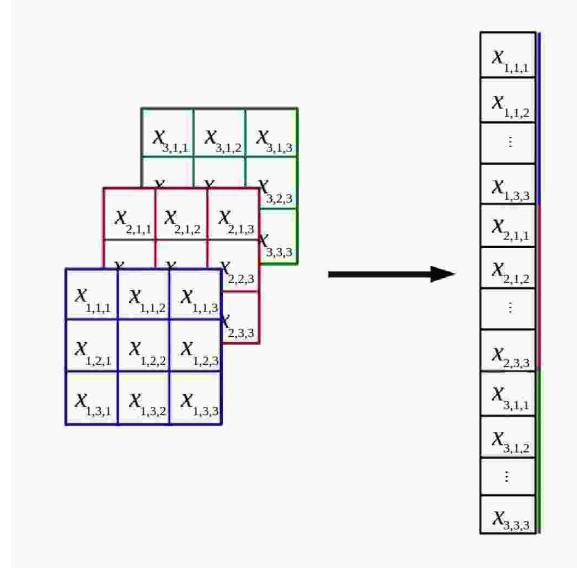


Figure 14. Reshaping of the multi-channel signal $z^{(\ell)}$ of shape $(3, 3, 3)$ to form the vector $\mathbf{z}^{(\ell)}$ of dimension 27.

Since all of the operations introduced above in standard DNNs are differentiable almost everywhere with respect to their parameters and inputs, given a training set and a loss function, one defines an update strategy for the weights Θ . This takes the form of an iterative scheme based on a first order iterative optimization procedure. Updates for the weights are computed on each input and usually averaged over *mini-batches* containing B exemplars with $B \ll N$. This produces an estimate of the correct update for Θ and is applied after each mini-batch. Once all the training instances of \mathcal{D} have been seen, after N/B mini-batches, this terminates an *epoch*. The dataset is then shuffled and this procedure is performed again. Usually a network needs hundreds of epochs to converge. For any given iterative procedure, the updates are computed for all the network parameters by *backpropagation* (Hecht-Nielsen et al., 1988), which follows from applying the chain rule of calculus. Common policies are Gradient Descent (GD) (Rumelhart et al., 1988) being the simplest application of backpropagation, Nesterov Momentum (Bengio et al., 2013) that uses the last performed updates in order to accelerate convergence and finally more complex adaptive methods with internal hyper-parameters updated based on the weights/updates statistics such as Adam (Kingma & Ba, 2014), Adadelta (Zeiler, 2012), Adagrad (Duchi et al., 2011), RMSprop (Tieleman & Hinton, 2012), etc.

We now precisely describe the convolution, pooling, skip-connection, and recurrent operators one can use to create the mapping f_Θ . For the activation operator, there are no more details to add beyond the description in the main text.

B.1. Convolution operator

A *convolution operator* is defined as

$$f_C^{(\ell)}(\mathbf{z}^{(\ell-1)}(\mathbf{x})) = \mathbf{C}^{(\ell)}\mathbf{z}^{(\ell-1)}(\mathbf{x}) + b_C^{(\ell)} \quad (51)$$

where a special structure is defined on $\mathbf{C}^{(\ell)}$ so that it performs multi-channel convolutions on the vector $\mathbf{z}^{(\ell-1)}(\mathbf{x})$.

To highlight this fact, we first remind the reader of the multi-channel convolution operation performed on the unflatenned input $z^{(\ell-1)}(\mathbf{x})$ of shape $(C^{(\ell-1)}, I^{(\ell-1)}, J^{(\ell-1)})$ given a filter bank we denoted here as $\psi^{(\ell)}$ composed of $C^{(\ell)}$ filters, each of which is a 3D tensor of shape $(C^{(\ell-1)}, I_C^{(\ell)}, J_C^{(\ell)})$. Hence $C^{(\ell-1)}$ represents the filters depth which equals to the number of channels of the input, and $(I_C^{(\ell)}, J_C^{(\ell)})$ the spatial size of the filters. The application of the linear filters $\psi^{(\ell)}$ on the signal

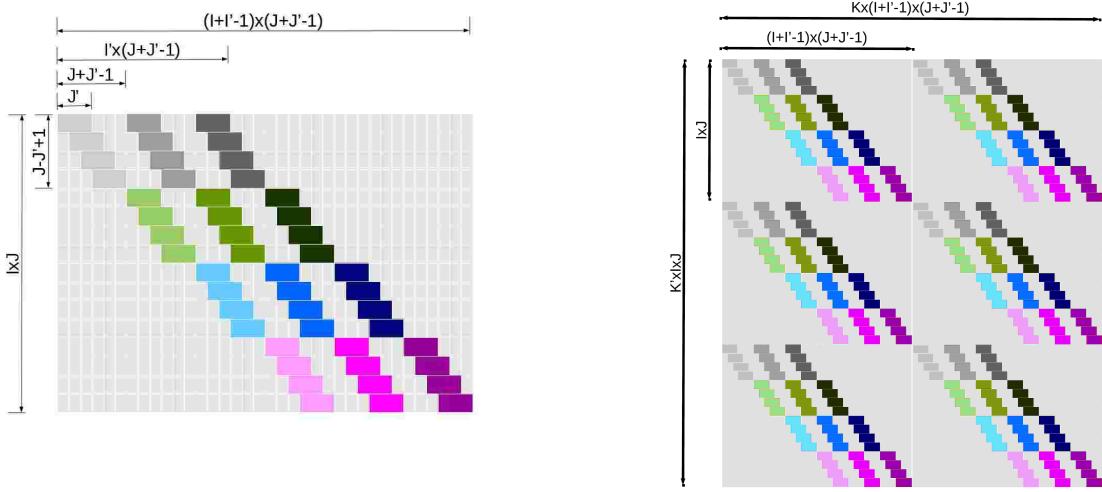


Figure 15. Left: depiction of one convolution matrix $\psi_{k,l}$. Right: depiction of the super convolution matrix C .

forms another multi-channel signal as

$$\begin{aligned} [\psi^{(\ell)} \star z^{(\ell-1)}(x)]_{c,i,j} &= \sum_{k=1}^{C^{(\ell-1)}} [\psi^{(\ell)}]_{c,k,\cdot,\cdot} \star [z^{(\ell-1)}(x)]_{k,\cdot,\cdot} \\ &= \sum_{k=1}^{C^{(\ell-1)}} \sum_{m=1}^{I_C^{(\ell)}} \sum_{n=1}^{J_C^{(\ell)}} [\psi^{(\ell)}]_{c,k,m,n} [z^{(\ell-1)}(x)]_{k,i-m,j-n}, \end{aligned} \quad (52)$$

where the output of this convolution contains $C^{(\ell)}$ channels, the number of filters in $\psi^{(\ell)}$. Then a bias term is added to each output channel, shared across spatial positions. We denote this bias term as $\xi \in \mathbb{R}^{C^{(\ell)}}$. As a result, to create channel c of the output, we first perform a 2D convolution of each channel $k = 1, \dots, C^{(\ell-1)}$ of the input with the impulse response $[\psi^{(\ell)}]_{c,k,\cdot,\cdot}$, then sum those outputs element-wise over k and finally add the bias, leading to $z^{(\ell)}(x)$ as

$$[z^{(\ell)}(x)]_{c,\cdot,\cdot} = \sum_{k=1}^{C^{(\ell-1)}} \left([\psi^{(\ell)}]_{c,k,\cdot,\cdot} \star [z^{(\ell-1)}(x)]_{k,\cdot,\cdot} \right) + \xi_c. \quad (53)$$

In general, the input is first transformed in order to apply some boundary conditions such as zero-padding, symmetric or mirror. Those are standard padding techniques in signal processing (Mallat, 1999). We now describe how to obtain the matrix $C^{(\ell)}$ and vector $b_C^{(\ell)}$ corresponding to the operations of Eq. 53 but applied on the flattened input $z^{(\ell-1)}(\mathbf{x})$ to produce the output vector $z^{(\ell)}(\mathbf{x})$. The matrix $C^{(\ell)}$ is obtained by replicating the filter weights $[\psi^{(\ell)}]_{c,k,\cdot,\cdot}$ into the circulant-block-circulent matrices $\Psi_{c,k}^{(\ell)}$, $c = 1, \dots, C^{(\ell)}$, $k = 1, \dots, C^{(\ell-1)}$ (Jayaraman et al., 2009) and stacking them into the super-matrix $C^{(\ell)}$

$$C^{(\ell)} = \begin{bmatrix} \Psi_{1,1}^{(\ell)} & \Psi_{1,2}^{(\ell)} & \dots & \Psi_{1,C^{(\ell-1)}}^{(\ell)} \\ \Psi_{2,1}^{(\ell)} & \Psi_{2,2}^{(\ell)} & \dots & \Psi_{2,C^{(\ell-1)}}^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{C^{(\ell)},1}^{(\ell)} & \Psi_{C^{(\ell)},2}^{(\ell)} & \dots & \Psi_{C^{(\ell)},C^{(\ell-1)}}^{(\ell)} \end{bmatrix}. \quad (54)$$

We provide an example in Figure 15 for $\Psi_{c,k}^{(\ell)}$ and $C^{(\ell)}$.

By sharing the bias across spatial positions, the bias term $b_C^{(\ell)}$ inherits a specific structure. It is defined by replicating $\xi_c^{(\ell)}$ on all spatial position of each output channel c .

