Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2009-049
CBCL-281

October 9, 2009

Notes on the Shannon Entropy of the Neural Response

Jake Bouvrie, Lorenzo Rosasco, Greg Shakhnarovich, and Steve Smale

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

CSAIL

Notes on the Shannon Entropy of the Neural Response

Jake Bouvrie[†], Lorenzo Rosasco[†], Greg Shakhnarovich[‡], Steve Smale[♯]

† *Center for Biological and Computational Learning, MIT*

‡ *Toyota Technological Institute at Chicago*

♯ *City University of Hong Kong and University of California, Berkeley*

{jvb, lrosasco}@mit.edu, gregory@tti-c.org, smale@math.berkeley.edu

September 29, 2009

Abstract

In these notes we focus on the concept of Shannon entropy in an attempt to provide a systematic way of assessing the discrimination properties of the neural response, and quantifying the role played by the number of layers and the number of templates.

1 Introduction

In a recent effort [1], we defined a distance function on a space of images which reflects how humans see the images. In this case, the distance between two images corresponds to how similar they appear to an observer. We proposed in particular a natural image *representation*, the neural response, motivated by the neuroscience of the visual cortex. The "derived kernel" is the inner product defined by the neural response and can be used as a similarity measure. A crucial question is that of the trade-off between invariance and discrimination properties of the neural response. In [1], we suggested that Shannon entropy is a useful concept towards understanding this question.

Here we substantiate the use of Shannon entropy [2] to study discrimination properties of the neural response as proposed in [1]. The approach sheds light on natural questions that arise in an analysis of the neural response: How should one choose the patch sizes? How many layers are appropriate for a given task? How many templates should be sampled? How do architectural choices induce invariance and discrimination properties? These are important and involved questions of broad significance. In this note, we suggest a promising means of clarifying the picture in simplified situations that can be potentially extended to more general settings and ultimately provide answers to the questions posed above.

This note is organized as follows. In Section 2 we begin by briefly recalling the definitions of the neural response, derived kernel, and Shannon entropy of the neural response. The reader is encouraged to consult [1] for a detailed treatment. In Section 3 we then study discrimination properties in terms of information-theoretic quantities in the case of two and three layer architectures defined on strings. Finally, we provide in Section 4 remarks which derive intuition from the preceding development and provide additional insight into the outstanding questions above.

2 Background

We first briefly recall the definition of the neural response following the development in [1], where a background discussion on Shannon entropy is also provided. The definition of neural response and the derived kernel is based on a recursion which defines a hierarchy of local kernels, and can be interpreted as a multi-layer architecture.

2.1 Neural Response

Consider an n layer architecture given by sub-patches $v_1 \subset v_2 \subset \cdots \subset v_n = Sq$. We use the notation $K_n = K_{v_n}$ and similarly $H_n = H_{v_n}$, $T_n = T_{v_n}$. Given a kernel K, we define a normalized kernel via $\widehat{K}(x,y) = K(x,y)/\sqrt{K(x,x)K(y,y)}$. The following definition of the derived kernel and neural response is given in [1], and we refer the reader to this source for more details.

Definition 2.1. Given a normalized, non-negative valued initial reproducing kernel $\widehat{K}_1$, the m layer derived kernel $\widehat{K}_m$, for $m = 2, \ldots, n$, is obtained by normalizing

$$K_m(f,g) = \langle N_m(f), N_m(g) \rangle_{L^2(T_{m-1})}$$

where

$$N_m(f)(t) = \max_{h \in H} \widehat{K}_{m-1}(f \circ h, t), \qquad t \in T_{m-1}$$

with $H = H_{m-1}$.

From the above definition we see that, the neural response is a map

$$\underbrace{f \in \mathrm{Im}(Sq)}_{\text{input}} \longmapsto \underbrace{\widehat{N}_{Sq}(f) \in L^2(T) = \mathbb{R}^{|T|}}_{\text{output}},$$

with $T = T_{m-1}$ and we let $\widehat{N}_m$ denote the normalized neural response given by $\widehat{N}_m = N_m/\|N_m\|_{L^2(T)}$. We can now define a thresholded variant of the neural response, along with the induced pushforward measure on the space of orthants. In the discussion that follows, we will study the entropy of this pushforward measure as well as that of the natural measure on the space of images.

