A Small LaTeX Article Template[*]

Your Name
Your Company / University

The Other Dude
His Company / University

June 24, 2017

Abstract

Short introduction to subject of the paper . . .

1 Introduction

The increasing availability of data in diverse domains has created a necessity to develop techniques and methods to discover knowledge from massive volumes of complex data, motivating many research works in databases, machine learning, and information retrieval communities. This has driven the development of scalable and efficient techniques to organize and retrieve this kind of data. Similarity search has been the traditional approach for information retrieval. Although several similarity search algorithms have been proposed to speed up similarity queries, most of them are either affected by the well-known "curse of dimensionality". Retrieve complex data causes stability problems when the data dimensionality is very high [3].

One of the few approaches that ensure an approximate solution with sublinear search cost for high-dimensional data is the Locality Sensitive Hashing (LSH) [1]. LSH is based on the idea that closeness between two objects is usually preserved by a random projection operation. In other words, if two objects are close together in their original space, then these two objects will remain close after a scalar projection operation. However, it presents some difficulties for approximate kNN queries, in particular, related to data domain parameter dependence and quality results. Therefore, in complex domains, in particular, in high dimensional data problems, an approximate solution with a solid theoretical analysis may be the best option in many application areas because of their efficiency in time and space.

On the other hand, in Machine Learning traditionally images are often described by the hand-craft visual features. However, these hand-craft features cannot well reveal the high-level semantic meaning (labels or tags) of images, and often limit the performance of image retrieval [7]. Inspired by recent advances in Convolutional Neural Network (CNN) [6], many methods solved the problem of precision of similarity retrieval by using CNN as feature extractor and then build a compact similarity-preserving hash code for fast image retrieval. Again, hashing is widely used for large-scale image retrieval as well as video and document searches because the compact representation of hash code is essential for data storage and reasonable for query searches [11]. However, some drawbacks based on these supervised hashing methods have not been solved entirely, as follows:

- There is a trade-off between classification error and quantization error: activations of lower layers are more general-purpose [12], so training is more effective. However lower layers have larger activations maps (many nodes), which are harder to encode which leads to a compromise.

- There is a dependency on parameter values for approximate similarity search schemes based on LSH, which determine the number of hash functions and number of hash tables.

[*]To your mother

This paper proposes a novel supervised hashing technique, named Deep frActal based Hashing (DAsH), designed to perform scalable approximate similarity search. The contributions of our work are as follows. First, we introduce and define a scheme based on CNN and optimized using fractal theory. To overcome the limitation of large activations on lower layers of CNN (output of the last convolutional layer) we reduce its dimensionality using autoencoders to the optimal sub-space. Then we index this new representation with LSH scheme. Second, we present a novel method, based on fractal theory, which allow us to can find the optimal number of hash functions for an approximate similarity search scheme based on LSH.

The paper is organized as follows. Section 2 summarizes the background for this work. Section 3 describes the proposed technique and Section 4 reports experimental results on real and synthetic datasets. Finally, we conclude in Section 5.

2 Locality Sensitive Hashing

Previous work [1] has explored the idea of hashing objects and grouping them into buckets with the goal of performing approximate similarity search within buckets associated with the query element. The idea behind LSH is that if two objects are close together in their original space, then these two objects will remain close after a scalar projection operation. Hence, let $h(x)$ be a hash function that maps a d-dimensional point x to a one-dimensional value. The function $h(x)$ is said to be *locality sensitive* if the probability of mapping two d-dimensional points x_1, x_2 to the same value grows as their distance $d(x_1, x_2)$ decreases.

LSH based methods report efficient results when adequate values for m (number of hash functions) and L (number of indexes) are chosen. The E^2-LSH algorithm find the best value for m and L by experimentally evaluating the cost of calculation for samples in the given dataset. Basically, the tuning parameter of LSH is chosen as a function of the dataset to minimize the running time of a query while the space requirement is within the memory bounds [10].

