SUPERVISED LEARNING OF LOW-RANK TRANSFORMS FOR IMAGE RETRIEVAL

Çağdaş Bilen, Joaquin Zepeda and Patrick Pérez

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{cagdas.bilen, joaquin.zepeda, patrick.perez}@technicolor.com

ABSTRACT

In this paper we propose a new method to automatically select the rank of linear transforms during supervised learning. Our approach relies on a sparsity-enforcing element-wise soft-thresholding operation applied after the linear transform. This novel approach to supervised rank learning has the important advantage that it is very simple to implement and incurs no extra complexity relative to linear transform learning. Furthermore, we propose a simple Stochastic Gradient Descent (SGD) implementation suitable for large scale learning, where SGD solvers have established themselves as the default workhorse.

We compare our method to various other metric learning techniques in the application of image retrieval. This is one of the remaining few areas where supervised learning of low-rank linear transforms has not been fully exploited. The main reason for this is the lack of adequate datasets that are large enough, and hence we further introduce a new dataset consisting of groups of matching images derived from Cable News Network (CNN) videos using geometric verification and manual selection to find matching frames with adequate variability.

Index Terms— Metric and similarity learning, image retrieval, soft-thresholding, stochastic gradient descent, Cable News Network (CNN) dataset

1. INTRODUCTION

Automated image search and comparison methods have become crucial when searching for relevant images in a large collection, especially with the increasing size of image collections and the related difficulty in manually annotating them. Many systems currently exist that enable such search approaches, including commercial web search engines [1] that admit an image as the query and return a ranked list of relevant web images; copyright infringement detection methods that are robust to image manipulations such as cropping, rotation, mirroring and picture-in-picture artifacts [2]; semantic search systems that enable querying of an unannotated private image collection based on visual concepts (*e.g.*, cat) [3, 4]; object detection systems that are robust to the image background content [5]; automatic face verification methods [6, 7]; and vision-based navigation systems used, for example, as part of the control mechanism of self-driving cars.

One important tool in this panoply of image search applications is the low rank linear transform of the form $\mathbf{Mx}$. Learning such linear transforms from supervised training sets endows metrics and similarity measures such as the Mahalanobis metric or the related inner product with task-tailored qualities. Penalty terms derived from the sparsity inducing ℓ_1-norm (*e.g.*, the trace or ℓ_{21} matrix norms) further exist that are continuous surrogates of the rank and are hence used to determine it automatically.

Indeed low-rank linear transforms have been used extensively in supervised learning. The Fisher vector faces approach [7], for example, jointly learns both a similarity and a metric on the same feature vectors and as part of the same objective for the application of face identification. Approaches that detect whether a pair of face images comes from the same individual (a problem called *face verification*) have also benefited greatly from such linear transforms. Guillaumin *et al.* [6], for example, formulate the metric learning problem for face verification as a logistic regression problem.

Generic image classification methods have equally benefited from supervised, low-rank linear transforms. The work of Weinberger *et al.* [8], for example, uses K-nearest neighbor classifiers based on a Mahalanobis metric learned using a training set of triplets consisting each of a reference image, a matching image, and a non-matching image. Mensink *et al.* [9] extended this approach to derive nearest class mean classifiers that had very low complexity when adding new classes relative to approaches employing one-versus-rest linear classifier that ideally need to be retrained with the addition of every new class.

One exception to the use of supervised measures is *image retrieval*, where the query is an image and the expected response consists of those images containing the same scene (or object), albeit under perspective, lighting or colorimetric transformations. Indeed, the more recent successful approaches for image retrieval employ unsupervised methods to build low-rank linear transforms consisting of subsets of the PCA basis [10, 11, 12] or random projections [2, 13]. One of the main reasons for the lack of supervised linear transforms in image retrieval is the lack of adequate training datasets, as the existing datasets [14, 2, 15] are meant to be used only as evaluation benchmarks and are hence too small and further biased towards specific content. Hence in the present work we introduce and exploit the CNN News Footage dataset that makes it possible to learn linear transforms for the image retrieval task.