2.2 Thresholded Neural Response

Denote by $\mathcal{O}$ the set of orthants in $L^2(T_{m-1}) = \mathbb{R}^{|T_{m-1}|}$ identified by sequences of the form $o = (\epsilon_i)_{i=1}^{|T|}$ with $\epsilon_i \in \{0,1\}$ for all i. If we assume that $\mathbb{E}[\widehat{N}_m(f)(t)] = 0$, then the map $\widehat{N}_m^* : \mathrm{Im}(v_m) \to \mathcal{O}$ can be defined by

$$\widehat{N}_m^*(f) := \Big(\Theta(\widehat{N}_m(f)(t))\Big)_{t \in T_{m-1}}$$

where $\Theta(x) = 1$ when $x > 1$ and is 0 otherwise. From this point on, we assume normalization and drop hats in the notation. Finally, we denote by $N^{**}\rho$ the push-forward measure induced by N^*, that is

$$N^{**}\rho(A) = \rho\big(\{f \in \mathrm{Im}(Sq) \mid N_v(f) \in A\}\big),$$

for any measurable set $A \subset L^2(T_{m-1})$.

2.3 Shannon Entropy of the Neural Response

We introduce the Shannon entropies relative to the measures ρ_v and $N_v^{**}\rho_v$. Consider the space of images $\mathrm{Im}(v) = \{f_1, \ldots, f_d\}$ to be finite. Then ρ_v reduces to $\{p_1, \ldots, p_d\} = \{\rho_v(f_1), \ldots, \rho_v(f_d)\}$. In this case the entropy of the measure ρ_v is

$$S(\rho_v) = \sum_i p_i \log \frac{1}{p}_i$$

and similarly,

$$S(N_v^{**}\rho_v) = \sum_{o \in \mathcal{O}} q_o \log \frac{1}{q}_o .$$

where $q_o = (N_v^{**}\rho_v)(o)$ is explicitly given by

$$(N_v^{**}\rho_v)(o) = \rho_v\Big(\big\{f \in \mathrm{Im}(v) \mid \big(\Theta(N_v(f)(t))\big)_{t \in |T|} = o\big\}\Big).$$

If $\mathrm{Im}(v)$ is not finite we can define the entropy $S(\rho_v)$ associated to ρ_v by considering a partition $\pi = \{\pi_i\}_i$ of $\mathrm{Im}(v)$ into measurable subsets. The entropy of ρ_v given the partition π is then given by

$$S_\pi(\rho_v) = \sum_i \rho_v(\pi_i) \log \frac{1}{\rho_v(\pi_i)}.$$

We can define $S_\pi(N_v^{**}\rho_v)$ similarly. The key quantity, to assess the discriminative power of the neural response N_v, is the *discrepancy*

$$\Delta S = S(\rho_v) - S(N_v^{**}\rho_v).$$

It is easy to see that

$$S(\rho_v) \geq S(N_v^{**}\rho_v) \tag{1}$$

so that $\Delta S \geq 0$. Furthermore, the discrepancy is zero if N^* is one to one (see remark below).

We add two remarks.

Remark 2.1. *Let X, Y be two random variables. We briefly recall the derivation of the inequality* (1)*, which we write here as $S(Y) \leq S(X)$. We use two facts: (a) $S(X,Y) = S(X)$ if $Y = f(X)$ with f deterministic, and (b) $S(X,Y) \geq S(Y)$ in general. To prove (a), write $P(Y = y, X = x) = p(Y = y|X = x)P(X = x) = \delta(y, f(x))P(X = x)$, and we sum over all $y = x$ in the definition of the joint entropy $S(X,Y)$.*

Remark 2.2. *Consider, for example, the finite partition $\pi = \{\pi_o\}_{o\in\mathcal{O}}$ on the space of images induced by N_v^*, with*

$$\pi_o = \big\{ f \in \mathrm{Im}(v) \mid \big(\Theta(N_v(f)(t))\big)_{t\in|T|} = o \big\}.$$

We might also consider only the support of ρ_v, which could be much smaller than $\mathrm{Im}(v)$*, and define a similar partition of this subset as*

$$\pi_o = \big\{ f \in \operatorname{supp} \rho_v \mid \big(\Theta(N_v(f)(t))\big)_{t\in|T|} = o \big\},$$

with $\pi = \{\pi_o \mid \pi_o \neq \emptyset\}$. One can then define a measure on this partition and corresponding notion of entropy.