2.1 Fractal Theory

A fractal is characterized by the self-similarity property, i.e., it is an object that presents roughly the same characteristics when analyzed over a broad range of scales [**?**]. From the Fractal Theory, the Correlation Fractal Dimension $\mathfrak{D}$ is particularly useful for data analysis, since it can be applied to estimate the intrinsic dimension of real datasets that exhibit fractal behavior, i.e., exactly or statistically self-similar datasets [4]. It has been shown that, given a set of N objects in a dataset with a distance function $d(x, y)$, the average number of k neighbors within a given distance r is proportional to r raised to $\mathfrak{D}$. Thus, the pair-count $PC(r)$ of pairs of elements within distance r follows the power law:

$$PC(r) = K_p \times r^{\mathfrak{D}} \tag{1}$$

where, K_p is a proportionality constant, and $\mathfrak{D}$ is the correlation fractal dimension of the dataset. Consequently, a fractal is defined by the self-similarity property, that is the main characteristic that represents exactly or statistically the similarity between the parts to the whole fractal.

2.2 Using Fractals to estimate LSH parameters

To tune the LSH parameters we used a property of the correlation fractal dimension $\mathfrak{D}$, which can describes statistically a dataset. Moreover, the correlation fractal dimension $\mathfrak{D}$ can be estimated in linear time as it is depicted in [**?**].

We are interested to find out the resolution scale $log(r)$ at which there are approximately k objects. Considering the line with slope $\mathfrak{D}$ passing at a point defined as $< log(r), log(Pairs(k)) >$ the constant K_d using the Equation 1 is:

$$\begin{gathered} log(PC(r)) = \mathfrak{D} \times log(r) + K_p \\ K_p = log(Pairs(k)) - \mathfrak{D} \times log(r) \end{gathered} \tag{2}$$

Considering another point $< log(R), log(Pairs(N)) >$, the constant K_d is defined as :

$$K_d = log(Pairs(N)) - \mathfrak{D} \cdot log(R) \tag{3}$$

Now, combining Equations 2, 3, we can define the radius r as:

$$r = R \cdot exp(\frac{log(Pairs(k)) - log(Pairs(N))}{\mathfrak{D}}) \tag{4}$$

Using the last equation 4 we find out that the optimal number of hash functions m for a Locality Sensitive Hashing (LSH) based index configured to retrieve the k nearest neighbors is proportional to the number of pairs at a distance r. This has sense, because an average number of k neighbors are within a given distance r. Then, we define:

$$m \approx log(PC(r)) \tag{5}$$

combining equations 5 and 1 we will obtain that $m \approx \mathfrak{D} \cdot log(r)$. Experimentally, we confirm out that the optimal m is:

$$m = (\lceil \mathfrak{D} + 1 \rceil) \cdot log(r) \tag{6}$$

3 Fractal Dimension and Dimensionality Reduction

The approaches used to reduce dimensionality depend on some parametric and non-parametric methods. Other methods reduce the original feature space by heuristic deletion of some of the attributes.

3.1 Principal component analysis

Principal Component Analysis (PCA) involves mathematical procedures that transforms a number of possible correlational variables into a small number of uncorrelated variables called *principal components*. The first component represents as much variability in the data as possible, and each successive component represents as much of the remaining variability as possible.

In PCA, the data is summarized as a linear combination of a set of orthonormal vectors. Let $\{x_i\}_{i=1}^{n}$ a sample of $\mathbb{R}^d$ with mean $\bar{x}$ and covariance $\sum$, with spectral decomposition $\sum = U \Lambda U^T$.
The main transformation component $y = U^T(x - \bar{x})$ produces a reference system in which the sample has a mean 0 and a covariance matrix Λ containing the eigenvalues of $\sum$. Now, we can discard the variables with small variance to project them on a subspace covered by the first main component, and get a good approximation of the original sample. The key property of PCA is to achieve the best linear mapping $x \in \mathbb{R}^d \rightarrow x^* \in \mathbb{R}^m$ of the sum of least squares errors in the reconstruction of data.