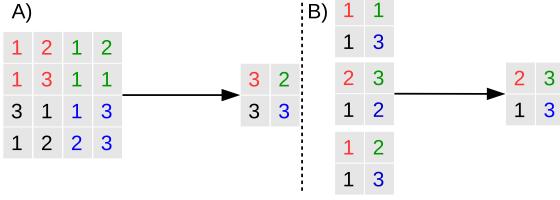


Figure 16. Depiction of A) spatial max-pooling (2, 2) with distinct regions per color, on each of them the max-pooling is applied, this is done here on an input of shape (1, 4, 4), B) channel pooling (3, 1, 1) done across channels (here 3) for each spatial position with same color coding applied on an input of shape (3, 2, 2).

B.2. Pooling operator

A *pooling* operator is a sub-sampling operation on its input according to a sub-sampling policy ρ and a collection of regions on which ρ is applied. We denote each region to be sub-sampled by $\mathcal{R}_d, d = 1, \dots, D^{(\ell)}$ where $D^{(\ell)}$ is the total number of pooling regions. Each region contains the set of indices on which the pooling policy is applied, leading to

$$\left[f_{\rho}^{(\ell)} \left(\mathbf{z}^{(\ell-1)}(\mathbf{x}) \right) \right]_k = \rho \left([\mathbf{z}^{(\ell-1)}(\mathbf{x})]_{\mathcal{R}_k} \right), \quad k = 1, \dots, D^{(\ell)} \quad (55)$$

where ρ is the pooling operator and $[\mathbf{z}^{(\ell-1)}(\mathbf{x})]_{\mathcal{R}_k} = \{[\mathbf{z}^{(\ell-1)}(\mathbf{x})]_d, d \in \mathcal{R}_k\}$. Usually one uses mean or max pooling defined as

- Max-Pooling: $\rho_{max}([\mathbf{z}^{(\ell-1)}(\mathbf{x})]_{\mathcal{R}_k}) = \max_{i \in \mathcal{R}_k} [\mathbf{z}^{(\ell-1)}(\mathbf{x})]_i$,
- Mean-Pooling: $\rho_{mean}([\mathbf{z}^{(\ell-1)}(\mathbf{x})]_{\mathcal{R}_k}) = \frac{1}{Card(\mathcal{R}_k)} \sum_{i \in \mathcal{R}_k} [\mathbf{z}^{(\ell-1)}(\mathbf{x})]_i$.

The regions \mathcal{R}_k can be of different cardinality $\exists d_1, d_2 | Card(\mathcal{R}_{d_1}) \neq Card(\mathcal{R}_{d_2})$ and can be overlapping $\exists d_1, d_2 | Card(\mathcal{R}_{d_1}) \cap Card(\mathcal{R}_{d_2}) \neq \emptyset$. However, in order to treat all input dimension, it is natural to require that each input dimension belongs to at least one region: $\forall k \in \{1, \dots, D^{(\ell-1)}\}, \exists d \in \{1, \dots, D^{(\ell)}\} | k \in \mathcal{R}_d$. The benefits of a pooling operator are three-fold. First, by reducing the output dimension, it allows for faster computation and less memory requirement. Second, it allows to greatly reduce the redundancy of information present in the input $\mathbf{z}^{(\ell-1)}(\mathbf{x})$. In fact, sub-sampling, even though linear, is common in signal processing after filter convolutions. Finally, in case of max-pooling, it allows to only backpropagate gradients through the pooled coefficient enforcing specialization of the neurons. The latter is the core of the winner-take-all strategy stating that each neuron specializes in what it performs best.

B.3. Skip connection

A *skip-connection* operator can be considered as a bypass connection added between the input of an operator and its output. Hence, it allows for the input of an operator such as a convolutional operator or FC-operator to be linearly combined with its own output. The added connections lead to better training stability and overall performances because there always exists a direct linear link from the input to all inner operators. Simply written, given an operator $f_{\theta^{(\ell)}}^{(\ell)}$ and its input $\mathbf{z}^{(\ell-1)}(\mathbf{x})$, the skip-connection operator is defined as

$$f_{\text{skip}}^{(\ell)} \left(\mathbf{z}^{(\ell-1)}(\mathbf{x}); f_{\theta^{(\ell)}}^{(\ell)} \right) = W_{\text{skip}}^{(\ell)} \mathbf{z}^{(\ell-1)}(\mathbf{x}) + f_{\theta^{(\ell)}}^{(\ell)} \left(\mathbf{z}^{(\ell-1)}(\mathbf{x}) \right) + b_{\text{skip}}^{(\ell)} \quad (56)$$

where the linear connection can be cast into a convolutional one by replacing the variables.

B.4. Recurrent operator

A *recurrent operator* which aims to act on time-series. It is defined as a recursive application along time $t = 1, \dots, T$ by transforming the input as well as using its previous output. The most simple form of this operator is a fully recurrent

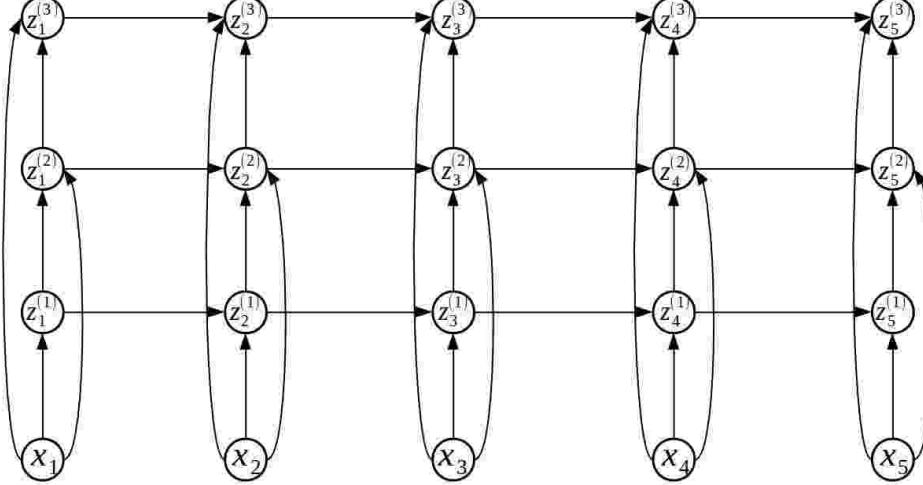


Figure 17. Depiction of a simple recursive neural network (RNN) with 3 layers. Connections highlight input-output dependencies.

operator defined as

$$\mathbf{z}^{(1,t)}(x) = \sigma \left(W^{(in,h_1)} x^t + W^{(h_1,h_1)} \mathbf{z}^{(1,t-1)}(x) + b^{(1)} \right) \quad (57)$$

$$\mathbf{z}^{(\ell,t)}(x) = \sigma \left(W^{(in,h_\ell)} x^t + W^{(h_{\ell-1},h_\ell)} \mathbf{z}^{(\ell-1,t)}(x) + W^{(h_\ell,h_\ell)} \mathbf{z}^{(\ell,t-1)}(x) + b^{(\ell)} \right) \quad (58)$$

while some applications use recurrent operators on images by considering the series of ordered local patches as a time series, the main application resides in sequence generation and analysis especially with more complex topologies such as LSTM (Graves & Schmidhuber, 2005) and GRU (Chung et al., 2014) networks. We depict the topology example in Figure 17.

C. Details on Spline Operators

C.1. Spline functions

Definition 2. Let $\Omega = \{\omega_1, \dots, \omega_R\}$ be a partition of \mathbb{R}^D and let $\Phi = \{\phi_1, \dots, \phi_R\}$ be a collection of local mappings with $\phi_{\omega_r} : \mathbb{R}^D \rightarrow \mathbb{R}$, Then we define multivariate spline with partition Ω and local mappings Φ the following

$$s[\Phi, \Omega](x) = \sum_{r=1}^R \phi_r(x) \mathbb{1}_{\{x \in \omega_r\}} \quad (59)$$

$$= \phix, \quad (60)$$

where the input dependent selection is abbreviated via

$$\phi[x] := \phi_r \text{ s.t. } x \in \omega_r. \quad (61)$$

If the local mappings ϕ_r are affine we have $\phi_r(x) = \langle [\alpha]_{r,.}, x \rangle + [\beta]_r$ with $\alpha \in \mathbb{R}^{R \times D}$ and $\beta \in \mathbb{R}$. We denote this functional as a multivariate affine spline:

$$s[\alpha, \beta, \Omega](x) = \sum_{r=1}^R (\langle [\alpha]_{r,.}, x \rangle + [\beta]_r) \mathbb{1}_{\{x \in \omega_r\}} \quad (62)$$

$$= \langle \alpha[x], x \rangle + \beta[x]. \quad (63)$$

C.2. Spline operators

A natural extension of spline functions is the **spline operator** (SO) that we denote $S : \mathbb{R}^D \rightarrow \mathbb{R}^K, K > 1$. We present here a general definition and propose in the next section an intuitive way to construct spline operators via a collection of multivariate splines, the special case of current DNs.