3 Shannon Entropy of the Neural Response: the String Case

Let A be an alphabet of k distinct letters so that $|A| = k$. Consider three layers $u \subset v \subset w$, where $\mathrm{Im}(u) = A$, $\mathrm{Im}(v) = A^m$ and $\mathrm{Im}(w) = A^n$, with $1 < m < n$. The kernel $K_u = \widehat{K}_u$ on single characters is simply, $K_u(f,g) = 1$, if $f = g$ and 0 otherwise. The template sets are $T_u = A$ and $T_v = A^m$.

3.1 Explicit Expressions for N and K

We specialize the definitions of the neural response and derived kernel in the case of strings.
The neural response at layer v is defined by

$$N_v(f)(t) = \max_{h\in H_u} \left\{ \widehat{K}_u(f \circ h, t) \right\},$$

and is a map $N_v : A^m \to \{0,1\}^k$. The norm of the neural response is

$$\|\widehat{N}_v(f)\| =: a(f) = \#\text{ distinct letters in } f.$$

From the definition of the derived kernel we have that

$$K_v(f,g) =: a(f,g) = \#\text{ distinct letters common to } f \text{ and } g.$$

The normalized kernel can then be written as

$$\widehat{K}_v(f,g) = \frac{a(f,g)}{\big(a(f)a(g)\big)^{1/2}}.$$

The neural response at layer w then satisfies

$$N_w(f)(t) = \frac{e(f,t)}{a(t)^{1/2}},$$

with

$$e(f,t) := \max_{h \in H_v} \frac{a(f \circ h, t)}{a(f \circ h)^{1/2}}.$$

This is the maximum fraction of distinct letters in m-substrings of f that are shared by t. Finally the derived kernel at layer w satisfies

$$\widehat{K}_w(f,g) = \frac{\sum_{t \in T_v} \frac{e(f,t)e(g,t)}{a(t)}}{\sum_{t \in T_v} \frac{e(f,t)^2}{a(t)} \sum_{t \in T_v} \frac{e(q,t)^2}{a(t)}}.$$

We are interested in knowing whether the neural response is injective up to reversal and checkerboard. If N_v^* is injective, then inequality' (1) holds with equality. We can consider N_v^* as acting on the set of equivalence classes of $\mathrm{Im}(v)$ defined by the strings and their reversals, and if n is odd, a checkerboard when applicable (see [1] for a discussion concerning the checkerboard pattern). Here injectivity of N_v^* is with respect to the action on equivalence classes. The following result is easy to prove.

Proposition 3.1. *N_v^* is injective if and only if $\mathrm{Im}(v)$ contains strings of length 2.*

3.2 Orthant Occupancy

We consider a 2-layer architecture and let $k = |A| > m$. As before, $\mathrm{Im}(v)$ contains strings of length m, and $\mathrm{Im}(u)$ contains single characters. The number of non-empty orthants is

$$\sum_{\ell=1}^{m} \binom{k}{\ell}.$$

The "all zero" orthant is always empty (strings must use at least one letter in the alphabet). Let $\mathcal{O}_p$ denote the set of orthants corresponding to strings of $p < m$ distinct letters, that is

$$\mathcal{O}_p = \big\{ o \in \mathcal{O} \mid \sum_{i=1}^{k} \epsilon_i = p \big\}.$$

Let $\lambda_o(p,m)$ denote the number of strings mapped into the orthant $o \in \mathcal{O}_p$. Then

$$\lambda_o(p,m) = k^m q_o.$$

If the measure ρ_v is uniform then $\lambda_o(p,m)$ is the same for all $o \in \mathcal{O}_p$ and we drop the subscript on λ. In the uniform case we have the following recursive formula

$$\lambda(p,m) = p^m - \sum_{j=1}^{p-1} \binom{p}{j} \lambda(j,m),$$

with $\lambda(1,m) = 1$.