In this paper we propose a new metric/similarity learning formulation that uses soft-thresholding as the mechanism to learn the rank of $\mathbf{M}$. Our algorithm uses Stochastic Gradient Descent (SGD), as SGD solvers have established themselves as the workhorse of large scale learning [16, 17, 18]. It has further been established theoretically and empirically that, for a fixed number of iterations, SGD solvers achieve lower generalization cost than batch-based gradient descent methods [18, 19, 20]. It is important to note that existing, SGD-compatible, rank-penalization methods derived from the ℓ_1 norm require complexity-increasing tricks such as decomposing the penalized variables into differences of non-negative variables [20], and that these tricks often result in unstable sparse supports. Our approach does not increase the learning complexity relative to fixed-rank learning (wherein $\mathbf{M}$ is set to be rectangular), yet learns

the rank automatically, similarly to rank-penalized methods. It further benefits from the stability and straight-forwardness of support selection that characterizes soft thresholding-based approximators.

We test our proposed algorithm using existing datasets as well as the new CNN News Footage dataset, comparing it against existing algorithms for retrieval that employ more complex mechanisms to achieve good retrieval performance. Our algorithm is shown to result in very good performance despite the simplicity of the approach and the low-complexity of the learning process.

The rest of this paper is organized as follows: In Section 2, we provide an overview of image descriptors and some methods for distance and correlation learning. We then formally introduce the problem we address in Section 3 and present our proposed rank-learning method in Section 4. In Section 5, we then present the CNN News Footage dataset and evaluate our method experimentally on this dataset and others. Finally, conclusions are discussed in Section 6.

2. BACKGROUND

Many supervised learning methods exist that use different objectives to define their linear transformation $\mathbf{M}$ [21]. We now present several of these that are particularly suited to the image retrieval application that we address in the experiments section.

Discriminative Component Analysis (*DCA*) [22] is a simple distance based metric learning algorithm that is based on maximizing and minimizing the approximate total variances among the pairs of dissimilar and similar items respectively. Chechik *et al.* [23] propose *OASIS*, a proximal stochastic algorithm for an objective based on triplets $(\mathbf{x},\mathbf{y},\mathbf{z})$ that enforces $\mathbf{x}^\top\mathbf{M}\mathbf{y} > \mathbf{x}^\top\mathbf{M}\mathbf{z} - 1 + \xi$ using the common approach wherein slack variables ξ are obtained from the hinge loss. Neither DCA nor the OASIS approach can estimate a low rank linear transformation unless the rank of their learned $\mathbf{M}$ is set by hand which creates an important handicap during the training phase.

Some more recent approaches have addressed this problem by learning the linear transform, $\mathbf{M}$, while enforcing terms that penalize high rank matrices. The *Metric Learning to Rank* (*MLR*) approach of [24, 25] borrows the Structural Support Vector Machine (SSVM) formulation of [26] by letting the ranking over the dataset be the structure to infer and the learned $\mathbf{M}$ be the learned SSVM classifier. The authors constrain $\mathbf{M}$ to be positive definite and further substitute the margin-enhancing ℓ_2 penalty term $\mathrm{Tr}(\mathbf{M}^\top\mathbf{M})$ with the rank-constraining trace norm $\mathrm{Tr}(\mathbf{M})$, thus learning the rank of their metric as part of the optimization. Semi-supervised Metric Learning Paradigm with Hyper Sparsity (*SERAPH*) [27] is a probabilistic approach which is based on maximizing the entropy of the probability distribution of the projected data while also enforcing sparsity in the distribution. Similar to MLR, SERAPH also estimates a positive semi-definite linear transform while penalizing the trace to estimate a low rank transformation. Hence both these methods suffer from having to perform singular value decomposition (SVD) with computational complexity of $O(n^3)$[1] at every iteration which results in poor scalability for the data dimension.

3. PROBLEM DEFINITION

We consider the problem of efficiently comparing and ranking items taken from a large set with a certain distribution. An example of such a setup can be the comparison of images or a set of image features.

