

HOW SHOULD WE EVALUATE SUPERVISED HASHING?

Alexandre Sablayrolles

Matthijs Douze

Nicolas Usunier

Hervé Jégou

Facebook AI Research

ABSTRACT

Hashing produces compact representations for documents, to perform tasks like classification or retrieval based on these short codes. When hashing is supervised, the codes are trained using labels on the training data. This paper first shows that the evaluation protocols used in the literature for supervised hashing are not satisfactory: we show that a trivial solution that encodes the output of a classifier significantly outperforms existing supervised or semi-supervised methods, while using much shorter codes. We then propose two alternative protocols for supervised hashing: one based on retrieval on a disjoint set of classes, and another based on transfer learning to new classes. We provide two baseline methods for image-related tasks to assess the performance of (semi-)supervised hashing: without coding and with unsupervised codes. These baselines give a lower- and upper-bound on the performance of a supervised hashing scheme.

1. INTRODUCTION

Traditional hashing aims at reproducing a target metric, such as the cosine similarity, based on compact codes. Large databases are then be stored memory and queried efficiently. Algorithms are tuned on a *learn* set, and evaluation is done by looking for nearest neighbors of a *query* set in a *database*.

Recent works have proposed to use annotated data, in the form of labeled images, to improve the hashing quality. Indeed, even if hash codes yield noisy reconstructed vectors, being able to discriminate classes from these reconstructions is a desirable property. In the literature, the proposed evaluation protocol for this property involves two datasets: a *train* set, with known or partially known labels and a *query* set with unknown labels. The true positives are defined as images from the train set belonging to the same class as the query. This setting will be referred to as semi-supervised hashing (SSH).

In the particular case of supervised hashing (SH), the labels of the train set are known. We can train a classifier and use it to classify queries, returning for each query all elements of the assigned class. This trivial baseline, which is not considered in overlooked in most works published on the topic, outperforms state-of-the-art methods. This shows that the evaluation protocol is flawed: it only requires to discriminate between known classes and not to reconstruct vectors in a semantically meaningful way.

This paper makes the following contributions:

- We show that both SH and SSH are well addressed by a trivial encoding on the output of a classifier, which outperforms the results reported in the literature.
- We propose two tasks and corresponding baselines to assess the performance of (semi-)supervised hashing: transfer to retrieval and transfer learning. They correspond to real use-cases.
- We show that, in the case of transfer learning, it is better to insert the layer producing a compact code in the middle of the network. In contrast, existing methods routinely encode the last activation layer.

2. RELATED WORK

We distinguish three classes of methods for supervised hashing: triplet loss hashing [1, 2, 3, 4], pairwise similarity-based and label-based. Often, pairwise similarity and label information are equivalent, because pairwise similarity is defined as sharing the same label, and reciprocally labels are equivalence classes of pairwise similarity relations. However they are treated differently because constructing pairwise similarity matrices scales quadratically with the number of labeled samples, limiting these algorithms to small labeled sets.

Pairwise similarity based. Binary Reconstructive Embeddings (BRE) [5] minimize the distortion between the distance matrix in the original space and the Hamming distances of the codes. BRE is extended to the supervised case by replacing the distance matrix by a pairwise similarity matrix. Following this work, different techniques use both the similarity matrix and the feature space: semi-supervised hashing [6], supervised hashing with kernels (KSH) [7], Semantic Hashing [8], Minimal Loss Hashing [9], fast supervised hashing [10].

Label based. Supervised discrete hashing (SDH) [11] and supervised quantization (SQ)[12] integrate the labels in a classification loss, along with a hashing loss. Recent work also explore deep architectures [13, 14] and augmented Lagrangian [15] for supervised hashing.

Transfer learning. Indexing based on attributes or unrelated classes is standard [16, 17]. Torresani et al. [17] remark that, “Without the novel-category requirement, the problem is trivial: the search results can be precomputed by running the known category detector on each database image [...] and

storing the results as inverted files”. We apply this remark to hashing and extend it to the semi-supervised setting.

3. SUPERVISED HASHING: A SIMPLE BASELINE

This section describes the protocols used in the literature for SSH and SH, and discuss how a simple strategy efficiently solves the corresponding problems.