Definition 3. A spline operator is a mapping $S : \mathbb{R}^D \rightarrow \mathbb{R}^K$ defined by a collection of local mappings $\Phi^S = \{\phi_r^S : \mathbb{R}^D \rightarrow \mathbb{R}^K, r = 1, \dots, R\}$ associated with a partition of \mathbb{R}^D denoted as $\Omega^S = \{\omega_r^S, r = 1, \dots, R\}$ s.t.

$$\begin{aligned} S[\Phi^S, \Omega^S](x) &= \sum_{r=1}^R \phi_r^S(x) \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= \phi^Sx, \end{aligned}$$

where we denoted the region specific mapping associated to the input x by $\phi^S[x]$.

Definition 4. A special case occurs when the mappings ϕ_r^S are affine. We thus define in this case the **affine spline operator** (ASO) which will play an important role for DN analysis. In this case, $\phi_r^S(x) = [A]_{.,r,.}x + [B]_{.,r}$, with $A \in \mathbb{R}^{K \times R \times D}, B \in \mathbb{R}^{K \times R}$. As a result, a ASO can be rewritten as

$$\begin{aligned} S[A, B, \Omega^S](x) &= \sum_{r=1}^R ([A]_{.,r,.}x + [B]_{.,r}) \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= A[x]x + b[x]. \end{aligned}$$

Such operators can be defined via a collection of multivariate splines. Given K multivariate spline functions $s[\phi(k), \Omega(k)] : \mathbb{R}^D \rightarrow \mathbb{R}, k = 1, \dots, K$, their respective output is "stacked" to produce an output vector of dimension K . The internal parameters of each of the K multivariate spline are $\Omega(k)$, a partition of \mathbb{R}^D with $\text{Card}(\Omega(k)) = R_k$ and $\Phi(k) = \{\phi(k)_1, \dots, \phi(k)_{R_k}\}$. Stacking their respective output to form an output vector leads to the induced spline operator $S[(s[\Phi(k), \Omega(k)])_{k=1}^K]$.

Definition 5. The spline operator $S : \mathbb{R}^D \rightarrow \mathbb{R}^K$ defined with K multivariate splines $(s[\Phi(k), \Omega(k)])_{k=1}^K$ with $s[\Phi(k), \Omega(k)] : \mathbb{R}^D \rightarrow \mathbb{R}$ is defined as

$$S[(s[\Phi(k), \Omega(k)])_{k=1}^K](x) = \begin{bmatrix} s[\Phi(1), \Omega(1)](x) \\ \vdots \\ s[\Phi(K), \Omega(K)](x) \end{bmatrix}. \quad (64)$$

The use of K multivariate splines to construct a SO does not provide directly the explicit collection of mappings and regions Φ^S, Ω^S . Yet, it is clear that the SO is jointly governed by all the individual multivariate splines. Let us first present some intuitions on this fact. The spline operator output is computed with each of the K splines having activated a region specific functional depending on their own input space partitioning. In particular, each of the region ω_r^S of the input space leading to a specific joint configuration ϕ_r^S is the one of interest, leading to Ω^S and Φ^S . We can thus write explicitly the new regions of the spline operator based on the ensemble of partition of all the involved multivariate splines as

$$\Omega^S = \left(\bigcup_{(\omega(1), \dots, \omega(K)) \in \Omega(1) \times \dots \times \Omega(K)} \left\{ \bigcap_{k \in \{1, \dots, K\}} \omega(k) \right\} \right) \setminus \{\emptyset\}. \quad (65)$$

We can denote the number of region associated to this SO as $R^S = \text{Card}(\Omega^S)$. From this, the local mappings of the SO ϕ_r^S correspond to the joint mappings of the splines being activated on ω_r^S :

$$\phi^S[\omega_r^S](x) = \begin{bmatrix} \phi(1)[\omega_r^S](x) \\ \vdots \\ \phi(K)[\omega_r^S](x) \end{bmatrix}, \quad (66)$$

with $\phi(k)[\omega_r^S] = \phi(k)_q \in \Phi_k$ s.t. $\omega_r^S \subset \omega(k)_q$. In fact, for each region ω_r^S of the SO there is a unique region $\omega(k)_q$ for each of the splines $k = 1, \dots, K$, such that it is a subset as $\omega_r^S \subset \omega(k)_q$ and it is disjoint to all others $\omega_r^S \cap \omega(k)_l = \emptyset, \forall l \neq q$. In other word we have the following property:

$$\forall(r, k) \in \{1, \dots, R^S\} \times \{1, \dots, K\}, \exists! q_k \in \{1, \dots, R_k\} \text{ s.t. } \omega_r^S \cap \omega(k)_l = \begin{cases} \omega_r^S, & l = q_k \\ \emptyset, & l \neq q_k \end{cases} \quad (67)$$

as we remind $\omega_r^S \cap \omega(k)_l = \omega_r^S \iff \omega_r^S \subset \omega(k)_{q_k}$.

This leads to the following SO formulation

$$\begin{aligned} S \left[(s[\Phi(k), \Omega(k)])_{k=1}^K \right] (x) &= \sum_{r=1}^{R^S} \begin{bmatrix} s[\Phi(1), \Omega(1)](x) \\ \vdots \\ s[\Phi(K), \Omega(K)](x) \end{bmatrix} \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= \sum_{r=1}^{R^S} \begin{bmatrix} \phi(1)[\omega_r^S](x) \\ \vdots \\ \phi(K)[\omega_r^S](x) \end{bmatrix} \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= \sum_{r=1}^{R^S} \phi^S[\omega_r^S](x) \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= \phi^Sx. \end{aligned} \quad (68)$$

We can now study the case of affine splines leading to ASOs. The affine property allows notation simplifications. It is defined as

$$\begin{aligned} S \left[(s[\alpha(k), \beta(k), \Omega(k)])_{k=1}^K \right] (x) &= \sum_{r=1}^{R^S} \begin{bmatrix} s[\alpha(1), \beta(1), \Omega(1)](x) \\ \vdots \\ s[\alpha(K), \beta(K), \Omega(K)](x) \end{bmatrix} \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= \sum_{r=1}^{R^S} \begin{bmatrix} \langle \alpha(1)[\omega_r^S], x \rangle + b(K)[\omega_r^S] \\ \vdots \\ \langle \alpha(1)[\omega_r^S], x \rangle + b(K)[\omega_r^S] \end{bmatrix} \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= \sum_{r=1}^{R^S} \left(\begin{bmatrix} \alpha(1)[\omega_r^S]^T \\ \vdots \\ \alpha(K)[\omega_r^S]^T \end{bmatrix} x + \begin{bmatrix} \beta(1)[\omega_r^S] \\ \vdots \\ \beta(K)[\omega_r^S] \end{bmatrix} \right) \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= \sum_{r=1}^{R^S} (A[\omega_r^S]x + B[\omega_r^S]) \mathbb{1}_{\{x \in \omega_r^S\}} \\ &= A[x]x + B[x]. \end{aligned} \quad (69)$$

The collection of matrices and biases and the partitions completely define an ASO.

D. Details on Deep Networks as Spline Operators

D.1. Deep Recurrent Networks

We can derive one step of a standard fully recurrent neural network (Graves, 2013) as

$$\begin{aligned} \mathbf{z}_{RNN}^{(1,t)} &= A_\sigma^{(1,t)} [W_{in}^{(1)}x^t + W_{rec}^{(1)}\mathbf{z}_{RNN}^{(1,t-1)} + b^{(1)}] (W_{in}^{(1)}x^t + W_{rec}^{(1)}\mathbf{z}_{RNN}^{(1,t-1)} + b^{(1)}) + b_\sigma^{(1,t)}, \\ \mathbf{z}_{RNN}^{(\ell,t)} &= A_\sigma^{(\ell,t)} [W_{in}^{(\ell)}x^t + W_{rec}^{(\ell)}\mathbf{z}_{RNN}^{(\ell,t-1)} + W_{up}^{(\ell)}\mathbf{z}_{RNN}^{(\ell-1,t)} + b^{(\ell)}] (W_{in}^{(\ell)}x^t + W_{rec}^{(\ell)}\mathbf{z}_{RNN}^{(\ell,t-1)} \\ &\quad + W_{up}^{(\ell)}\mathbf{z}_{RNN}^{(\ell-1,t)} + b^{(\ell)}) + b_\sigma^{(\ell,t)}, \quad \ell > 1. \end{aligned}$$

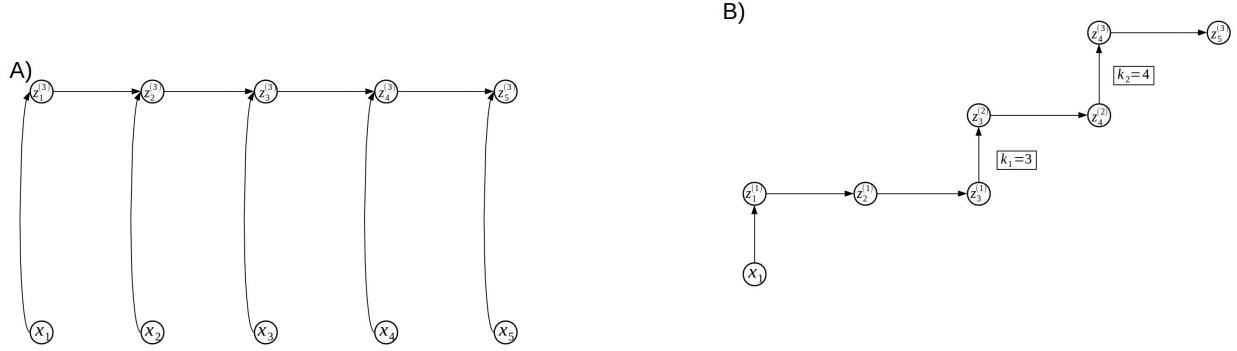


Figure 18. At any given layer of an RNN, there always exists a direct input to a representation path and recursion.