3.2 Factor Analysis

Factor analysis is a statistical method for modeling the covariance structure of high dimensional data using a small number of latent variables [**?**]. The data are assumed to be a linear combination of uncorrelated Gaussian sources (factors). After the linear combination, each component of the data vector is also assumed to be corrupted with additional Gaussian noise. (see Figure 1)

The objective of the factor analysis is to find Λ and Ψ which improves the model of the x structure. The factorial variables z of the correlation model between the elements of x, while the variable u counts for an independent noise in each element of x. The m factors play the same role of principal components in the PCA. We first subtract the information from the data and then show the date as:

$$x - \mu = \Lambda z + u \tag{7}$$

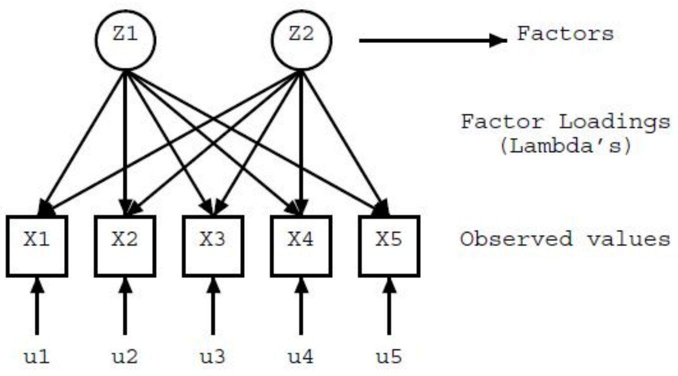


Figure 1: Simple route diagram for a factor analysis model

3.3 Self Supervised Multi Layer Perceptrons

The supervised MLP architecture, or also called autoencoder, implements a mapping using two layers of linear perceptrons with d input, m hidden units and d training outputs to replicate the input to the output layer minimizing the error Of the sum of squares with *back-propagation*. This approach is called *self-supervised*, referring to the fact that during training each vector output sample is identical to the vector input sample.

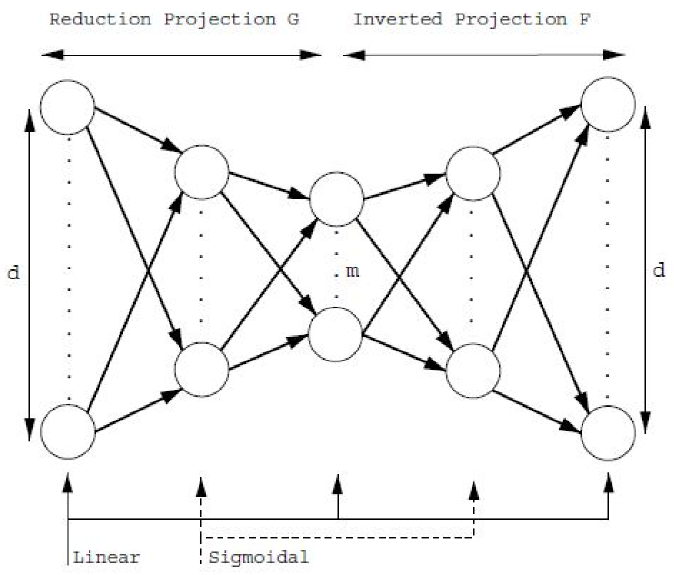


Figure 2: The self-supervised MLP architecture

A botleneck MLP with a unique hidden layer physically performs a linear PCA even with non-linear hidden units. In fact, in order to effectively implement a non-linear dimensionality reduction, the mapping function F and G must be both non-linear. The process of dimensionality reduction consists to find the functions F and G which are approximately functions inverse to each other. Since the inverse of a nonlinear function can not be linear, therefore, if any of the functions is linear, the other must also be linear.

Linear self-supervised MLP can be extended to implement a non-linear PCA, using non-linear activation functions and more hidden layers (see Figure 2).

4 SUPERVISED HASHING: A SIMPLE BASELINE

In this section, we describe the protocols used in the literature for SSH and SH and explain how a simple strategy efficiently solves the similar problems.

4.1 Evaluation protocols of SSH and SH

SSH consists in index a dataset of N images $\mathcal{I}_{train}$, of which a subset $\mathcal{I}_{label} \subseteq \mathcal{I}_{train}$ is labeled. But, SH is the extreme case $\mathcal{I}_{label} = \mathcal{I}_{train}$. If we have an unlabeled query image q, the system have to return an ordered list of images from the $\mathcal{I}_{train}$. To evaluate, we have a dataset of queries; the evaluator knows all labels of the queries as well as the labels in $\mathcal{I}_{train}$, also in the SSH setting, then an image is considered correct if it and query have the same label. The performance is measured by precision or mean average precision (mAP).