[1] n indicates the larger dimension of the matrix $\mathbf{M}$.

Let I represent one of many items that we would like to compare. Let it also be given that every data item I is associated with a data vector $\mathbf{y} \in \mathbb{R}^N$. For the example of comparing images, the items represent the images to be compared whereas the data vectors are features extracted from each image for easier processing. We assume that there exists an unknown scalar similarity function $\mathrm{S}^*(I_1, I_2)$ for every item pair (I_1, I_2). Even though the function $\mathrm{S}^*()$ is not known, a set of constraints over $\mathrm{S}^*()$ and a training set $\{I_i, i = 1,\ldots,T\}$ (with corresponding data vectors $\mathbf{y}_i \in \mathcal{T}$) are assumed to be known. In this paper we consider pairwise constraints, each of which is defined by a pair of data points and an indicator variable as

$$\{C_{\text{pair},p}\}_{p=1}^m = \{(i_p, j_p, \gamma_p) \,|\, \gamma_p = \left\{ \begin{array}{c} 1 \text{ if } \mathrm{S}^*(I_{i_p}, I_{j_p}) \geq s_c \\ -1 \text{ if } \mathrm{S}^*(I_{i_p}, I_{j_p}) < s_c \end{array} \right\} \quad (1)$$

for an unknown constant s_c so that the variable γ_p is 1 if two data points are sufficiently similar (or in the same cluster) and -1 if not. Such pairwise constraints are relevant to a task such as matching. Without loss of generality, we define the task of matching as finding a function, $\mathrm{S}(\mathbf{v}_1, \mathbf{v}_2)$, given the constraints $\{C_{\text{pair},p}\}_{p=1}^m$, such that for any given pair of query items and their corresponding data vectors, $(\mathbf{y}_{q_1}, \mathbf{y}_{q_2})$, the function $\mathrm{S}()$ satisfies $\mathrm{S}(\mathbf{y}_{q_1}, \mathbf{y}_{q_2}) \geq 0$ if $\mathrm{S}^*(I_{q_1}, I_{q_2}) \geq s_c$ and $\mathrm{S}(\mathbf{y}_{q_1}, \mathbf{y}_{q_2}) < 0$ otherwise.

The task of matching can be described as determining whether a given query data belongs to a cluster in a dataset. For example, in face recognition systems, the given facial picture of a person is compared to other facial data of the same person within the database to perform verification. It is also possible to perform matching between two given data points even though these points belong to a cluster different from the observed clusters in the database.

4. PROPOSED METHOD

In this paper, we propose a new approach for learning Mahalanobis metric transformations for the purpose of matching or ranking images. The proposed approach, which is called DAta SHrinking for metric learning (DASH in short), can automatically learn a low-rank Mahalanobis metric transformation with low computational complexity and can be easily used for large scale datasets. The proposed approach can be utilized both for Mahalanobis metric learning with distance among the pairs computed as

$$D_{\mathbf{M}}(\mathbf{y}_1, \mathbf{y}_2) = (\mathbf{y}_1 - \mathbf{y}_2)^\top \mathbf{M}^\top \mathbf{M} (\mathbf{y}_1 - \mathbf{y}_2) \quad (2)$$

or with correlation among the pairs computed as [2]

$$S_{\mathbf{M}}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{y}_1^\top \mathbf{M}^\top \mathbf{M} \mathbf{y}_2. \quad (3)$$

4.1. Data Shrinking for Distance Metric Learning (DASH-Distance)

Let us assume a set of pairs of data points (each pair known to be similar or dissimilar) is given as described in (1). In order to learn a Mahalanobis metric transformation matrix $\mathbf{M}$ that can be used to estimate the distances between any given two data descriptors as in (2), we propose to minimize an objective function of the form:

$$\mathbf{M}_{\text{dist}} = \underset{\mathbf{M}}{\text{argmin}} \sum_{p=1}^{m} \ell\left(-\gamma_p \|\mathbf{z}_p\|^2 + d(\gamma_p - 1)/2\right) \quad (4)$$
$$\text{s.t. } \mathbf{z}_p = \mathrm{S}_\lambda(\mathbf{M}(\mathbf{y}_{i_p} - \mathbf{y}_{j_p})),$$

