Understanding Memory-Equivalent Capacity in Unsupervised Learning

Jin Park
University of California, Berkeley
jpark96@berkeley.edu

Abstract

Recent works in bounding the representational power of neural networks have established theoretical tools to measure machine learning models[1] [2] [3]. However, these frameworks are restricted to supervised models. In this paper, we attempt to understand memory-equivalent capacity in an unsupervised learning context. We introduce the framework of metric spaces and extend the definition of LM and MK dimensions to this framework. We utilize the siamese network to determine the upper bounds of the ability to overfit metric spaces. We empirically observe a linear scaling of LM and MK capacities, verifying our theoretical bounds.

1. Introduction

In order to create robust and interpretable machine learning algorithms, quantitative descriptions must be created to analyze how well an algorithm can perform with a particular dataset and how well it can generalize to new datasets. While statistical learning theory, information theory, and other theoretical frameworks address many of these concerns, these theoretical frameworks are limited to describing discriminative learners, such as SVMs, decision trees [4], and neural networks [11].

Recent successes of unsupervised learning methods from image retrieval [6], natural language processing [7], and transfer learning [8] has created a demand for a new paradigm of machine learning methods. These unsupervised learning methods include autoencoders, generative adversarial networks, and siamese networks. These methods have surpassed its discriminative counterparts in most tasks and has the benefit of portability. The parameters learned from unsupervised learning methods tend to generalize more easily to other tasks [7], saving computation, whereas those of supervised methods are rigid, task-specific, and prone to overfitting.

In this article, we attempt to provide a framework for analyzing the intellectual capacity of such methods. The framework is rooted in the Shannon communication model presented by Friedland [3] and extends the notion of memorizing functions to memorizing the pairwise distance functions of a metric space. The two critical numbers LM and MK dimensions prove to be consistent theoretically and empirically in the context of metric learning.

2. Background

2.1. Metric learning

Learning distributions from few to no data points is still an open problem with many applications in image retrieval [6], robotics, and transfer learning [8]. These methods attempt to learn a *metric space*, in which similar points are closer in distance and and dissimilar points are further apart. Formally, we can define a metric space as follows.

Definition 2.1. (Metric space) Given a set of n points and a distance function $\delta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, a metric space is an ordered pair (X, δ) such that

1. $\delta(x, y) = 0 \iff x = y$
2. $\delta(x, y) = \delta(y, x)$
3. $\delta(x, z) \leq \delta(x, y) + \delta(y, z)$

The distance function can be completely defined by an adjacency matrix $A \in \mathbb{R}^n \times \mathbb{R}^n$ in which each entry measures the pairwise distance between two points, $A_{ij} = \delta(x_i, x_j)$.

One popular technique to learn metric spaces is siamese learning, in which a neural network tries to estimate the similarity metric $\delta(x_i, x_j)$. Recent success in few-shot, multiclass learning has garnered interest in models that attempt to define a metric space rather than predict labels directly.

There is theoretical work that attempts to bound the representational power of metric embeddings techniques. Johnson-Lindenstrauss [9] and Bourgain [10] set bounds for metric embeddings in L_p space through a *distortion* quantity, a measure of how much distances expand and contract in the metric space. FRT sets similar distortion bounds for tree metrics. However, there is not any known work for

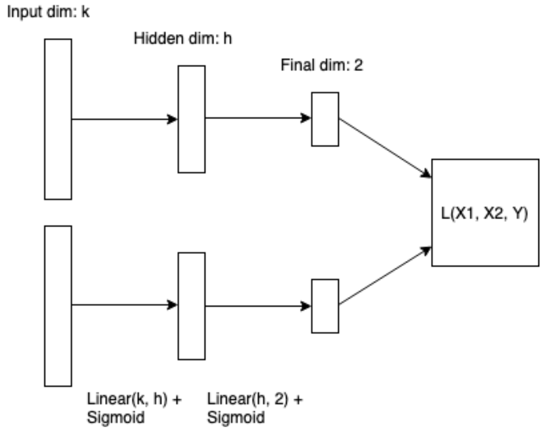

Figure 1. Network architecture of siamese network. Datapoints are projected from the input dimension k to a final dimension of 2. The final output is compared with the contrastive loss function $L(X1, X2, Y)$. Parameters are shared between the two sub-networks.

measurements in the representational power of parameterized metric embeddings such as those generated by neural networks.

2.2. Memory-equivalent capacity

The representational power of discriminative learners have been thoroughly analyzed in many works. Vapnik and Chervonenkis introduced the VC dimension, defined as the largest natural number of samples that can be shattered by a hypothesis space [1]. Rademacher complexity [2], a measure for dataset complexity, has also been used to understand the properties of parameterized discriminative learners like neural networks [11].

In this work, we extend two critical numbers introduced by Friedland [3], the LM and MK dimension.

Definition 2.2. (Lossless Memory dimension) The lossless memory dimension D_{LM} is the maximum integer number D_{LM} such that for any dataset with cardinality $n \leq D_{LM}$ and points in random position, all possible labelings of this dataset can be represented with a function in the hypothesis space.