Then, given a query q, if the i^{th} image is correct for q we define $\delta(q,i) = 1$, and 0 otherwise. The precision at (rank) k is given by $P(q,k) = \frac{1}{k}\sum_{i=1}^{k}\delta(q,i)$. Denoting by $cl(q) = \sum_{i=1}^{N}\delta(q,i)$ the total number of correct images in $\mathcal{I}_{train}$, and the average precision at k is $AP(q,k) = \frac{1}{cl(q)}\sum_{i=1}^{k}\delta(q,i)P(q,i)$. The mAP at k (or simply mAP when $k = N$) is the mean AP over all test queries.

4.2 Retrieval through class probability estimation

The information recovery [**?**] and learning to rank that the best prediction for precision at k is given by placing items $x \in \mathcal{I}_{train}$ n according to their probability of being correct for the query. This result extends to the optimization of mAP.
Optimal ranking for SH. In the particular setup of SH where the system knows the labels of the images in $\mathcal{I}_{train}$, the probability that a picture x with label y is correct is the likelihood $\mathbb{P}(y \mid q)$ that the query image has label y. The important point in this part is that the probability of x being correct for q only depends on the label of x. Thus, ordering the C labels so that $\mathbb{P}(y \mid q) \geq \cdots \geq \mathbb{P}(c_C \mid q)$, the best ranking is to return the complete dataset of $\mathcal{I}_{train}n$ with label c_1 first, followed by all the pictures with label c_2, and so on.

In practice, $\mathbb{P}(. \mid q)$ is unfamiliar, but we can train a classifier on $\mathcal{I}_{label} = \mathcal{I}_{train}$ which outputs probability con predict $\hat{\mathbb{P}}(c \mid q)$ for every label c, and process the optimal ranking according to $\hat{\mathbb{P}}(c \mid q)$ Such probability estimates are shown by, e.g., multiclass logistic regression or a Convolutional Neural Network (CNN) with a softmax output layer. Labels of $\mathcal{I}_{train}$ are stored on $\lceil log2(C) \rceil$ e bits or in an inverted file

Relationship between classification accuracy and ranking performance. If we show the classification accuracy as p, then the resulting mAP is at least p. When the classifier foretells the class of q properly, all images of that class will be ranked first, and the resulting AP(q) is 1; this happens on a proportion p of the queries. So, the classification accuracy is a lower bound on the mAP.

Optimal ranking for SSH. In the more general setup of SSH, we do not know the label of some images in $\mathcal{I}_{train}$. Yet, considering the (real) probabilities $\mathbb{P}(c \mid q)$ and $\mathbb{P}(c \mid x)$, the probability that x is correct for q is given by $\sum_{c=1}^{C}\mathbb{P}(c \mid q)\mathbb{P}(c \mid x)$ it is the probability that both q and x have the same label, assuming conditional self-sufficiency of the labels of the query and the image. Notice that this is the dot product between the dependent label probability vectors of q and x.$Then, given probability measures for the labels of queries and images, which are taken on \mathrm{I}_{label}$, we consider two retrieval algorithms:

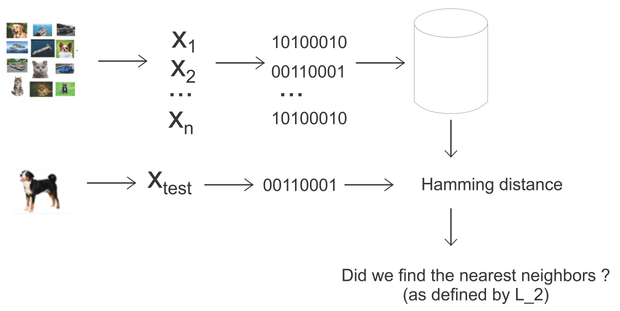
- **Classifier topline:** : For each image $x$of $\mathcal{I}_{train}$, saves a vector $u(x)$ equal to either (1) the one-hot encoding vector of the label of x if $x \in \mathcal{I}_{label}$, or (2) the full vector $\hat{\mathbb{P}}(. \mid x)$). Rank images x according to the dot product $\langle\hat{\mathbb{P}}(. \mid q), u(x)\rangle$ This strategy corresponds to the best strategy, but needs to save the probability vectors for each image in $\mathcal{I}_{train}\backslash\mathcal{I}_{label}$
- **Classifier hashed:** Here we hash the probability vector. The first hashing method that we evaluate, is the one-hot strategy, which saves the index of the maximal activation on