[2] Throughout this paper we assume the descriptors have an additional last row of value 1 so that a translation vector is also learned within $\mathbf{M}$, which is essential for the correlation formulation.

where d is the desired separation among the distances of similar pairs and dissimilar pairs. The function $\ell()$ can be set as the hinge loss defined as $\ell(x) = \max(-(x-\alpha),0)$ or the logistic loss that is equal to $\ell(x) = \log(\exp(-\alpha x)+1)/\alpha$, in both of which the parameter $\alpha > 0$ controls the value of penalty at $x = 0$ that is needed to avoid an all-zero solution for $\mathbf{M}$. Note that when the function $S_\lambda()$ is the identity operator, the optimization problem defined in (4) can be solved to learn a Mahalanobis metric operator $\mathbf{M}$ from the given set of pairs. The large distances between the pairs of similar items as well as the small distances between the pairs of dissimilar items are penalized during the optimization, where the parameter d controls the separation between the distances among the similar pairs and the distances among the dissimilar pairs. In this case, a low rank $\mathbf{M}$ can be learned either by specifying the dimensions of the matrix manually, or including penalization terms that are known to enforce low rank transformations as described in Section 2.

As an alternative, we propose to use a sparsifying function as S_λ such that $S_\lambda(\mathbf{x}) \approx \mathbf{x}$ and $S_\lambda(\mathbf{x})$ is sparse. An example of such a function is the element-wise soft thresholding function defined component-wise as $S_\lambda(x) = \max(|x|-\lambda,0)\text{sign}(x)$ with a sparsifying parameter λ, which is in fact the solution to the optimization problem $\text{argmin}_\mathbf{z} \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1$. The motivations of using such an approach is discussed in more detail in Section 4.3.

4.2. Data Shrinking for Correlation Metric Learning (DASH-Correlation)

Similar to learning linear projection for distances, it is also possible to apply the proposed approach for learning linear projection for correlations. For this purpose, similar to (4), we propose to minimize an objective function of the form:

$$\mathbf{M}_{\text{corr}} = \underset{\mathbf{M}}{\text{argmin}} \sum_{p=1}^{m} \ell(\gamma_p \mathbf{z}_{j_p}^\top \mathbf{z}_{i_p}) \text{ s.t. } \mathbf{z}_t = S_\lambda(\mathbf{M}\mathbf{y}_t),\ t = 1,\ldots,T \quad (5)$$

4.3. Motivation for using a Sparsifying Function and Other advantages

The proposed algorithms rely on the fact that, when the projected descriptors (in DASH-Correlation) or the difference of projected descriptors (in DASH-Distance) are constrained to be sparse while learning $\mathbf{M}$, the energy of the projected coefficients are forced to lie in a subset of the support. Therefore the rows of the projected coefficients that are the least frequently utilized can inherently be removed resulting in a low rank $\mathbf{M}$. Some other advantages of the proposed approaches are as follows:

- Even though the objective functions are non-convex, both of the optimizations in (4) and (5) can be easily performed with stochastic gradient descent (SGD) [20] since the required sub-gradient is easy to compute for any given pair of data points. Therefore the optimizations can be carried out even when the projections are learned over very large datasets.[3]
- The linear transformation, $\mathbf{M}$, is not constrained for scale. Hence no matter the value of sparsifying parameter λ, the sparsity of the result depends only on the best possible performance reached by the optimization. Therefore the proposed algorithms are not sensitive to the selection of λ. If so desired, a constraint on the norm of $\mathbf{M}$ can be added to better control the resulting sparsity of the resulting coefficients.
- A low rank projection can easily be obtained without specifying the rank manually prior to the optimization or without any constraints on the singular values of the projection matrix during learning stage. Furthermore, the rows of the learned projection matrix inherently contribute in varying degrees to the distance (or correlation) estimation, and this degree of contribution can be determined after the learning stage. This property not only helps to determine the optimum dimensionality reduction best suited for the dataset, but also makes it possible to change the desired rank after the optimization without the need to learn $\mathbf{M}$ from scratch.