3.1. Evaluation protocols of SSH and SH

The task of SSH consists in indexing a dataset of N images $\mathcal{I}_{\text{train}}$, of which a subset $\mathcal{I}_{\text{label}} \subseteq \mathcal{I}_{\text{train}}$ is labeled. SH is the extreme case $\mathcal{I}_{\text{label}} = \mathcal{I}_{\text{train}}$. Given an unlabeled query image q , the system must return an ordered list of images from the $\mathcal{I}_{\text{train}}$. For evaluation purposes, a dataset of queries is given; the labels of the queries as well as all labels in $\mathcal{I}_{\text{train}}$ are known to the evaluator, even in the SSH setting, and an image is deemed correct if it has the same label as the query. The performance is measured in terms of precision or mean average precision (mAP), which we now describe.

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

3.2. Retrieval through class probability estimation

It is well known in information retrieval [18] and learning to rank that the optimal prediction for precision at k is given by ranking items $x \in \mathcal{I}_{\text{train}}$ according to their probability of being correct for the query. This result extends to the optimization of mAP.

Optimal ranking for SH. In the specific setup of SH where the system knows the labels of the images in $\mathcal{I}_{\text{train}}$, the probability that an image x with label y is correct is the probability $\mathbb{P}(y|q)$ that the query image has label y . The important point here is that the probability of x being correct for q only depends on the label of x . Thus, ordering the C labels so that $\mathbb{P}(c_1|q) \geq \dots \geq \mathbb{P}(c_C|q)$, the *optimal* ranking is to return all images of $\mathcal{I}_{\text{train}}$ with label c_1 first, followed by all images with label c_2 , and so on.

In practice, $\mathbb{P}(\cdot|q)$ is unknown, but we can train a classifier on $\mathcal{I}_{\text{label}} = \mathcal{I}_{\text{train}}$ which outputs probability estimates $\hat{\mathbb{P}}(c|q)$ for every label c , and compute the optimal ranking according to $\hat{\mathbb{P}}(\cdot|q)$. Such probability estimates are given by, e.g., multiclass logistic regression or a Convolutional Neural Network (CNN) with a softmax output layer. Labels of $\mathcal{I}_{\text{train}}$ are stored on $\lceil \log_2(C) \rceil$ bits or in an inverted file [17].

Relationship between classification accuracy and ranking performance. If the classifier has classification accuracy p , then the resulting mAP is *at least* p : whenever the classifier correctly predicts the class of q , all images of that class will be ranked first and the resulting $\text{AP}(q)$ is 1; this happens on a proportion p of the queries. Thus 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 $\mathcal{I}_{\text{train}}$. Yet, considering the (true) conditional label probabilities $\mathbb{P}(c|q)$ and $\mathbb{P}(c|x)$, the probability that x is correct for q is given by $\sum_{c=1}^C \mathbb{P}(c|q)\mathbb{P}(c|x)$: it is the probability that both q and x have the same label, assuming conditional independence of the labels of the query and the image. Notice that this is the dot product between the conditional label probability vectors of q and x . Then, given probability estimates $\hat{\mathbb{P}}$ for the labels of queries and images, which are obtained on $\mathcal{I}_{\text{label}}$, we consider two retrieval algorithms:

Classifier topline: For each image x of $\mathcal{I}_{\text{train}}$, store a vector $u(x)$ equal to either (1) the one-hot encoding vector of the label of x if $x \in \mathcal{I}_{\text{label}}$, or (2) the full conditional probability vector $\hat{\mathbb{P}}(\cdot|x)$. Rank images x according to the dot product $\langle \hat{\mathbb{P}}(\cdot|q), u(x) \rangle$. This strategy corresponds to the optimal strategy, but requires storing the probability vectors for images in $\mathcal{I}_{\text{train}} \setminus \mathcal{I}_{\text{label}}$.

Classifier hashed: Here we hash the conditional probability vector. The first hashing method that we evaluate, is the one-hot strategy, which stores the index of the maximal activation on $\lceil \log_2(C) \rceil$ bits. This approach, denoted **Classifier+one-hot** in what follows, returns all images of the strongest class first. The second encoding, referred to as **Classifier+LSH**, is locality-sensitive hashing (LSH) with tight frames [19], a simple non data-adaptive hashing scheme. This LSH method produces binary vectors that are compared with Hamming distances. Therefore it can be used as drop-in replacements for the competing binary encoding methods.