$\lceil log_2(C) \rceil$ bits. This approach, expressed **Classifier+one-hot** in what results, returns all images of the strongest class first. The second encoding, referred to as Classifier+LSH, is locality-sensitive hashing (LSH) with tight frames [5], a simple non data adaptive hashing scheme. This LSH method provides binary vectors that are compared with Hamming distances. Consequently it can be used as drop-in replacements for the competing binary encoding methods.

5 How should we evaluate supervised hashing?

5.1 HASHING FOR NEAREST NEIGHBOR SEARCH

Compress a set of vectors $(x_i)_{i=1}^{b}, x_i \in \mathbb{R}^d$



5.2 SUPERVISED HASHING

Compress a set of vectors and their labels $((x_i, y_2))_{i=1}^{n}, x_1 \in \mathbb{R}^d, y_1 \in \{1, ..., L\}$

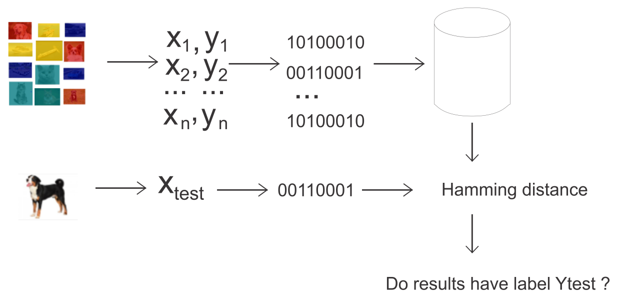


Figure 3: HASHING FOR NEAREST NEIGHBOR SEARCH

- Supervised hashing $[1, 2]$: labels y known for all x in the reference set
- Semi-supervised hashing $[3, 4]$: labels y known for only n_label samples

5.3 WHY NOT JUST ENCODE THE CLASS ID?

Supervised hashing with classification baseline

- Trivial binary encoding of the class id,$e.g. y = 9 \rightarrow 1001$

$$x_1, y_1 \quad 1001 \tag{8}$$
$$x_2, y_2 \quad \rightarrow \quad 1001 \tag{9}$$
$$\ldots \quad \ldots \tag{10}$$
$$x_n, y_n \quad 0011 \tag{11}$$

Table 1: Results on CIFAR-10

Features	n_label	$n_anchors$	Method	bits	mAP
GIST	59,000	1,000	SQ[1]	64	0.704
			SQ[1]	128	0.712
			One-hot	4	0.762
GIST	5,000	1,000	SDH[3]	64	0.402
			One-hot	4	0.377
			LSH	64	0.430
			Topline	-	0.578
GIST	1,000	300	KSH[4]	12	0.232
			KSH[4]	48	0.284
			One-hot	4	0.270
			LSH	48	0.309
			Topline	-	0.350
DEEP	50,000	-	$\vdots$	$\vdots$	$\vdots$
AlexNet			DSH[2]	12	0.616
			DSH[2]	48	0.621
			One-hot	4	0.870

Table 2: Results on ImageNet

Features	Method	bits	mAP@1500
VGG	SQ[1]	128	0.620
VGG	one-hot	10	0.664

- Train classifier on pairs $(x_1, y_1), \dots, (x_n; y_n)$ and predict y_test with the classifier
  - Guaranteed performance: $mAP \geq$ classifier accuracy

5.4 EXTENSION TO SEMI-SUPERVISED HASHING

$$\begin{aligned}\mathbb{P}(x\ correct for\ q) &= \sum_{j=label} \mathbb{P}(j \mid x)\mathbb{P}(j \mid q) \\ &= (\mathbb{P}(j \mid x), \mathbb{P}(j \mid q)) \\ &= \underbrace{(\hat{\mathbb{P}}(j \mid x),}_{classifier} \hat{\mathbb{P}}(j \mid q))\end{aligned} \tag{12}$$