Figure 19. In addition of all the possible input paths going through the hidden layers there are a combinatorially large number of paths, each one fully determined by the succession of forward in time or upward in layer successions.

By the double recursion of the formula (in time and in depth) we first proceed by writing the time unrolled RNN mapping as

$$\begin{aligned}
z_{RNN}^{(1,T)}(x) &= \sum_{t=T}^1 \left(\prod_{k=T}^{t+1} A_\sigma^{(1,k)} W_{rec}^{(1)} \right) A_\sigma^{(1,t)} (W_{in}^{(1)} x^t + b_\sigma^{(1,t)} + A_\sigma^{(1,t)} b^{(1)}) \\
&= \sum_{t=T}^1 \left(\prod_{k=T}^{t+1} A_\sigma^{(1,k)} W_{rec}^{(1)} \right) A_\sigma^{(1,t)} W_{in}^{(1)} x_t + \sum_{t=T}^1 \left(\prod_{k=T}^{t+1} A_\sigma^{(1,k)} W_{rec}^{(1)} \right) [b_\sigma^{(1,t)} + A_\sigma^{(1,t)} b^{(1)}] \\
z_{RNN}^{(\ell,T)}(x) &= \sum_{t=T}^1 \left(\prod_{k=T}^{t+1} A_\sigma^{(\ell,k)} W_{rec}^{(\ell)} \right) A_\sigma^{(\ell,t)} W_{in}^{(\ell)} x^t \\
&\quad + \sum_{t=T}^1 \left(\prod_{k=T}^{t+1} A_\sigma^{(\ell,k)} W_{rec}^{(\ell)} \right) [b_\sigma^{(\ell,t)} + A_\sigma^{(\ell,t)} b^{(\ell)} + A_\sigma^{(\ell,t)} W_{up}^{(\ell)} z_{RNN}^{(\ell-1,t)}(x)], \quad \ell > 1.
\end{aligned}$$

The formula unrolled in time is still recursive in depth. While the exact unrolled version would be cumbersome for any layer ℓ , we propose a simple way to find the analytical formula based on the possible paths an input can take till the final time representation of layer ℓ . To do so, see Figures 18, 19.

Hence we can decompose all those paths by blocks of forward interleave with upward paths. With this, we can see that the possible paths are all the paths from the input to the final nodes, and they can not go back in time nor down in layers. Hence, they are all the possible combinations for forward in time or upward in depth. We can thus find the exact output formula

below for RNN where we focus on the template matching part for clarity (thus omitting the bias terms from the formula):

$$\mathbf{z}_{RNN}^{(2,T)}(x) = \sum_{t=1}^T \sum_{k_1=t}^T \left(\prod_{q=T}^{k_1+1} A_\sigma^{(2,q)} W_{rec}^{(2)} \right) A_\sigma^{(2,k_1)} W_{up}^{(2)} \left(\prod_{q=t+1}^{k_1-1} A_\sigma^{(1,q)} W_{rec}^{(1)} \right) A_\sigma^{(1,t)} W_{in}^{(1)} x^t \quad (70)$$

$$\begin{aligned} \mathbf{z}_{RNN}^{(3,T)}(x) &= \sum_{t=1}^T \sum_{k_1=t}^T \sum_{k_2 \geq k_1}^T \left(\prod_{q=T}^{k_2+1} A_\sigma^{(3,q)} W_{rec}^{(3)} \right) A_\sigma^{(3,k_2)} W_{up}^{(3)} \left(\prod_{q=k_1}^{k_2-1} A_\sigma^{(2,q)} W_{rec}^{(2)} \right) A_\sigma^{(2,k_1)} W_{up}^{(2)} \\ &\quad \left(\prod_{q=t+1}^{k_1-1} A_\sigma^{(1,q)} W_{rec}^{(1)} \right) A_\sigma^{(1,t)} W_{in}^{(1)} x^t \end{aligned} \quad (71)$$

$$\begin{aligned} \mathbf{z}_{RNN}^{(4,T)}(x) &= \sum_{t=1}^T \sum_{k_1=t}^T \sum_{k_2 \geq k_1}^T \sum_{k_3 \geq k_2}^T \left(\prod_{q=T}^{k_3+1} A_\sigma^{(4,q)} W_{rec}^{(4)} \right) A_\sigma^{(4,k_3)} W_{up}^{(4)} \left(\prod_{q=T}^{k_2+1} A_\sigma^{(3,q)} W_{rec}^{(3)} \right) A_\sigma^{(3,k_2)} W_{up}^{(3)} \\ &\quad \left(\prod_{q=k_1}^{k_2-1} A_\sigma^{(2,q)} W_{rec}^{(2)} \right) A_\sigma^{(2,k_1)} W_{up}^{(2)} \left(\prod_{q=t+1}^{k_1-1} A_\sigma^{(1,q)} W_{rec}^{(1)} \right) A_\sigma^{(1,t)} W_{in}^{(1)} x^t \end{aligned} \quad (72)$$

...

E. Proofs

The proofs of any propositions and theorems not provided below are trivial.

E.1. Proofs of Propositions 1–5

Those propositions are trivial by definition and uniqueness of a piecewise affine convex mapping and their link to the rewritten operators. Also, see Sec. E.2 as well as (Boyd & Vandenberghe, 2004) for the convex properties and conditions.

E.2. Proof of Proposition 6 (DN layers are MASOs)

We now provide examples to better illustrate the DN layer reparametrization as a composition of DN operators.

Affine Transform+Activation: Any MASO following an affine transformation can be computed through the definition of a new reparametrized MASO with

$$\begin{aligned} S[A_\sigma^{(\ell)}, b_\sigma^{(\ell)}](W^{(\ell)}\mathbf{x} + b_W^{(\ell)}) &= \begin{bmatrix} \max_{r=1,\dots,R_1} [A_\sigma^{(\ell)}]_{1,r}^T (W^{(\ell)}\mathbf{x} + b_W^{(\ell)}) + [b_\sigma^{(\ell)}]_{1,r} \\ \dots \\ \max_{r=1,\dots,R_K} [A_\sigma^{(\ell)}]_{K,r}^T (W^{(\ell)}\mathbf{x} + b_W^{(\ell)}) + [b_\sigma^{(\ell)}]_{K,r} \end{bmatrix} \\ &= \begin{bmatrix} \max_{r=1,\dots,R_1} (W^{(\ell)T} [A_\sigma^{(\ell)}]_{1,r})^T \mathbf{x} + [A_\sigma^{(\ell)}]_{1,r}^T b_W^{(\ell)} + [b_\sigma^{(\ell)}]_{1,r} \\ \dots \\ \max_{r=1,\dots,R_K} (W^{(\ell)T} [A_\sigma^{(\ell)}]_{K,r})^T \mathbf{x} + [A_\sigma^{(\ell)}]_{K,r}^T b_W^{(\ell)} + [b_\sigma^{(\ell)}]_{K,r} \end{bmatrix} \\ &= S[A, b](\mathbf{x}), \end{aligned} \quad (73)$$

with $[A^{(\ell)}]_{k,r,\cdot} = W^{(\ell)T} [A_\sigma^{(\ell)}]_{k,r,\cdot}$ and $[b^{(\ell)}]_{k,r} = [b_\sigma^{(\ell)}]_{k,r} + \langle b_W^{(\ell)}, [A_\sigma^{(\ell)}]_{k,r,\cdot} \rangle$

Linear skip-connection: Any MASO with a skip-connection can be computed via definition of a new reparametrized MASO with

$$S[A^{(\ell)}, b^{(\ell)}](\mathbf{x}) + \mathbf{x} = \begin{bmatrix} \max_{r=1, \dots, R_1} [A^{(\ell)}]_{1,r,\cdot}^T \mathbf{x} + [b^{(\ell)}]_{1,r} \\ \dots \\ \max_{r=1, \dots, R_K} [A^{(\ell)}]_{K,r,\cdot}^T \mathbf{x} + [b^{(\ell)}]_{K,r} \end{bmatrix} + \mathbf{x} \quad (74)$$

$$= \begin{bmatrix} \max_{r=1, \dots, R_1} ([A^{(\ell)}]_{1,r,\cdot} + \mathbf{e}_1)^T \mathbf{x} + [b^{(\ell)}]_{1,r} \\ \dots \\ \max_{r=1, \dots, R_K} ([A^{(\ell)}]_{K,r,\cdot} + \mathbf{e}_D)^T \mathbf{x} + [b^{(\ell)}]_{K,r} \end{bmatrix} \quad (75)$$

$$= S[A', b'](\mathbf{x}) \quad (76)$$

$$(77)$$

with $[A']_{k,r,\cdot} = [A^{(\ell)}]_{k,r,\cdot} + \mathbf{e}_r$ and $[b']_{k,r} = [b^{(\ell)}]_{k,r}$