The LM dimension bounds how well a network can memorize all label configurations. In practice, the LM dimension corresponds to the lowest number of parameters that can completely overfit all instance of D_{LM} random points.

Definition 2.3. (MacKay dimension) The MacKay dimension D_{MK} is the maximum integer D_{MK} such that for any dataset with cardinality $n \leq D_{MK}$ and points in random position at least 50% of all possible labelings of these datasets can be represented with a function in the hypothesis space.

Empirically, the MK dimension corresponds to the lowest number of parameters needed to memorize 50% of D_{MK} random points.

3. Memory-Equivalent Capacity of Metric Learners

In this section, we prove that a metric learner has the same LM capacity as an equivalent discriminative learner. The proof for MK capacity easily follows.

Define a metric learner for a set of D_{LM} points as $g = I(\delta(x_i, x_j) < m)$ where m is the margin constant, I is an indicator function, and $\delta(x_i, x_j) = |f(x_i) - f(x_j)|$ is the distance function. The equivalent discriminative learner is f.

Theorem 1. The LM dimension of a metric learner g is equal the LM dimension of the equivalent discriminative learner f.

Proof. Say we have a set of $\{x_i\}_{i=1}^{n}$ points where $n = D_{LM}$. A perfectly memorized metric learner would be able to memorize an adjacency matrix A where $A_{ij} = g(x_i, x_j) = I(|f(x_i) - f(x_j)| < m)$ and I is an indicator function. Assuming the class distributions are sufficiently apart and concentrated, $|\mu_0 - \mu_1| >> m$ and $m >> \sigma$, such that transitivity holds, $(g(x_i, x_j)$ and $g(x_j, x_k)) \implies g(x_i, x_k)$, a single row in this adjacency matrix determines the rest of the matrix. Memorizing a single row in the adjacency matrix is equivalent to memorizing $A_j = g(c, x_j)$, which has the same LM dimension as a network with the equivalent discriminative learner f. The memorized adjacency matrix can define a metric space that perfectly matches the data. Thus, we have completed our proof. □

Theorem 2. The MK dimension of a metric learner g is equal the MK dimension of the equivalent discriminative

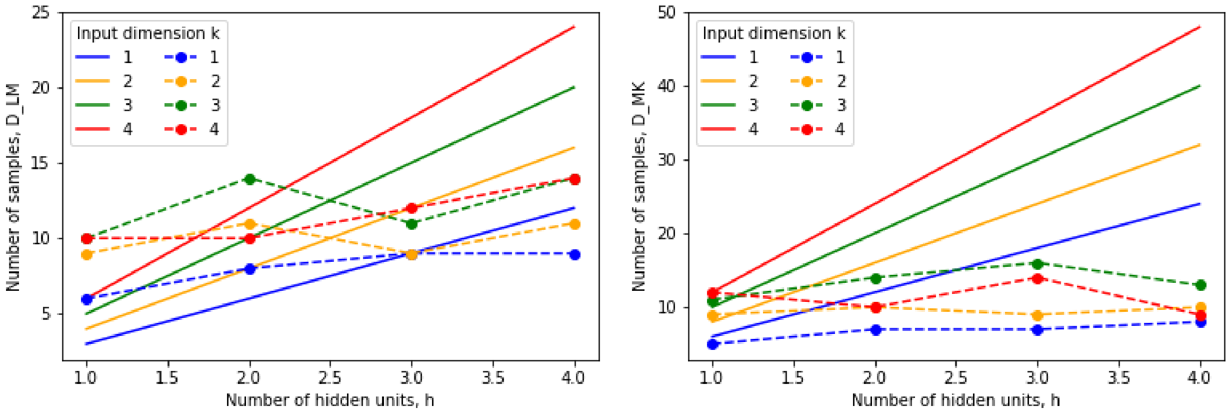

Figure 2. Experimental results for LM dimension (left) and MK dimension (right). Displayed are the functional dependency on k (top) and on h (bottom). The solid lines depict the theoretical boundaries whereas the respective dotted lines display our empirical results. The black lines display the number of samples where not all labelings are tested anymore but a random sample, which makes the empirical results less reliable.

learner f.

This proof follows the same outline as above.

4. Experiments and Results

In order to test the memory-equivalent capacity of a metric learner, we fit a siamese network to toy and real datasets. We empirically verify the LM and MK dimensions of a siamese network by observing the maximum number of label configurations the network can fit completely and 50%, respectively.

4.1. Implementation

A siamese network consists of two networks with shared weights and the contrastive loss function. The networks project two datapoints into the same nonlinear subspace. Once the datapoints are projected, they are passed through a *contrastive loss function*, defined by

$$L(W, Y, X_1, X_2) = \begin{cases} \frac{1}{2}(D_W)^2 & Y = 0 \\ \frac{1}{2}\max\{(0, m - D_W)\}^2 & Y = 1 \end{cases}$$

where label Y indicates whether X_1 and X_2 are similar and D_W is $\delta(X_1, X_2)$. Intuitively, this loss function ”pulls” similar points points together to minimize D_W while the dissimilar points are ”pushed apart” to minimize $m - D_W$.