[3]Thanks to the use of SGD, the computational complexity and memory requirements of the algorithm per iteration is significantly low, however this does not mean that the overall speed of the algorithm is higher than other approaches, since the number of iterations required for convergence can be much higher. However SGD results in a more flexible algorithm that can be used with large databases where other algorithms may not be feasible.

Fig. 1: Example images from CNN News Footage dataset, the images labeled to be in the same group are shown with the same color of background.

5. CNN DATASET AND THE EXPERIMENT RESULTS

CNN News Footage dataset. In order to test the performance of the metric learning algorithms, we introduce a new dataset called CNN News Footage dataset. The CNN News Footage dataset is a public dataset composed of images extracted from videos of CNN news broadcasts over the years. The images extracted from a footage of the same scene in a news report are grouped together under a single label and marked as similar. The dataset is composed of 17000 images in 5500 groups with various images being related to each other by camera actions such as panning and zooming, as well as objects being shot from multiple angles and occluded at times. Some example images from the CNN News Footage dataset can be seen in Figure 1. As a comparison, the Holidays dataset [2] which is composed of images of similar properties contains around 1500 images and 500 groups which is too small for supervised learning. [4]

CNN News Footage splits. For testing the algorithms, 200 groups are randomly selected as test groups while the rest of the groups are used to generate 55000 training pairs. VLAD feature vectors [10] with PCA dimension of 512 are used as image descriptors. The pairs of similar items for training are taken to be pairs of images from the same group, while pairs of dissimilar items for training are chosen among the image pairs from different groups with the most similar descriptors. Two separate test sets of pairs are used for evaluation of algorithm performance. The pairs of similar items for both test sets are generated only from the 200 test groups. The pairs of dissimilar items of the first test set are selected among the image pairs with most similar descriptors within the 200 test groups

[4]The CNN News Footage dataset can be obtained by contacting the authors of this paper.

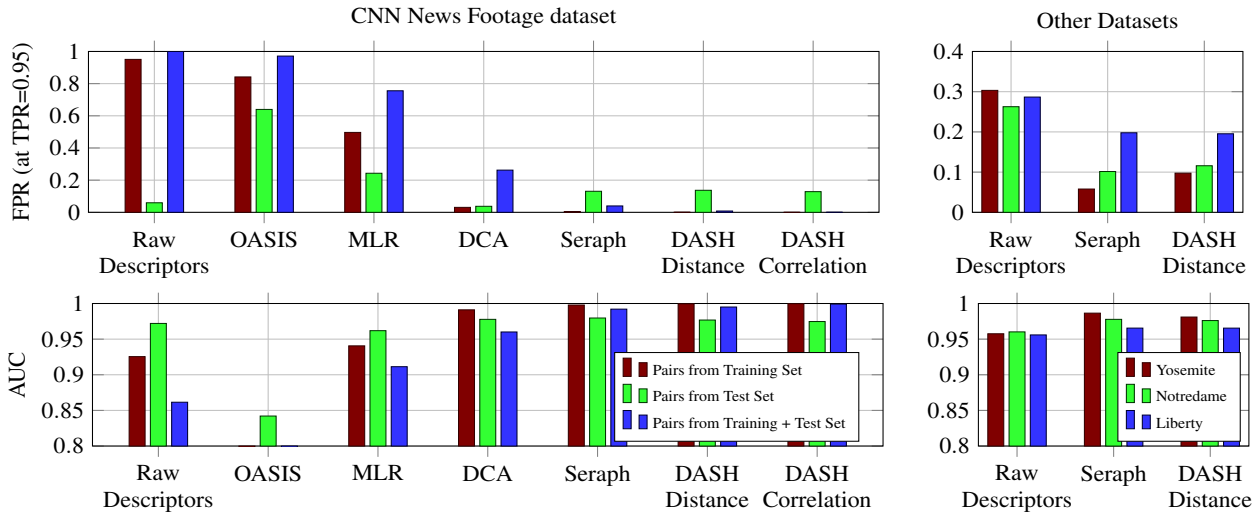

Fig. 2: Comparison of proposed methods (*DASH*) against other metric learning algorithms with respect to False Positive Rate (FPR) (measured at True Positive Rate of 0.95) and Area Under Curve (AUC).

whereas the pairs of dissimilar items of the second test set are selected among the image pairs with most similar descriptors within the entire 5500 groups.