ResNet layer: Any ResNet layer can be computed via definition of a new reparametrized MASO with

$$\begin{aligned} S[A_\sigma^{(\ell)}, b_\sigma^{(\ell)}](\mathbf{C}\mathbf{x} + b_C) + \mathbf{C}_{\text{skip}}\mathbf{x} + b_{\text{skip}} &= \begin{bmatrix} [\mathbf{C}_{\text{skip}}]_1^T \mathbf{x} + [b_{\text{skip}}]_1 + \max_{r=1, \dots, R_1} [A_\sigma^{(\ell)}]_{1,r,\cdot}^T (\mathbf{C}\mathbf{x} + b_C) + [b_\sigma^{(\ell)}]_{1,r} \\ \dots \\ [\mathbf{C}_{\text{skip}}]_K^T \mathbf{x} + [b_{\text{skip}}]_K + \max_{r=1, \dots, R_K} [A_\sigma^{(\ell)}]_{K,r,\cdot}^T (\mathbf{C}\mathbf{x} + b_C) + [b_\sigma^{(\ell)}]_{K,r} \end{bmatrix} \\ &= \begin{bmatrix} \max_{r=1, \dots, R_1} (\mathbf{C}^T [A_\sigma^{(\ell)}]_{1,r,\cdot} + [\mathbf{C}_{\text{skip}}]_{1,\cdot})^T \mathbf{x} + [A_\sigma^{(\ell)}]_{1,r,\cdot}^T b_C + [b_\sigma^{(\ell)}]_{1,r} + [b_{\text{skip}}]_1 \\ \dots \\ \max_{r=1, \dots, R_K} (\mathbf{C}^T [A_\sigma^{(\ell)}]_{K,r,\cdot} + [\mathbf{C}_{\text{skip}}]_{K,\cdot})^T \mathbf{x} + [A_\sigma^{(\ell)}]_{K,r,\cdot}^T b_C + [b_\sigma^{(\ell)}]_{K,r} + [b_{\text{skip}}]_K \end{bmatrix} \\ &= S[A^{(\ell)}, b^{(\ell)}](\mathbf{x}) \end{aligned} \quad (78)$$

with $[A^{(\ell)}]_{k,r,\cdot} = \mathbf{C}^T [A_\sigma^{(\ell)}]_{k,r,\cdot} + [\mathbf{C}_{\text{skip}}]_{k,\cdot}$ and $[b^{(\ell)}]_{k,r} = [A_\sigma^{(\ell)}]_{k,r,\cdot}^T b_C + [b_\sigma^{(\ell)}]_{k,r} + [b_{\text{skip}}]_k$.

E.3. Proof of Proposition 8

This result is direct from the fact that all the mappings are piecewise affine. Hence the derivative of the prediction or any neuron output with respect to the output gives the slope of the projection, itself corresponding to the definition of saliency maps. This is exact for piecewise affine mappings and will be the same for any input in the same input space region. However for other DNs (e.g., sigmoid), the derivative is only exact at the point at which it is computed and corresponds to a Taylor approximation of the DN mapping.

E.4. Proof of Proposition 10

We now derive the Lipschitz constant of the softmax nonlinearity. We note that for any differentiable mapping f we have the following upper bound

$$\|f(x) - f(y)\|_2^2 \leq \max_x \|Df(x)\|_F^2 \|x - y\|_2^2, \quad (79)$$

with Df the total derivative. Thus we now analyze $\max_x \|Df(x)\|_F^2$ to find the softmax Lipschitz constant. We thus obtain by taking f the softmax nonlinearity with input $x := \mathbf{z}^{(L)}$

$$\max_{p \in \Delta_D} \|Df(p)\|_F^2 = \max_{p \in \Delta_D} \sum_{i=1}^D \sum_{j=1, j \neq i}^D p_i^2 p_j^2 + \sum_{i=1}^D p_i^2 (1 - p_i)^2 \quad (80)$$

where we used Δ_D the simplex of dimension D defined as

$$\Delta_D = \{x \in \mathbb{R}^{D+} \mid \sum_i x_i = 1\}. \quad (81)$$

We now aim at maximizing this functional, instead of maximizing w.r.t. the input $\mathbf{z}^{(L)}$ we directly maximize w.r.t. the induced density. The value of the maximum is the same. Note that we have a bijection between $\mathbf{z}^{(L)}$ and p . The Lagrangian

is the augmented loss function with the constrain as

$$\mathcal{L}(p) = \sum_{i=1}^D \sum_{j=1, j \neq i}^D p_i^2 p_j^2 + \sum_{i=1}^D p_i^2 (1 - p_i)^2 + \lambda \left(\sum_{i=1}^D p_i - 1 \right), \quad (82)$$

and with $p_i = \frac{e^{[\mathbf{z}^{(L)}]_i}}{\sum_i e^{[\mathbf{z}^{(L)}]_i}}$. We now seek the stationary points

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_k} &= 4p_k \sum_{j=1, j \neq k}^D p_j^2 + 2p_k(1 - p_k)^2 - 2(1 - p_k)p_k^2 + \lambda \\ &= 4p_k \sum_{j=1, j \neq k}^D p_j^2 + 2p_k - 4p_k^2 + 2p_k^3 - 2p_k^2 + 2p_k^3 + \lambda \\ &= 4p_k \sum_{j=1, j \neq k}^D p_j^2 + 2p_k - 6p_k^2 + 4p_k^3 + \lambda \\ &= 4p_k \sum_{j=1}^D p_j^2 + 2p_k - 6p_k^2 + \lambda, \quad \forall k \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{i=1}^D p_i - 1. \end{aligned} \quad (83)$$

We now seek to solve the system $\nabla \mathcal{L} = 0$. Since we have $\frac{\partial \mathcal{L}}{\partial p_k} = 0 \forall k$, it is clear that $\sum_{k=1}^D \frac{\partial \mathcal{L}}{\partial p_k} = 0$, leading to

$$\sum_{k=1}^D \frac{\partial \mathcal{L}}{\partial p_k} = 0 \quad (84)$$

$$\implies \sum_{k=1}^D \left(4p_k \sum_{j=1}^D p_j^2 + 2p_k - 6p_k^2 + \lambda \right) = 0 \quad (85)$$

$$\implies 4 \sum_{j=1}^D p_j^2 + 2 - 6 \sum_{j=1}^D p_j^2 + D\lambda = 0 \quad (86)$$

$$\implies D\lambda = 2 \sum_{j=1}^D p_j^2 - 2 \quad (87)$$

$$\implies \lambda = \frac{2}{D} \left(\sum_{j=1}^D p_j^2 - 1 \right). \quad (88)$$

Plugging this into the gradient of \mathcal{L} , we arrive at the updated system of equations

$$\begin{aligned} 4p_k \sum_{j=1}^D p_j^2 + 2p_k - 6p_k^2 + \frac{2}{D} \left(\sum_{j=1}^D p_j^2 - 1 \right) &= 0, \quad \forall k \\ \sum_{i=1}^D p_i &= 1 \end{aligned} \quad (89)$$

which in vector form leads to

$$\mathbf{p}(2\|\mathbf{p}\|_2^2 + 1) - 3\mathbf{p} \circ \mathbf{p} = (1 - \|\mathbf{p}\|_2^2)\text{vec}(1/D), \quad (90)$$

where \circ denotes the Hadamard product. It is clear that \mathbf{p} must be constant across its dimensions; this constraint leads to $p = \text{vec}(1/D)$. At the point we have the value of the maximum being equal to

$$\begin{aligned}
\mathcal{L}(\text{vec}(1/D)) &= \sum_{i=1}^D \sum_{j=1, j \neq i}^D \frac{1}{D^4} + \sum_{i=1}^D \frac{1}{D^2} \left(1 - \frac{1}{D}\right)^2 \\
&= \frac{D-1}{D^3} + \frac{1}{D} \left(1 - \frac{1}{D}\right)^2 \\
&= \frac{D-1}{D^3} + \frac{1}{D} - \frac{2}{D^2} + \frac{1}{D^3} \\
&= \frac{D-1+D^2-2D+1}{D^3} \\
&= \frac{D(D-1)}{D^3} \\
&= \frac{D-1}{D^2}
\end{aligned} \tag{91}$$

Note also that the softmax generalizes the sigmoid. In the case of the sigmoid (softmax with $D^{(L)} = 2$ or 2 classes), we directly obtain that the maximum of the gradient occurs at the point $p_1 = p_2 = 0.5$. \square

E.5. Proof of Proposition 12

We aim to minimize the cross-entropy loss function for a given input \mathbf{x}_n belonging to class y_n . We abbreviate $[A[\mathbf{x}_n]]_{c,.}$ as $A[\mathbf{x}_n]_c$. We also have the constraint $\sum_{c=1}^C \|A[\mathbf{x}_n]_c\|^2 \leq K$. The loss function is thus convex on a convex set. It is thus sufficient to find a extremum point. We denote the augmented loss function with the Lagrange multiplier as

$$l(A[\mathbf{x}_n]_1, \dots, A[\mathbf{x}_n]_C, \lambda) = -\langle A[\mathbf{x}_n]_{y_n}, \mathbf{x}_n \rangle + \log \left(\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle} \right) - \lambda \left(\sum_{c=1}^C \|A[\mathbf{x}_n]_c\|^2 - K \right). \tag{92}$$