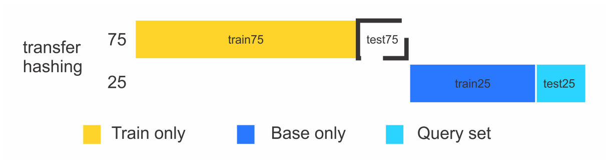
- Train $\hat{\mathbb{P}}$ on labelled images
- Compute $\hat{\mathbb{P}}(\cdot \mid x) \in [0,1]^L$ for x unlabelled
- Compress $\hat{\mathbb{P}}(\cdot \mid x)$ with one-hot / LSH

Encoding schemes

- One-hot:$(0.15, 0.07, 0.08,$ **0.7**$)30011$
- LSH:$(0.15, 0.07, 0.08, 0.7)(sign(w_i^T \cdot v + b_i))_{i=1}^{n_bits}$
- Top line: no compression

5.5 HOW CAN WE AVOID THIS BIASED PROTOCOL?

5.5.



- Test on classes **never seen at train time** classes in 4 folds, each with 25% of classes,Both setups:
  - Train hash functions on train75
  - Encode train25 with hash functions
- **Setup 1: Retrieval with hash codes**
  - Use train25 as reference set
  - Use test25 as queries
- **Setup 2: Classification on hash codes**
  - Train classifier using train25 labels
  - Evaluate accuracy on test25

5.6 UNSUPERVISED BASELINE FOR PROPOSED PROTOCOL

- **Experimental setting**
  - Experiments with CIFAR-10 using AlexNet
  - Unsupervised PQ codes [5] with 4 bytes
- **Setup 1: Retrieval with hash codes**
  - PQ codes with asymmetric comparison
  - Higher layers are better
  - 4 bytes enough for most of performance
  - Inner product on softmax gets the best result
- **Setup 2: Classification on hash codes**
  - Drop in accuracy due to encoding
  - Lower layers more generic$\rightarrow$better accuracy
  - Lower layers high dimensional$\rightarrow$larger gap between PQ and full vector
  - $\rightarrow$ Trade-off encoding/accuracy

6 EVALUATION PROTOCOLS

In this section, we describe the two evaluation tasks proposed in [9], namely retrieval of unseen classes, and transfer learning to new classes. They correspond to application cases on large datasets. The two protocols differ only in the evaluation metric: ranking versus class accuracy.

Dataset definition. at test time we use separate classes from a standard classification dataset. When learning the hashing function 75% of the classes are assumed to be known, and the 25% remaining classes are used to evaluate the encoding/hashing scheme. We call train75/test75 the train/test images of the 75% classes and train25/test25 the remaining ones.

Protocol 1: Retrieval of unseen classes, to index train25 and use test25 as queries, we use the hashing scheme. We use the labels of train25 for evaluation only. This setup is like an instance search approach except that the class labels give the ground-truth. The train75 - train25 - test25 split is the supervised equivalent of the learn - database - query split in unsupervised hashing.

Protocol 2: Transfer learning, the authors proposed to train a new CNN from scratch with the same structure as the top of the original network, using the stored train25 descriptors. The goal is to maximize the transfer accuracy on test25.

7 Experiments

In this section, we are interested in answering the following question: (a) How accurate is our model in estimating the LSH parameters using the fractal dimension; (b) How does our DAsH method improve the other LSH implementations in terms of *querying performance* and *precision*. The performance of DAsH method was compared to two well-known approximate search methods, namely Multi-probe LSH [8], LSH-Forest [2], ITQ [**?**], and LOPQ [**?**]. All of the experiments were performed on a workstation with Intel core i7 3.0Ghz (12 cores) CPU and 64Gb RAM which is supplied with four Geforce GTX 1080 GPU with 8Gb VRAM each one.