Other datasets. We have also compared the performance of best performing algorithms in CNN News Footage dataset on Yosemite, Notredame and Liberty datasets given in [28]. Dense SIFT descriptors of length 1152 (composed of 9 blocks for each SIFT descriptor) are used as image descriptors. The algorithms are trained using 500000 training pairs from Yosemite dataset and then tested by 100000 test pairs from all three datasets.

Baseline algorithms. The performance of the proposed algorithms, *DASH-Distance* and *DASH-Correlation*, are compared against the methods discussed in Section 2 (OASIS [23], DCA, [22], MLR [24] and SERAPH [27]). These methods are chosen because they carry out supervised learning of linear transforms for general ranking applications, and hence they are immediately adapted to the supervised image retrieval task we address. Note that, in all cases, we use the code made available by the authors. For the fixed rank algorithms (DCA and OASIS), the rank of the linear transform is manually selected whereas for the rank learning algorithms (MLR and SERAPH) the parameters are adjusted so that the learned linear transform is at the desired rank. For the proposed approaches the rank of the linear projection is adjusted after the learning stage. The target rank for CNN News Footage dataset is selected as 330 while the target rank for the other datasets are selected as 400.

For the *DASH* algorithm, the parameters α for the log-loss and the separation distance d are set empirically through a number of Monte-Carlo simulations. The soft thresholding parameter, λ is also selected empirically, however the algorithm is noted to be not sensitive to the selection of this parameter as described in Section 4.3.

Comparison of the performances of all the methods on the CNN database can be seen in the left column of Figure 2. The performance of using raw descriptors are also provided for comparison. The performance is measured in the criterion of false positive rate (corresponding to a true positive rate of 0.95) on top and in the criterion of area under curve (AUC) on the bottom. It can be seen in the results that the performance of *SERAPH* and *DASH* (both correlation and distance) are better than the other methods in general and very close to each other. The performance of *DCA* is also close, however this algorithm requires specifying manually the dimensionality reduction of the projection matrix. The comparison of *SERAPH* and *DASH* for the Yosemite, Notredame and Liberty datasets is also shown in the right column of Figure 2.

It is observed that the low rank projections learned on the CNN News Footage dataset do not necessarily perform well on other benchmark datasets such as the ones introduced in [2, 14, 15], which due to the different overall characteristics of the images in the datasets. The performance results within the CNN News Footage dataset also suggest that the retrieval task is not very challenging for the tested algorithms. In order to overcome this limitation, we consider extending the CNN News Footage dataset with more images having higher diversity and greater retrieval difficulty as future work.

All the presented results show that the proposed approach can match the best performance among the other methods in the literature with the added benefit of simple and low complexity optimization as well as automatic learning of the dimensionality reduction best suited for the data at hand. Even though the proposed optimization problems are non-convex, the minimization is observed to be stable. Furthermore, the proposed optimization approach is robust to the selection of some of the parameters such as the sparsifying threshold, λ, since the scale of the learned projection is not constrained and the best compromise of the scale is automatically learned to balance the resulting dimensionality reduction and sparsity *vs.* the matching performance for the training pairs.

6. CONCLUSIONS

In this paper we have presented a new approach to learn linear projections for image matching and ranking. The proposed approach can learn the dimensionality reduction in the projection best suited for the task without any computationally complex steps during learning such as singular or eigenvalue decomposition. It can also be easily applied for large datasets thanks to the use of stochastic gradient descent with very computationally cheap updates.

