SPARSE DISTRIBUTED REPRESENTATIONS AND WITNESS COMPLEXES

MIRKO KLUKAS

Abstract. The starting point for this note is a white paper of the science startup *Numenta* (http://numenta.com/). At the core of the paper lies the model of a sequence memory that enables the prediction of elements in a time series based on its history of predecessors. An essential ingedient in the model are *sparse distributed representations*.

The goal of this paper is (a) to briefly recall what sparse representations are, and (b) how to encode real-world data as sparse representations. We roughly follow Numenta's approach and connect it to a construction well-known in computational topology called *witness complex*.

Contents

1. Sparse Distributed Representations

1.1. **The space of patterns.** Let X denote a discrete set of n elements, where n is a positive integer. Think of it as an (a priori unordered) collection of bits, $\{1, \dots, n\}$ say. Denote by $\mathcal{P} = \mathcal{P}_n$ the power set of X, i.e.

$$\mathcal{P} := \big\{p : p \subset X\big\}.$$

We will refer to elements in $\mathcal{P}$ as **patterns**. Another intuitive example is to think of X as representing a family of pixels, and a pattern representing the collection of black pixels of an black-and-white image. There are two natural operations on $\mathcal{P}$, namely the union of sets $\cup$, and the intersection of sets $\cap$. The latter in fact gives rise to a notion of **similarity** of two patterns p and p' in terms of their intersection or overlap. We define the **overlap count** of p, p' as

$$\omega(p, p') := |p \cap p'|.$$

Date: May 13, 2016.

Intuitively: the bigger the overlap the more **similar** the two patterns are. This is a somewhat vague notion – since we do not incorporate the individual sizes of the patterns – but for now we do not want to bother too much about that, when we restrict our attention to *sparse patterns*, in §1.3 below, this notion of similarity becomes more acurate.

1.2. **Patterns as binary vectors.** Note that we can naturally identify the power set of n elements $\mathcal{P}$ with $\{0,1\}^n$, the boolean algebra of 2^n elements, as follows: recall that $\mathcal{P}$ denotes the power set of a set of n distinct elements, $\{1,\dots,n\}$ say. Then one easily checks that the desired isomorphism from $\{0,1\}^n$ into $\mathcal{P}$ is given by

$$(v_1,\dots,v_n) \mapsto \{i : v_i = 1\}.$$

With respect to the above identification the union of sets $\cup$ corresponds to the componentwise or-operation $\vee$ on $\{0,1\}^n$, and the intersection of sets $\cap$ corresponds to the componentwise and-operation $\wedge$ on $\{0,1\}^n$. Therefore we indeed have

$$\big(\mathcal{P};\cup,\cap\big) \cong \big(\{0,1\}^n;\vee,\wedge\big).$$

Note that we could understand $\{0,1\}^n$ as embedded into n-dimensional Euclidean Space $\mathbb{R}^n$. Then the above notion of similarity corresponds to the euclidean scalar product of two vectors. This viewpoint allows us to pull back any distance, or scalarproduct, defined on $\mathbb{R}^n$ to $\mathcal{P}$. We can pull back the metric induced by the L^1-norm on $\mathbb{R}^n$, for instance, which yields the the **Hamming distance** on $\{0,1\}^n$ (interpreted as binary strings). We may come back to that later.

1.3. **Sparse patterns and their natural metric.** The central objects in the present paper are *sparse representations* in the sense of [3] and [4]. For some fixed positive integer $k \ll n$ we call a pattern $p \in \mathcal{P}$ **sparse** if the number of elements in p satisfies $|p| = k$. Finally we denote by $SDR = SDR(k,n)$ the set of sparse patterns in $\mathcal{P}$ and refer to it as **sparse distributed representations**, i.e. we define

$$SDR := \big\{p \in \mathcal{P} : |p| = k\big\}.$$

The terminology anticipates (and only makes sense in combination with) a *sparse encoder* satisfying certain properties. We describe a particular construction of such an encoder in §2. It is easy to check that the notion of similarity ω, defined above, gives rise to an honest metric $d = d_k$ on the subset of sparse patterns SDR defined by

$$d_H(p,q) := k - |p \cap q|.$$

This is equivalent to the Hamming distance (up to a constant factor of 2) on $\mathcal{P}$ restricted to the set of sparse patterns. Let us also define a softer version of sparse representations by including patterns with less or equal than k elements, i.e. we define

$$SDR_{\leq k} := \big\{p \in \mathcal{P} : |p| \leq k\big\}.$$

Remark 1 (Sparseness after Olshausen and Field)**.** In [5] *sparseness* turned up within a probabilistic framework where the authors try to match the probability distribution over images observed in nature by a generative linear model. However, in contrast to our approach, the coefficients in [5] are allowed to take values in $\mathbb{R}$. The notion of sparseness then corresponds to assuming a certain prior probability distribution over these coefficients, whose density is shaped to be unimodal and peaked at zero with heavy tails (cf. Figure 2), implying that coefficents are mostly

inactive. However there is no restriction on the actual number of active units. If we would like to emphasize this difference, we will refer to a sparse vector in the sense presented here as a *binary sparse vector*. However it should always be clear from the context what definition is referred to.