We first conduct experiments on eight widely used datasets using hand-crafted features (AUDIO, CITIES, EIGENFACES, HISTOGRAMS, MGCOUNTY, RANDOMWALK, SYNTH16D, SYNTH6D, VIDEO)[1] to evaluate our proposed method for estimating the LSH parameters. Beside hand-crafted features, we also show the effectiveness of our methods when deep features are extracted by the deep Convolutional Neural Networks (CNN), we conduct this experiment on three datasets (MNIST[2], CIFAR-10[3], SVHN[4]) to evaluate our in terms of querying performance, meap average precision (mAP), and precision. The following describes the details of the experiments and results.

7.1 Experiment 1: Tunning LSH Parameters

LSH based methods report efficient results when adequate values for m (number of hash functions) are chosen. To evaluate the effectiveness of the presented approach to tune the LSH parameters using fractal dimension, we worked on a variety of synthetic and real dataset. Table **??** summarizes the main features and parameters of the datasets, including the number of elements N, number of attributes d, their intrinsic (fractal) dimension $\mathfrak{D}$, the LSH parameters computed using two approaches: the Andoni [5] algorithm and our proposal based on fractal dimension, and the total computation time for tune the LSH index (in seconds). The experiment results for the number of hash functions m show that the estimations given by Equation 6 are comparable with those obtained with the E2LSH algorithm proposed by Andoni using up to $10X$ less time.

7.2 Retrieval Performance

The aim of this experiment is to measure the total time spent retrieving the k-nearest neighbor objects. The data structures being compared were tested with an specific values for queries. Thus we use $k = 1000$ when compute the Mean Average Precision(mAP) metric and $k = 25$ when compute the precision metric $(P(\%))$.

[1] https://github.com/joselhuillca/fractal_dataset
[2] http://yann.lecun.com/exdb/mnist/
[3] https://www.cs.toronto.edu/~kriz/cifar.html
[4] http://ufldl.stanford.edu/housenumbers/
[5] http://www.mit.edu/~andoni/LSH/

Table 3: My caption

	SMALL SPACE						ORIG. SPACE
AGNEW							
Dim	16	32	64	128	256	512	8704
Fractal Dim	4.2656	4.5797	5.7044	20.5952	24.7651	24.7651	30.4943
CIFAR10							
Dim	16	32	64	128	256	512	4096
Fractal Dim	1.4436	5.4327	7.2287	23.2534	25.5753	25.5753	1.6494
ISBI							
Dim	16	32	64				4096
Fractal Dim	11.6047	16.1283	16.1283				16.1283
MNIST							
Dim	16	32	64	128	256		800
Fractal Dim	4.3912	4.1542	22.5753	22.5753	22.5753		22.5753
SVHN							
Dim	16	32	64	128			1152
Fractal Dim	6.5841	4.6553	28.3359	28.3359			28.3359

8 Conclusions

References

[1] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122, 2008.

[2] Mayank Bawa, Tyson Condie, and Prasanna Ganesan. LSH forest: self-tuning indexes for similarity search. In *Proceedings of the 14th international conference on World Wide Web*, pages 651–660, Chiba, Japan, 2005.

[3] Christian Böhm, Stefan Berchtold, and Daniel A. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3):322–373, 2001.

[4] Christian Cesar Bones, Luciana A. S. Romani, and Elaine P. M. de Sousa. Clustering multivariate data streams by correlating attributes using fractal dimension. *JIDM*, 7(3):249–264, 2016.

[5] Hervé Jégou, Teddy Furon, and Jean-Jacques Fuchs. Anti-sparse coding for approximate nearest neighbor search. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2029–2032. IEEE, 2012.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[7] Zechao Li, Jing Liu, Jinhui Tang, and Hanqing Lu. Robust structured subspace learning for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(10):2085–2098, October 2015.

[8] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe LSH: efficient indexing for high-dimensional similarity search. In *Proceedings of the International Conference on Very Large Data Bases*, pages 950–961, Vienna, Austria, 2007.

[9] Alexandre Sablayrolles, Matthijs Douze, Hervé Jégou, and Nicolas Usunier. How should we evaluate supervised hashing? *arXiv preprint arXiv:1609.06753*, 2016.

[10] Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*, chapter Locality-Sensitive Hashing Using Stable Distributions, pages 55–67. The MIT Press, 2006.

[11] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. In *CVPR*.

[12] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.