A new dataset for image ranking called CNN News Footage dataset is also introduced, which has a much larger size than any other comparable dataset that currently exists. Extending this dataset to have greater diversity in order to provide more challenge is considered as future work.

7. REFERENCES

[1] Google Inc., "Google Image Search," . 1

[2] Hervé Jégou, Matthijs Douze, and Cordelia Schmid, "Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search," in *European Conference on Computer Vision*, 2008. 1, 3, 4

[3] Google Inc., "Google Plus Photos," . 1

[4] Ken Chatfield and Andrew Zisserman, "VISOR : Towards On-the-Fly Large-Scale Object Category Retrieval," . 1

[5] Josef Sivic and Andrew Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, 2003, pp. 2–9. 1

[6] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? Metric learning approaches for face identification," in *International Conference on Computer Vision*, Sept. 2009. 1

[7] Karen Simonyan, Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Fisher Vector Faces in the Wild," in *British Machine Vision Conference*, 2013. 1

[8] Kilian Q. Weinberger and Lawrence K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *The Journal of Machine Learning Research*, pp. 207–244, 2009. 1

[9] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka, "Distance-based image classification: generalizing to new classes at near-zero cost.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2624–37, Nov. 2013. 1

[10] Jonathan Delhumeau, Philippe-Henri Gosselin, Hervé Jégou, and Patrick Pérez, "Revisiting the VLAD image representation," in *Proceedings of ACM International Conference on Multimedia*, New York, New York, USA, 2013, vol. 21, pp. 653–656, ACM Press. 1, 3

[11] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier, "Large-scale Image Retrieval with Compressed Fisher Vectors," in *Computer Vision and Pattern Recognition*. June 2010, pp. 3384–3391, Ieee. 1

[12] Hervé Jégou and Andrew Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Computer Vision and Pattern Recognition*, 2014. 1

[13] Hervé Jégou, Florent Perronnin, Matthijs Douze, Sánchez Jorge, Pérez Patrick, and Cordelia Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2011. 1

[14] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition*, 2007. 1, 4

[15] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," 2006. 1, 4

[16] Alex Krizhevsky, I. Sutskever, and Geoffrey Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Neural Information Processing Systems*, 2012, pp. 1–9. 1

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe : Convolutional Architecture for Fast Feature Embedding," in *ACM Multimedia*, 2014. 1

[18] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro, "Pegasos : Primal Estimated sub-GrAdient SOlver for SVM," in *International Conference of Machine Learning*, 2007. 1

[19] Shai Shalev-Shwartz and Nathan Srebro, "SVM Optimization : Inverse Dependence on Training Set Size," in *International Conference on Machine Learning*, 2008. 1

[20] Leon Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, Grégoire Montavon, Geneviève Orr, and Klaus-Rober Müller, Eds., vol. 1. Springer, 2 edition, 2012. 1, 3

[21] Brian Kulis, "Metric learning: A survey," *Foundations and Trends in Machine Learning*, 2012. 2

[22] Steven C H Hoi, Wei Liu, Michael R. Lyu, and Ma Wei-Ying, "Learning distance metrics with contextual constraints for image retrieval," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, no. c, pp. 2072–2078, 2006. 2, 4

[23] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio, "Large Scale Online Learning of Image Similarity Through Ranking," *J. Mach. Learn. Res.*, vol. 11, pp. 1109–1135, 2010. 2, 4

[24] Brian Mcfee and Gert Lanckriet, "Metric Learning to Rank," *Icml*, pp. 775–782, 2010. 2, 4

[25] Joseph J. Lim, C. Lawrence Zitnick, and Piotr Dollar, "Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3158–3165, June 2013. 2

[26] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, "Large Margin Methods for Structured and Interdependent Output Variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005. 2

[27] Gang Niu, Bo Dai, Makoto Yamada, and Masashi Sugiyama, "SERAPH: Semi-supervised Metric Learning Paradigm with Hyper Sparsity," , no. September 2015, 2011. 2, 4

[28] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sept. 2014. 4