The sufficient KKT conditions are thus

$$\frac{dl}{dA[\mathbf{x}_n]_1} = -1_{\{y_n=1\}} x + \frac{e^{\langle A[\mathbf{x}_n]_1, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \mathbf{x}_n - 2\lambda A[\mathbf{x}_n]_1 = 0 \tag{93}$$

\vdots

$$\frac{dl}{dA[\mathbf{x}_n]_C} = -1_{\{y_n=C\}} x + \frac{e^{\langle A[\mathbf{x}_n]_C, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} x - 2\lambda A[\mathbf{x}_n]_C = 0 \tag{94}$$

$$\frac{\partial l}{\partial \lambda} = K - \sum_{c=1}^C \|A[\mathbf{x}_n]_c\|^2 = 0. \tag{95}$$

(96)

We first proceed by identifying λ as follows

$$\left. \begin{array}{l} \frac{dl}{dA[\mathbf{x}_n]_1} = 0 \\ \vdots \\ \frac{dl}{dA[\mathbf{x}_n]_C} = 0 \end{array} \right\} \Rightarrow \sum_{c=1}^C A[\mathbf{x}_n]_c^T \frac{dl}{dA[\mathbf{x}_n]_c} = 0 \tag{97}$$

$$\Rightarrow -\langle A[\mathbf{x}_n]_{y_n}, x \rangle + \sum_{c=1}^C \frac{e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle - 2\lambda \sum_{c=1}^C \|A[\mathbf{x}_n]_c\|^2 = 0 \tag{98}$$

$$\Rightarrow \lambda = \frac{1}{2K} \left(\sum_{c=1}^C \frac{e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle - \langle A[\mathbf{x}_n]_{y_n}, \mathbf{x}_n \rangle \right). \tag{99}$$

Now we plug λ into $\frac{dl}{dA[\mathbf{x}_n]_k}$, $\forall k = 1, \dots, C$ to obtain

$$\begin{aligned} \frac{dl}{dA[\mathbf{x}_n]_k} &= -1_{\{y_n=k\}}x + \frac{e^{\langle A[\mathbf{x}_n]_k, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \mathbf{x}_n - 2\lambda A[\mathbf{x}_n]_k \\ &= -1_{\{y_n=k\}}\mathbf{x}_n + \frac{e^{\langle A[\mathbf{x}_n]_k, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \mathbf{x}_n - \frac{1}{K} \sum_{c=1}^C \frac{e^{\langle A[\mathbf{x}_n]_c, x \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle A[\mathbf{x}_n]_k \\ &\quad + \frac{1}{K} \langle A[\mathbf{x}_n]_{y_n}, \mathbf{x}_n \rangle A[\mathbf{x}_n]_k. \end{aligned} \tag{100}$$

We now leverage the fact that $A[\mathbf{x}_n]_i = A[\mathbf{x}_n]_j, \forall i, j \neq y_n$ to simplify the notation to

$$\begin{aligned} \frac{dl}{dA[\mathbf{x}_n]_k} &= \left(\frac{e^{\langle A[\mathbf{x}_n]_k, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, x \rangle}} - 1_{\{k=y_n\}} \right) \mathbf{x}_n - \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle A[\mathbf{x}_n]_k \\ &\quad + \frac{1}{K} \left(1 - \frac{e^{\langle A[\mathbf{x}_n]_{y_n}, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \right) \langle A[\mathbf{x}_n]_{y_n}, \mathbf{x}_n \rangle A[\mathbf{x}_n]_k \\ &= \left(\frac{e^{\langle A[\mathbf{x}_n]_k, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} - 1_{\{k=y_n\}} \right) \mathbf{x}_n - \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, x \rangle}} \langle A[\mathbf{x}_n]_i, x \rangle A[\mathbf{x}_n]_k \\ &\quad + \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \langle A[\mathbf{x}_n]_{y_n}, \mathbf{x}_n \rangle A[\mathbf{x}_n]_k \\ &= \left(\frac{e^{\langle A[\mathbf{x}_n]_k, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, x \rangle}} - 1_{\{k=y_n\}} \right) \mathbf{x}_n + \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \langle A[\mathbf{x}_n]_{y_n} - A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle A[\mathbf{x}_n]_k. \end{aligned} \tag{101}$$

We now proceed by using the proposed optimal solutions $A^*[\mathbf{x}_n]_c, c = 1, \dots, C$ and demonstrate that they lead to an extremum point that, by nature of the problem, is the global optimum. We denote by i any index different from y_n . We first consider the case $k = y_n$:

$$\begin{aligned} \frac{dl}{dA[\mathbf{x}_n]_{y_n}} &= - \left(1 - \frac{e^{\langle A[\mathbf{x}_n]_{y_n}, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, x \rangle}} \right) x + \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, x \rangle}} \langle A[\mathbf{x}_n]_{y_n} - A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle A[\mathbf{x}_n]_{y_n} \\ &= -(C-1) \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} x + \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, x \rangle}} \langle A[\mathbf{x}_n]_{y_n} - A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle A[\mathbf{x}_n]_{y_n} \\ &= \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, x \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} (-K\mathbf{x}_n + \langle A[\mathbf{x}_n]_{y_n} - A[\mathbf{x}_n]_i, x \rangle A[\mathbf{x}_n]_{y_n}) \\ &= \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, x \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \left(-K\mathbf{x}_n + \sqrt{\frac{C-1}{C}} K\mathbf{x}_n + \sqrt{\frac{K}{C(C-1)}} \mathbf{x}_n, \mathbf{x}_n \right) \sqrt{\frac{C-1}{C}} K\mathbf{x}_n \\ &= \frac{(C-1)}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} (-K\mathbf{x}_n + K\|\mathbf{x}_n\|^2 \mathbf{x}_n) \\ &= 0. \end{aligned} \tag{102}$$

Now, we consider the other cases $k \neq y_n$:

$$\frac{dl}{dA[\mathbf{x}_n]_i} = \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \mathbf{x}_n + \frac{C-1}{K} \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \langle A[\mathbf{x}_n]_{y_n} - A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle A[\mathbf{x}_n]_i \quad (103)$$

$$= \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x}_n \rangle}} \left(\mathbf{x}_n + \frac{C-1}{K} \langle A[\mathbf{x}_n]_{y_n} - A[\mathbf{x}_n]_i, \mathbf{x}_n \rangle A[\mathbf{x}_n]_i \right) \quad (104)$$

$$= \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x} \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x} \rangle}} \left(\mathbf{x} - \frac{C-1}{K} \langle \sqrt{\frac{C-1}{C}} K \mathbf{x}_n + \sqrt{\frac{K}{C(C-1)}} \mathbf{x}_n, \mathbf{x} \rangle \sqrt{\frac{K}{C(C-1)}} \mathbf{x}_n \right) \quad (105)$$

$$= \frac{e^{\langle A[\mathbf{x}_n]_i, \mathbf{x} \rangle}}{\sum_{c=1}^C e^{\langle A[\mathbf{x}_n]_c, \mathbf{x} \rangle}} (\mathbf{x}_n - \|\mathbf{x}_n\|^2 \mathbf{x}_n) \quad (106)$$

$$= 0. \quad (107)$$

This analysis is per-example, but it is clear that minimizing the per-example loss minimizes the global (dataset) loss. \square

E.6. Proof of Theorems 1,2

The proofs follow easily by leveraging the definition of the nonlinearity coupled with the results from Section E.2 and (Boyd & Vandenberghe, 2004).

E.7. Proof of Theorem 3

The proof leverages a result from (Breiman, 1993) that is similar but targeted at a functional f . Thanks to the independence of the K max-affine splines being used internally to construct the 2 MASOs, the exact same proof can be used here applied independently on each of the output $[f(\mathbf{x})]_k$ as

$$\begin{aligned} \|f(\mathbf{x}) - f_\Theta(\mathbf{x})\|_2^2 &= \sum_k ([f(\mathbf{x})]_k - [f_\Theta(\mathbf{x})]_k)^2 \\ &\leq \frac{\sum_k \eta_k}{R} \end{aligned} \quad (108)$$

as the same number of regions is present for each of the max-affine spline from a MASO. \square