We now give an explicit expression for the discrepancy $S(\rho_v) - S(N_v^{**}\rho_v)$. If ρ_v is uniform

$$S(\rho_v) = m \log k.$$

With little work we have that

$$\begin{aligned}
S(N_v^{**}\rho_v) &= -\sum_j^m \binom{k}{j} \frac{\lambda(j,m)}{k^m} \log \frac{\lambda(j,m)}{k^m} \\
&= -\sum_j^m \binom{k}{j} \frac{\lambda(j,m)}{k^m} \log \lambda(j,m) + \underbrace{\log k^m}_{S(\rho_v)} \underbrace{\sum_j^m \binom{k}{j} \frac{\lambda(j,m)}{k^m}}_{=1},
\end{aligned}$$

and we obtain the following explicit expression for the discrepancy

$$\Delta S = S(\rho_v) - S(N_v^{**}\rho_v) = \sum_j^m \binom{k}{j} \frac{\lambda(j,m)}{k^m} \log \lambda(j,m).$$

This quantity can be seen as a weighted average, writing $\Delta S = \sum_j^m b_j \log \lambda(j,m)$ and noting that $\sum_j b_j = 1$.

3.2.1 Non-recursive Occupancy Formula (2-layer Case)

Alternatively, we can use multinomial coefficients to describe the number of m-strings mapped into the $\binom{k}{p}$ orthants with exactly $p < m$ ones as follows:

$$\begin{aligned}
\lambda(p,m) &= \sum_{r_1,\dots,r_p} \binom{m}{r_1,\dots,r_p} \\
&= p! S(m,p) \\
&= \sum_{t=1}^p (-1)^{p+t} \binom{p}{t} t^m
\end{aligned}$$

where the first summation is taken over all sequences of *positive* integer indices $r_1,\dots,r_p$ such that $\sum_{i=1}^p r_i = m$. The number of terms in this summation is the number of p-part compositions of m[1] and is given by $\binom{m-1}{p-1}$. The $S(m,p)$ are Stirling numbers of the second kind [2], and the final equality follows from direct application of Stirling's Identity. Note that since $S(m,1) = 1$, we can verify that $\lambda(1,m) = S(m,1) = 1$.

From the multinomial theorem, we also have that

$$\sum_{\{r_i \geq 0 : r_1 + \cdots + r_p = m\}} \binom{m}{r_1,\dots,r_p} = (1 + \cdots + 1)^m = p^m = \sum_{k=1}^p \lambda(k,m)$$

which checks with the previous recursive definition.

4 Final Remarks

We add some remarks concerning the application of the above ideas towards understanding the role of the patch sizes and layers in inducing discrimination and invariance properties.

- In a 3-layer network, $u \subset v \subset w$, $N_w^*(f)(t) \to 1$ in probability as $n \to \infty$, for all $t \in T_v$, with T_v exhaustive and $f \sim \rho_w$ with ρ_w uniform. As the string f gets infinitely long, then the probability we find a given template in that string goes to 1. Note that the rate is very slow: for example, there are $(k-1)^n$ possible strings which do not include a given letter which would appear in many templates.

[1] A p-part composition of m is a solution to $m = r_1 + \cdots + r_p$ consisting of positive integers.

[2] $S(m,p)$ counts the number of ways one can partition sets of size m into p nonempty subsets.

- The above also implies that the entropy $S(N_w^{**}\rho_w) \to 0$ as $n \to \infty$ since all images f are mapped to the orthant of all ones.

- Consider a 3 layer architecture: $u = \mathrm{Str}(1) \subset v = \mathrm{Str}(m) \subset w = \mathrm{Str}(n)$ with n fixed, all translations, and exhaustive template sets.

  Question: Which choice of m maximizes $S(N_w^{**}\rho_w)$?

  Intuition: For large m most of the probability will concentrate near the "sparse" orthants – the orthants characterized by many zeros– because the probability of finding a long template in f is low. For small m, most of the probability mass will fall in the orthants with many ones – where a large number of templates match pieces of f. In both cases, the entropy is low because the number of orthants with few 1's or few 0's is small. For some intermediate choice of m, the entropy should be maximized as the probability mass becomes distributed over the many orthants which are neither mostly zeros nor mostly ones. (consider the fact that $\binom{a}{a/2} \gg \binom{a}{1}$ or $\binom{a}{a-1}$).

In this note, we have shown the possibility of mathematically analyzing the discriminatory power of the neural response in simple cases, via entropy. It is our hope that the methods suggested here can be extended and ultimately leveraged to understand in concrete terms how parametric and architectural choices influence discrimination and invariance properties.

References

[1] S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. "Mathematics of the Neural Response", *Foundations of Computational Mathematics*, June 2009 (online).

[2] T. M. Cover and J. A. Thomas. *Elements of information theory.* John Wiley and Sons, Inc., 1991.