4. EXPERIMENTS: CLASSIFIERS ON SH AND SSH

Here we experiment with the classifier based retrieval method on SSH. We use off-the-shelf classifiers, whose accuracies are not necessarily the current state of the art. However, we show that they perform better than SSH methods of the literature. We consider two datasets: CIFAR10 and ImageNet.

CIFAR10 [20] is a dataset of 32x32 color natural images that contains 60,000 images divided into 10 classes. For deep methods, we compare against DSH in the SH setting, and use the provided train-test split of CIFAR10 to train an AlexNet. For non-deep methods, we follow the GIST-based protocol of [7, 11, 12]. We hold out 1,000 query images (100 per class)

Table 1. Retrieval (mAP): CIFAR10, SH and SSH protocols.

Features	n_{label}	h	Method	bits	mAP
GIST (SH)	59,000	1,000	SQ [12]	64	0.704
			VDSH [14]	16	0.650
			SQ [12]	128	0.712
			Classifier+one-hot	4	0.762
GIST (SSH)	5,000	1,000	SDH [11]	64	0.402
			Classifier+one-hot	4	0.377
			Classifier+LSH	64	0.430
			Classifier topline	-	0.578
GIST (SSH)	1,000	300	KSH [7]	12	0.232
			KSH [7]	48	0.284
			Classifier+one-hot	4	0.270
			Classifier+LSH	48	0.309
			Classifier topline	-	0.350
Deep (SH) AlexNet	50,000	-	DSH [21]	12	0.616
			DSH [21]	48	0.621
			Classifier+one-hot	4	0.870

and index the remaining 59,000 images, a variable number n_{label} of them being labelled (following the experimental protocols of the papers we compare with).

We start from the 512D GIST descriptors of the images; then we sample h of the labeled images $(a_i)_{i=1}^h$ as anchors. Images are represented by their Gaussian features $[\exp(-\|x - a_i\|_2^2/2\sigma^2)]_{i=1..h} \in \mathbb{R}^h$, with $\sigma = \frac{1}{N} \sum_{i=1}^N \min_{j=1, \dots, h} \|x_i - a_j\|_2$. We fit a Logistic Regression classifier on these features. We cross-validate the regularization parameter on 10% of the train set.

Results are shown in Table 1. We compare our approach to methods in the literature, using the numbers reported in the cited papers. With $\lceil \log_2(C) \rceil$ bits, mAP results are almost as good as the state of the art, while being 4-8 times more compact. With the same code size, simple LSH encodings outperform competing methods by a large margin.

Although the work of [22] was both deep and a SSH setup, their evaluation metric differs from our definition of mAP and thus we have not included them in our comparison.

ImageNet (ILSVRC 2012), contains over 1.2 million natural images of 1,000 categories [23]. The training set is used for learning and indexing purposes. The 50,000 validation images are used as queries. In [12], the images are represented by activations of a VGG16 network [24].

The authors of [12] experiment on ImageNet, in the SH setting. They use a CNN trained on ImageNet, and then use the train labels to train their quantization method. We use the same classifier from Caffe (VGG16) to classify query images, and store train labels as 1-hot vectors. The results in Table 2 show that our baseline method is more accurate and an order of magnitude more compact.

Table 2. Results on ImageNet with the SH protocol.

Descriptors	Method	bits	mAP @1500
VGG	SQ [12]	128	0.620
VGG	Classifier+one-hot	10	0.664

5. PROPOSED EVALUATION PROTOCOLS

The previous section has shown that existing protocols fail to capture desirable properties of supervised hashing schemes. In this section, we propose two evaluation tasks, namely retrieval of unseen classes, and transfer learning to new classes. They correspond to application cases on large and growing user-generated datasets, where classifiers are trained on fluctuating training sets or for new labels. For computational reasons the features cannot be recomputed when the classes evolve, and mid-level features must be compressed in a way that preserves their semantic information. The two protocols we consider differ only in the evaluation metric: ranking versus class accuracy.

Dataset definition. In both tasks, we start from a standard classification dataset but we use separate classes at test time, similar to a metric learning setup [25]. 75% of the classes are assumed to be known when learning the hashing function, 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.