E.8. Proof that Activation Optimization Leads to MASO DNs

Proof. (Unser, 2018) model a DN as

$$\underline{z}^{(\ell)}(\mathbf{x}) = \left(\underline{f}_\sigma^{(\ell)} \circ \underline{f}_W^{(\ell)} \circ \cdots \circ \underline{f}_\sigma^{(\ell)} \circ \underline{f}_W^{(\ell)} \right) (\mathbf{x}), \quad (109)$$

with optimal nonlinearities for each operator and dimension defined as

$$\underline{\sigma}_k^{(\ell)}(u) = b_{1,k,\ell} + b_{2,k,\ell} u + \sum_{r=1}^R a_{r,k,\ell} \max(u - \tau_{r,k,\ell}, 0). \quad (110)$$

leading to a formulation in which $\underline{z}^{(\ell)}(\mathbf{x}) = \underline{f}_\sigma^{(\ell)}(\underline{f}_W^{(\ell)}(\underline{z}^{(\ell-1)}(\mathbf{x})))$, $\underline{f}_W^{(\ell)} : \mathbb{R}^{D_\sigma^{(\ell-1)}} \rightarrow \mathbb{R}^{D_W^{(\ell)}}$ and $\underline{f}_\sigma^{(\ell)} : \mathbb{R}^{D_W^{(\ell-1)}} \rightarrow \mathbb{R}^{D_\sigma^{(\ell)}}$. Such a composition of operators can be cast into a composition of MASO operators. First, due to the linear combinations of nonlinearities in the nonlinear layer, we first have to split this nonlinear operator $\underline{f}_\sigma^{(\ell)}$ into $\underline{f}_{W_\sigma}^{(\ell)} \circ \underline{f}_\sigma^{(\ell)}$ as it is clear from the above that after the nonlinearity with the max application, the linear combination is performed by a linear operator. With this definition and letting $\underline{f}_W^{(\ell)} = \underline{f}_W^{(\ell)}$ we obtain in accordance with our layer definition from 1, the same DN from above but defined as

$$f_\Theta(\mathbf{x}) = \left(\underline{f}_{W_\sigma}^{(L)} \circ \underline{f}_\sigma^{(L-1)} \circ \underline{f}_W^{(L-1)} \circ \underline{f}_{W_\sigma}^{(L-1)} \circ \cdots \circ \underline{f}_\sigma^{(2)} \circ \underline{f}_W^{(2)} \circ \underline{f}_{W_\sigma}^{(2)} \circ \underline{f}_\sigma^{(1)} \circ \underline{f}_W^{(1)} \right) (\mathbf{x}) \quad (111)$$

each of the above DN operators can now be cast as a MASO as follows:

<p><i>smallCNN</i></p> <pre>Conv2DLayer(layers[-1], 32, 3, pad='valid') Pool2DLayer(layers[-1], 2) Conv2DLayer(layers[-1], 64, 3, pad='valid') Pool2DLayer(layers[-1], 2) Conv2DLayer(layers[-1], 128, 1, pad='valid') Pool2DLayer(layers[-1], 2)</pre>	<p><i>largeCNN</i></p> <pre>Conv2DLayer(layers[-1], 96, 3, pad='same') Conv2DLayer(layers[-1], 96, 3, pad='full') Conv2DLayer(layers[-1], 96, 3, pad='full') Pool2DLayer(layers[-1], 2) Conv2DLayer(layers[-1], 192, 3, pad='valid') Conv2DLayer(layers[-1], 192, 3, pad='full') Conv2DLayer(layers[-1], 192, 3, pad='valid') Pool2DLayer(layers[-1], 2) Conv2DLayer(layers[-1], 192, 3, pad='valid') Conv2DLayer(layers[-1], 192, 1) Pool2DLayer(layers[-1], 2)</pre>
---	--

Figure 20. Description of the *smallCNN* and *largeCNN* topologies.

Partition	Number of Nonzero Volume Regions			
	Initialization (BN)	(BN)	Trained (BN)	(BN)
$\omega_{\text{in}}^{(1)}$	90	90	681	801
$\omega_{\text{in}}^{(2)}$	6	8	8	8
$\omega_g^{(2)}$	98	98	800	1068

Table 5. Number of nontrivial partition regions having nonzero volume for the input signal space partition induced by $\omega_{\text{in}}^{(1)}$ from layer 1 and $\omega_{\text{in}}^{(2)}$ from 2 of the toy DN from Figure 7 for leaky RELU.

- $f_W^{(\ell)} = S[A_W^{(\ell)}, B_W^{(\ell)}]$ with $D_W^{(\ell)} = \underline{D}_W^{(\ell)}$, $R = 1$ and $[A_W^{(\ell)}]_{k,1,\cdot} = [W]_{k,\cdot}$ and $[B^{(\ell)}]_{k,1} = 0, \forall k$,
- $f_\sigma^{(\ell)} = S[A_\sigma^{(\ell)}, B_\sigma^{(\ell)}]$ with $D_\sigma^{(\ell)} = \underline{R}\underline{D}_W^{(\ell)} + \underline{D}_W^{(\ell)}$, $R = 2$ and $[A_\sigma^{(\ell)}]_{k,1,\cdot} = e_{k\%D_W^{(\ell)}} [A_\sigma^{(\ell)}]_{k,2,\cdot} = 0, \forall k$ and $[B^{(\ell)}]_{k,1} = \tau_{k\%D_W^{(\ell)}, k/D_W^{(\ell)}, \ell}, [B^{(\ell)}]_{k,2} = 0, k = 1, \dots, \underline{R}\underline{D}_W^{(\ell)}$ and $[B^{(\ell)}]_{k,1} = [B^{(\ell)}]_{k,2} = 0, k = \underline{R}\underline{D}_W^{(\ell)}, \dots, \underline{R}\underline{D}_W^{(\ell)} + \underline{D}_W^{(\ell)}$,
- $f_{W_\sigma}^{(\ell)} = S[A_{W_\sigma}^{(\ell)}, B_{W_\sigma}^{(\ell)}]$ with $D_{W_\sigma}^{(\ell)} = \underline{D}_\sigma^{(\ell-1)}$, $R = 1$ and $[A_{W_\sigma}^{(\ell)}]_{k,1,\cdot} = \sum_{r=1}^R e_{r+(k-1)R} a_{r,k,\ell} + e_{R\underline{D}_W^{(\ell)}+k}$ and $[B_{W_\sigma}^{(\ell)}]_{k,1} = b_{1,k,\ell}, \forall k$.

leading to the entire DN being a composition of MASO operators. To simplify notation, one could collapse together $f_\sigma^{(\ell)} \circ f_W^{(\ell)} \circ f_{W_\sigma}^{(\ell)}$ into a single MASO layer. \square

F. Deep Network Topologies and Datasets

The *smallCNN* and *largeCNN* architectures used in the experiments are detailed in Figure 20, where “Conv2DLayer(layers[-1], 192, 3, pad='valid')” denotes a standard 2D convolution with 192 filters of spatial size (3, 3) and with valid padding (no padding). ResNetD-W denotes the standard wide ResNet topology with depth D and width W .

Figure 21 is the block diagram of the toy DN analyzed and visualized in Section 6.4 (e.g., Figure 7).

We used the standard datasets MNIST, CIFAR10, CIFAR100 and SVHN. We employed standard training regimes except for MNIST, where we drew at random the training and test set.

G. More on Orthogonal Templates

See Figure 22 for additional results from our experiments with orthogonal template DNs.

H. Additional Partition Visualization and Statistics

We provide additional experimental results for the toy DN studied in Section 6.4 in Tables 5–8.

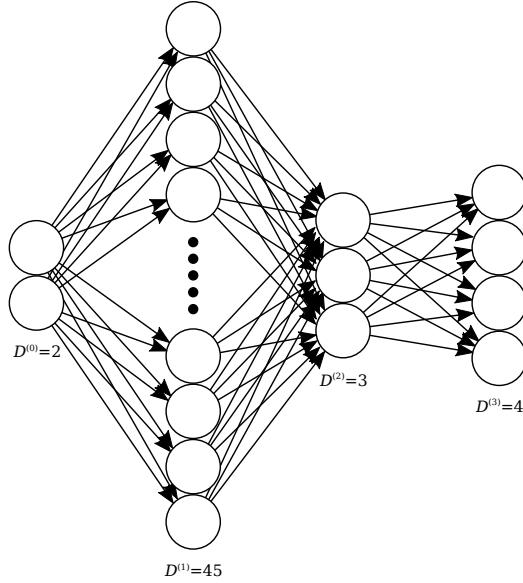


Figure 21. Block diagram of the toy DN used for the experiments in Figure 7.

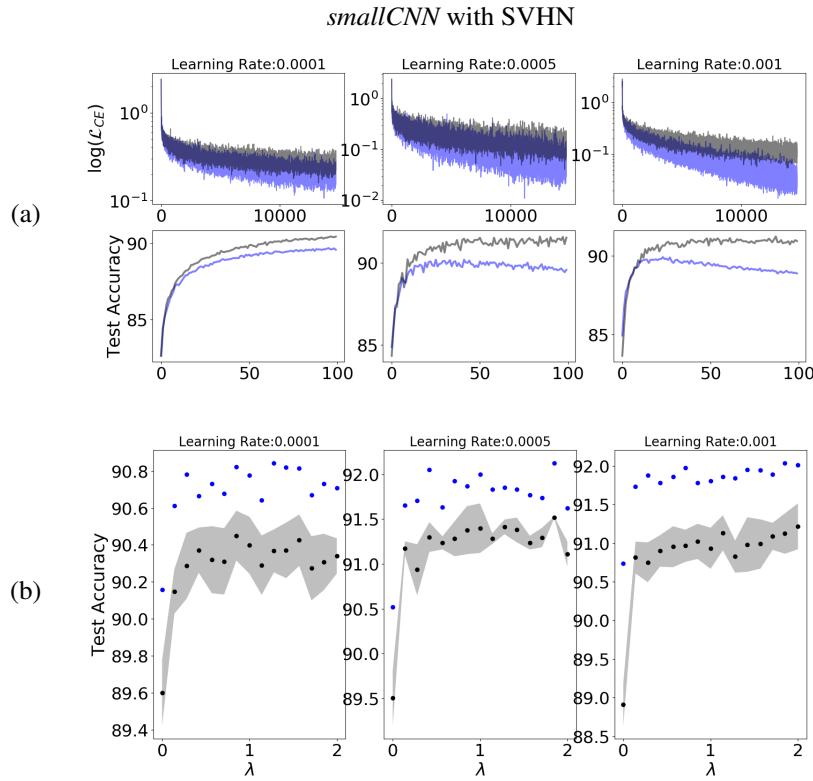


Figure 22. Further experiments on orthogonal templates for *smallCNN* with the SVHN dataset. (a) The presence of the orthogonal penalty loss shows greater accuracy as well as robustness to overfitting as opposed to standard DN setting. (b) Different values of the regularization parameter show the gain of having it greater than 0 hence the impact of the proposed penalty term.