Since the siamese network is passed in two inputs chosen from a given dataset, we must analyze all pairwise inputs for completeness. Given an input dimension k, there are 2^k possible singleton labelings. Since each set of pairwise labelings, i.e. adjacency matrix, is unique per singleton labeling, there are also 2^k possible pairwise labelings. Due to computation constraints, we sample this search space by sampling the pairwise inputs. We also ensure that the same number of similar pairs are passed in as dissimilar pairs to tackle class-imbalance issues.

4.2. Random Dataset

For the random dataset, we sampled a multivariate Gaussian with input dimensions of size $k = [1, 2, 3, 4]$. For $D_{LM} < 4$, we completely sampled all possible labelings by creating a superset of the $2^{D_{LM}}$ possible label configurations. For $D_{LM} \geq 4$, we sampled 2^3 label configurations from the $2^{D_{LM}}$ sized label space. We note that if $p\%$ of the search space cannot be overfitted, there is a probability of $(1 - 100p)^8$ that the overfitted space would not be found by our method. However, since we are simply trying to establish a relationship between the capacity and the likelihood of overfitting, we find that our empirical method suffices.

The siamese's shared network is a 3-layer fully-connected network with a hidden layer of size h and sigmoid activations, following the tail of the network presented in [siamese network for one-shot learning]. The network was trained with a batch size of 8 over 30 epochs. We use an Adam optimizer with a learning rate of 1.0 and learning rate decay of 0.1 per 10 epochs. For each dataset of $\min(2^{D_{LM}}, 2^3)$ different label configurations, we attempted 10 trails with each trial training a different model over 30 epochs. If none of the 10 models could overfit the particular label configuration, the capacity of the network would be the current number of samples minus one, since we could not overfit the current number of samples.

We found that there is indeed a relationship between the capacity of our model and the ability for the model to overfit. In both LM and MK dimensions, we see empirically that the capacity of our model increases as we increase input and hidden dimensions of our model. We note that we do surpass the theoretical bounds, but this may be due to the fact that we do not check all pairwise relationships when checking for overfitting; rather, we check the relationships between the pairwise inputs sampled during the procedure described in the Implementation section.

We also note that the MK capacity is not significantly higher than the LM capacity, contrary to what was shown in [3]. This may be due to our inability to mimic the label configuration sampling presented in [3]. Due to computational constraints, we use a maximum of 2^3 label configuration, whereas [3] uses 2^{15}. This significantly limits our ability to sample the entire configuration space, which may lead to the empirical inconsistencies we find here.

5. Conclusion

In this report, we extend the notion of LM and MK dimensions to metric learning. We proved that the information capacity of a metric learner is equal to its equivalent discriminative learner. We empirically verified our theoretical results by implementing a siamese network to learn our distance function and determining whether pairwise relationships can be completely learned. This is rather surprising, as there are $\binom{n}{2}$ possible pairwise relationships, meaning that there are $2^{\binom{n}{2}}$ possible label configurations.

Further work could create stronger results bounding the LM and MK capacities to the limit through larger samples and more efficient sampling techniques. There is also work that needs to be done to create the $D_{MK} = 2 * D_{LM}$ result described in Friedland's paper.

All code was implemented on Pytorch in the following repository `https://github.com/jpark96/capacity-of-metric-learners`.

References

[1] V. N. Vapnik. *The nature of statistical learning theory.*. Springer, 2000.

[2] P. L. Bartlett and S. Mendelson. *Rademacher and Gaussian Complexities: Risk Bounds and Structural Results.*. Journal of Machine Learning Research, 3:463482, 2001.

[3] Gerald Friedland, Mario Michael Krell *A Capacity Scaling Law for Artificial Neural Networks.*. Neural and Evolutionary Computing.

[4] O. Asian, O. T. Yildiz, and E. Alpaydin. *Calculating the VC-dimension of decision trees.*. In 24th International Symposium on Computer and Information Sciences, pages 193198. IEEE, sep 2009.

[5] O. Asian, O. T. Yildiz, and E. Alpaydin. *Calculating the VC-dimension of decision trees*. In 24th International Symposium on Computer and Information Sciences, pages 193198. IEEE, sep 2009.

[6] Filip Radenovic, Giorgos Tolias, Ondrej Chum *Fine-tuning CNN Image Retrieval with No Human Annotation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 10.1109/TPAMI.2018.2846566

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever *Language Models are Unsupervised Multitask Learners*. `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf`

[8] Gregory Koch, Richard Zemel Ruslan, Salakhutdinov *Siamese Neural Networks for One-shot Image Recognition*. Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

[9] Sanjoy Dasgupta, Anupam Gupta *An Elementary Proof of a Theorem of Johnson and Lindenstrauss*. 2002 Wiley Periodicals, Inc.

[10] Ohad Giladi, Assaf Naor, Gideon Schechtman *Bourgain's discretization theorem*. `arXiv:1110.5368` [math.FA]

[11] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. *Understanding deep learning requires rethinking generalization.* In International Conference on Learning Representations (ICLR), 2017.