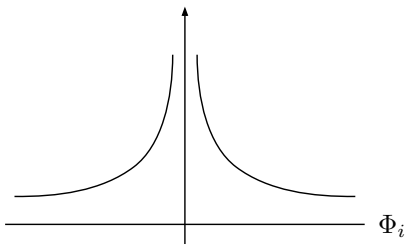

FIGURE 1. Probability density of sparse coefficients in the sense of Olshausen and Field: unimodal and peaked at zero with heavy tails.

2. Spatial pooling and weak witness complexes

In the present section I introduce a geometrical perspective on how to encode inputs sampled from a metric space as sparse representations. We follow an idea analogous to the approach in [3] and connect it to a construction well-known in computational topology called *witness complex* [1] (cf. also [2]).

Let (X, d) be a, not necessaryly discrete, metric space, e.g. $X = \mathbb{R}^m$ endowed with the Euclidean distance. Our goal is the construction of a map

$$\Phi\colon\ X \to SDR \subset \mathcal{P}_n.$$

We refer to such a map Φ as an (untrained) **spatial pooler** or (binary) **sparse encoder**. We will address different notions of what it means for an encoder to be *trained* on a collection of inputs $Y \subset X$ in another note. Let me first describe a way to construct such a map. Choose a family of n landmarks

$$L = \{l_1, \ldots, l_n\} \subset X.$$

These landmarks can be understood as *features*, or elements of a *codebook*, that is an overcomplete basis of the input space. Given an input $x \in X$ we assign to it the collection of indices $\Lambda(x) \subset \{1, \ldots, n\}$ of the k closest landmarks, i.e. we define

$$\Lambda(x) := \Lambda_L^{(k)}(x) := \{i_1, \ldots, i_k\},$$

such that $d(x, l_i) < d(x, L \setminus \{l_{i_1}, \ldots, l_{i_k}\})$ for each $i \in \Lambda(x)$. Here $d(x, Y)$ deotes the the minimal distance of a point x to the elements in a finite set Y. Note that in general this assignment is not well-defined since the set of k closest landmarks may not be unique – in practice however this can always be achieved by adding a tiny random perturbation to each landmark, or by fixing a rule of precedence, e.g. prefer the landmark with the lower index.[1]

[1]The latter approach for instance is taken in the spatial pooler implementation of Numenta's open source library *nupic* (https://github.com/numenta/nupic/tree/master/src/nupic/research)

Remark 2. (Landmarks in Numenta's spatial pooling) We can interpret the permanence values of the synaptic connections of a particular column in [3] as a landmark in $[0,1]^m$.

For each $i \in \Lambda(x)$ we call x a **weak witness** (or for simplicitly just **witness**) for l_i,i.e. x is a witness for each of the k closest landmarks – note that our definition of a witness slightly differs from the construction in [1], but it still captures the essence of the construction. (We also give a slight variation of this notion in §2.2 below.) For a lack of better imagination we refer to the collection of witnesses

$$W(l) := \{x : x \text{ is a witness for } l\}.$$

associated to a particular landmark l as its associated **(depth-k) witness cell**. Note that for $k = 1$ the witness cell of a landmark equals its associated *Voronoi cell*. We could also be interested in the the collection of common witnesses associated to a family of landmarks $l_{i_1}, ..., l_{i_q}$ and hence define

$$W(l_{i_1}, ..., l_{i_q}) := \bigcap_{i\in} W(l_i).$$

This obviously generalizes the above notion of a (depth-k) witness cell. Analogous to the simpler version, for $k = q$ this equals the *order-k Voronoi region* associated to $l_{i_1}, ..., l_{i_q}$. In §2.3 below we describe how we can use these notions to define a simplicial complex associated to a collection of inputs.

Obviously we can understand $\Lambda(x)$ as an element of $SDR(k, n)$ for each $x \in X$, and hence the assignment

$$\Phi\colon\ x \mapsto \Lambda(x)$$

defines the desired map. To *train* an encoder on a collection of inputs translates into the right choice of landmarks. This will be addressed in a follow-up paper.