Partition	Number of Nonempty Regions			
	Initialization (BN)	(BN)	Trained (BN)	(BN)
$\omega_{in}^{(1)}$	90	90	288	332
$\omega_{in}^{(2)}$	6	8	6	8
$\omega_g^{(2)}$	98	98	311	370

Table 6. Number of partition regions that are occupied by at least one training data point for the toy DN from Figure 7 for leaky RELU.

Partition	Number of Nonzero Volume Regions			
	Initialization (BN)	(BN)	Trained (BN)	(BN)
$\omega_{in}^{(1)}$	90	90	777	831
$\omega_{in}^{(2)}$	3	8	8	8
$\omega_g^{(2)}$	102	106	950	1115

Table 7. Number of nontrivial partition regions having nonzero volume for the input signal space partition induced by $\omega_{in}^{(1)}$ from layer 1 and $\omega_{in}^{(2)}$ from 2 of the toy DN from Figure 7 for the absolute value activation function.

Partition	Number of Nonempty Regions			
	Initialization (BN)	(BN)	Trained (BN)	(BN)
$\omega_{in}^{(1)}$	90	90	294	341
$\omega_{in}^{(2)}$	3	8	8	8
$\omega_g^{(2)}$	102	106	376	445

Table 8. Number of partition regions that are occupied by at least one training data point for the toy DN from Figure 7 for the absolute value activation function.

References

- Amos, Brandon, Xu, Lei, and Kolter, J Zico. Input convex neural networks. *arXiv preprint arXiv:1609.07152*, 2016.
- Arora, S., Bhaskara, A., Ge, R., and Ma, T. Provable bounds for learning some deep representations. *arXiv preprint arXiv:1310.6343*, 2013.
- Balestrieri, Randall and Baraniuk, Richard. A spline theory of deep networks. *Proc. Int. Conf. Mach. Learn. (ICML'18)*, Jul. 2018.
- Bengio, Yoshua, Boulanger-Lewandowski, Nicolas, and Pascanu, Razvan. Advances in optimizing recurrent networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8624–8628. IEEE, 2013.
- Bennett, JA and Botkin, ME. Structural shape optimization with geometric description and adaptive mesh refinement. *AIAA journal*, 23(3):458–464, 1985.
- Bishop, Christopher M. *Neural networks for pattern recognition*. Oxford university press, 1995.
- Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.
- Breiman, Leo. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, 39(3):999–1013, 1993.
- Bruna, Joan and Mallat, Stéphane. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.
- Cho, Kyunghyun, Van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- Chung, Junyoung, Gulcehre, Caglar, Cho, KyungHyun, and Bengio, Yoshua. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- Cohen, Nadav, Sharir, Or, and Shashua, Amnon. On the expressive power of deep learning: A tensor analysis. In Feldman, Vitaly, Rakhlin, Alexander, and Shamir, Ohad (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 698–728, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.
- Cybenko, George. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.
- De Boor, Carl, De Boor, Carl, Mathématicien, Etats-Unis, De Boor, Carl, and De Boor, Carl. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.
- de Brébisson, Alexandre and Vincent, Pascal. An exploration of softmax alternatives belonging to the spherical loss family. *arXiv preprint arXiv:1511.05042*, 2015.
- Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- Eldar, Yonina C and Oppenheim, Alan V. Orthogonal matched filter detection. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 5, pp. 2837–2840. IEEE, 2001.
- Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.
- Gersho, Allen and Gray, Robert M. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.
- Ghahramani, Zoubin and Jordan, Michael I. Factorial hidden markov models. In *Advances in Neural Information Processing Systems*, pp. 472–478, 1996.

- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 2672–2680. MIT Press, 2014.
- Goodfellow, Ian J., Warde-Farley, David, Mirza, Mehdi, Courville, Aaron, and Bengio, Yoshua. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.
- Graves, Alex. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.
- Graves, Alex and Schmidhuber, Jürgen. Framewise phoneme classification with bidirectional lstm networks. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 4, pp. 2047–2052. IEEE, 2005.
- Hannah, Lauren A and Dunson, David B. Multivariate convex regression with adaptive partitioning. *The Journal of Machine Learning Research*, 14(1):3261–3294, 2013.
- He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- Hecht-Nielsen, Robert et al. Theory of the backpropagation neural network. *Neural Networks*, 1(Supplement-1):445–448, 1988.
- Hinton, Geoffrey E and Zemel, Richard S. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pp. 3–10, 1994.
- Hochreiter, Sepp and Schmidhuber, Jürgen. Flat minima. *Neural Computation*, 9(1):1–42, 1997.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Jain, Anil K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.
- Jayaraman, S., Esakkirajan, S., and Veerakumar, T. Digital image processing tmh publication. *Year of Publication*, 2009.
- Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kurakin, Alexey, Goodfellow, Ian, and Bengio, Samy. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- LeCun, Yann. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- LeCun, Yann, Jackel, LD, Bottou, Léon, Cortes, Corinna, Denker, John S, Drucker, Harris, Guyon, Isabelle, Muller, UA, Sackinger, Eduard, Simard, Patrice, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276, 1995.
- Lu, Haihao and Kawaguchi, Kenji. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.
- Magnani, Alessandro and Boyd, Stephen P. Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1):1–17, 2009.
- Mallat, Stéphane. *A wavelet tour of signal processing*. Academic press, 1999.
- Montufar, Guido F, Pascanu, Razvan, Cho, Kyunghyun, and Bengio, Yoshua. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pp. 2924–2932, 2014.
- Nam, Woonhyun, Dollár, Piotr, and Han, Joon Hee. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, pp. 424–432, 2014.
- Nasrabadi, Nasser M and King, Robert A. Image coding using vector quantization: A review. *IEEE Transactions on communications*, 36(8):957–971, 1988.

- Nishikawa, Hiroaki. Accurate piecewise linear continuous approximations to one-dimensional curves: Error estimates and algorithms. 1998.
- Pal, Sankar K and Mitra, Sushmita. Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on neural networks*, 3(5):683–697, 1992.
- Papyan, Vardan, Romano, Yaniv, and Elad, Michael. Convolutional neural networks analyzed via convolutional sparse coding. *Journal of Machine Learning Research*, 18(83):1–52, 2017.
- Patel, Ankit B, Nguyen, Tan, and Baraniuk, Richard G. A probabilistic framework for deep learning. *NIPS*, 2016.
- Powell, Michael James David. *Approximation theory and methods*. Cambridge university press, 1981.
- Rabiner, Lawrence R and Gold, Bernard. Theory and application of digital signal processing. *Englewood Cliffs, NJ, Prentice-Hall, Inc.*, 1975. 777 p., 1975.
- Rister, Blaine and Rubin, Daniel L. Piecewise convexity of artificial neural networks. *Neural Networks*, 94:34–45, 2017.
- Roberts, Arthur Wayne. Convex functions. In *Handbook of Convex Geometry, Part B*, pp. 1081–1104. Elsevier, 1993.
- Romberg, J. K., Choi, H., and Baraniuk, Richard G. Bayesian tree-structured image modeling using wavelet-domain hidden markov models. In *Mathematical Modeling, Bayesian Estimation, and Inverse Problems*, volume 3816, pp. 31–45. International Society for Optics and Photonics, 1999.
- Rumelhart, David E, Hinton, Geoffrey E, Williams, Ronald J, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.
- Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- Soatto, S. and Chiuso, A. Visual representations: Defining properties and deep approximations. In *Proc. Int. Conf. Learn. Rep. (ICLR'16)*, Nov. 2016.
- Soudry, Daniel and Hoffer, Elad. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.
- Srivastava, Rupesh K, Greff, Klaus, and Schmidhuber, Jürgen. Training very deep networks. In *Advances in neural information processing systems*, pp. 2377–2385, 2015.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- Szegedy, Christian, Ioffe, Sergey, Vanhoucke, Vincent, and Alemi, Alexander A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, pp. 12, 2017.
- Targ, Sasha, Almeida, Diogo, and Lyman, Kevin. Resnet in resnet: generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016.
- Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.
- Tishby, Naftali and Zaslavsky, Noga. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pp. 1–5. IEEE, 2015.
- Unser, Michael. A representer theorem for deep neural networks. *arXiv preprint arXiv:1802.09210*, 2018.
- Veit, Andreas, Wilber, Michael J, and Belongie, Serge. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, pp. 550–558, 2016.
- Vincent, Pascal, Larochelle, Hugo, Bengio, Yoshua, and Manzagol, Pierre-Antoine. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103. ACM, 2008.

- Wei, Li-Yi and Levoy, Marc. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 479–488. ACM Press/Addison-Wesley Publishing Co., 2000.
- Xu, Yangyang and Yin, Wotao. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.
- Zeiler, Matthew D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
- Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.
- Zemel, Richard S. *A minimum description length framework for unsupervised learning*. University of Toronto, 1994.
- Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.