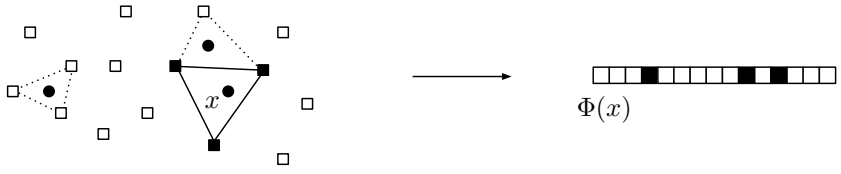

FIGURE 2. Three points in X on the left and the sparse representation $\Phi(x)$ (for $n = 15$ and $k = 3$) of one of them as a binary vector on the right. The vertices of the dotted triangles indicate the representations of the other two remaining points. The squares represent the landmarks.

2.1. **Witnesses with respect to similarity.** Suppose, instead of an honest metric we are given a notion of similarity on X expressed in terms of a non-negative function

$$\omega\colon\ X \times X \to \mathbb{R}_{\geq 0}.$$

Consider the m-dimensional cube $[0,1]^m$ endowed with the Euclidean dot product, for instance. Then the above construction of Λ also follows through if we look for the k most similar landmarks – instead of the k closest landmarks. To be more

precise we define $\Lambda(x) := \{i_1, \ldots, i_k\}$ such that $\omega(x, l_i) > \omega(x, l_j)$ for each $i \in \Lambda(x)$ and $j \notin \Lambda(x)$. (This is the approach taken in [3])

2.2. **Restricting the sight of witnesses.** Instead of assigning the k closest (or most similar respectively) landmarks to a point, we can further restrict the set the potential landmarks to be within a certain radius, $\theta > 0$ say. To be more precise, x is a **witness with sight θ for** l iff l is among the k closest landmarks and x is contained in the ball of radius θ centered at l, i.e. $d(x, l) < \theta$. The problem with this approach is that we have to ensure that we find k landmarks to form a complete sparse representation, i.e. one that has exactly k active bits. This can be achieved e.g. by adding random elements to the representation until completion, or by fixing a rule of precedence upon which we choose elements to complete the representation.[2]

A way around this is to establish a softer version of sparse representations, e.g. allow patterns with less or equal than k elements. This then yields a variation of the above sparse encoder with values in $SDR_{\leq k}$ defined by

$$\Lambda_\theta(x) := \big\{l : d(x, l) < \theta \text{ and } l \text{ is among the } k \text{ closest landmarks}\big\}.$$

2.3. **Approximation of the input space by simplicial complexes.** Suppose we are given a collection of inputs $Y \subset X$. Then there is a simplicial complex $\mathcal{W}(Y, L)$, the **(weak) witness complex** (cf. [2] or [1]), whose vertex set is given by L, and where $\Lambda = \{l_{i_1}, ..., l_{i_q}\}$ spans a q-simplex iff there is a common witness for all landmarks in Λ, i.e. if $W(l_{i_1}, ..., l_{i_q}) \neq \emptyset$. An obvious variation of this complex is obtained by restricting the set of witnesses to those with sight θ. We denote this complex by $\mathcal{W}_\theta(Y, L)$. This obviously defines a subcomplex of the (weak) witness complex. Usually we are interested in the subcomplexes of dimension less or equal than k, and we can understand the witness complexes as an simplicial approximation of the data $Y \subset X$.

References

[1] Gunnar Carlsson, *Topological estimation using witness complexes*, Eurographics Symposium on Point-Based Graphics (2004).

[2] ———, *Topology and data*, Bull. Amer. Math. Soc. (N.S.) **46** (2009), no. 2, 255–308, DOI 10.1090/S0273-0979-09-01249-X. MR2476414 (2010d:55001)

[3] Numenta, *Hierarchical Temporal Memory (HTM) Whitepaper* (2011), available at `http://numenta.com/learn/hierarchical-temporal-memory-white-paper.html`.

[4] Subutai Ahmad and Jeff Hawkins, *Properties of Sparse Distributed Representations and their Application to Hierarchical Temporal Memory*, ArXiv e-prints (2015), available at `http://arxiv.org/abs/1503.07469`.

[5] B. A. Olshausen and D. J. Field, *Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?*, Vision Research **37** (1997), 3311–3325.

[6] Yoshua Bengio and Lon Bottou, *Convergence Properties of the K-Means Algorithms*, Advances in Neural Information Processing Systems 7, 1995, pp. 585–592.

E-mail address: `mirko.klukas@gmail.com`

(Mirko Klukas) Institute of Science and Technology Austria (IST Austria), Am Campus 1, A 3400 Klosterneuburg, Austria

[2]This is essential what is done for instance in the spatial pooler implementation of Numenta's open source library *nupic* (https://github.com/numenta/nupic/tree/master/src/nupic/research) up from version 0.3.5