In practice, we shuffle the classes randomly and use 4 folds to define 4 of those splits. Performance measures are averaged over the folds. We split the classes of an existing dataset rather than combining different datasets, because the latter would introduce a lot of noise due to dataset bias [26]. In both protocols, test75 is not used at all.

As feature representations, we use the activation maps at a given level of a CNN trained on train75¹. The top-line is when these activation maps are stored completely. To evaluate the hashing, they are encoded using hashing methods that can reconstruct an approximation of the original features.

Protocol 1: Retrieval of unseen classes. In this setup, we use the hashing scheme to index train25 and use test25 as queries. For each query from test25, we retrieve nearest neighbors among train25 and then compute the mAP. The nearest neighbors are defined by the L2 distance between descriptors. This is a relevant distance measure for CNN activation maps [27]. The labels of train25 are used for evaluation only. This setup is like an instance search approach except that the ground-truth is given by the class labels. The train75 - train25 - test25 split is the supervised equivalent of the *learn - database - query* split in unsupervised hashing.

Protocol 2: Transfer learning. A new classifier with the same structure as the top of the original CNN is trained from

¹Other types of features, such as GIST are also possible.

Table 3. Retrieval performance (mAP, Protocol 1), when features are extracted at different layers in the network.

CIFAR-100					
Layer	conv3	fc1	fc2	fc3	softmax
Full	15.6%	16.3%	21.3%	22.2%	22.8%
PQ, M=4	16.6%	16.8%	21.0%	21.2%	22.0%

ImageNet					
Layer	fc1	fc2	fc3	softmax	+PQ, M=8
mAP	4.72%	10.89%	11.3%	13.52%	11.4%

Table 4. Accuracy in transfer (protocol 2) for CIFAR-100 when features are extracted at different layers.

Layer	conv3	fc1	fc2
Full	69.6%	61.8%	57.7%
PQ, M=4	43.4%	45.3%	47.4%

scratch using the stored train25 descriptors. The classification accuracy is reported on test25. The goal is to maximize the transfer accuracy on test25.

Compared to recomputing the features from the images, this approach offers two advantages. First, the features are stored in a compact and semantic way. Secondly, it avoids the computationally intensive computation of the low-level activations: in the case of AlexNet on CIFAR-10, 80% of the computation time is spent in the lower convolutional layers.

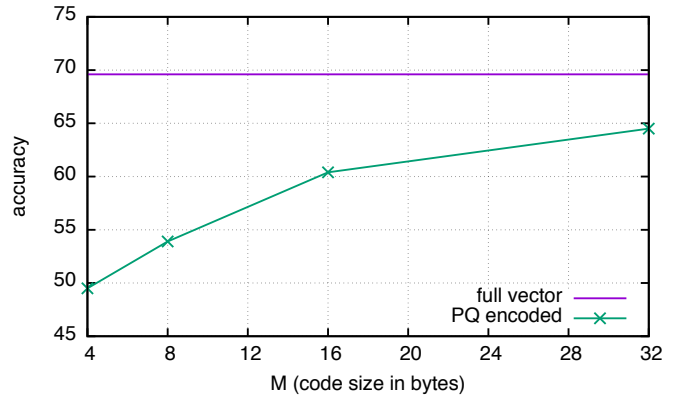
6. BASELINES ON THE TRANSFER-BASED EVALUATION PROTOCOL

We evaluate retrieval and classification methods based on hashed descriptors with our two protocols on CIFAR-100 [20], which is the same dataset as CIFAR-10 except that it is divided into 100 classes instead of 10), and ImageNet. We used the AlexNet [28] architecture, with 3 (resp 5) convolutional layers (“conv”) and 2 fully-connected (“fc”) ones for CIFAR-100 (resp ImageNet).

As an unsupervised baseline for hashing, we report the performance of the Product Quantizer (PQ) [29]. This is an efficient method that can reconstruct approximate vectors from the codes. The PQ parameter M is the number of quantizers and the number of bytes of the produced codes.

Retrieval. Retrieval on the classification layer is done using an inner product similarity, for all other layers we use L2 distance. Table 3 reports the results with full activations and PQ. The performance increases monotonically with the CNN level. This is consistent with prior observations [17] that attribute vectors from arbitrary classes are an efficient representation for global image matching, and further extends the

Fig. 1. Accuracy in transfer (Protocol 2), for layer conv3 on CIFAR-100, as a function of the number of bytes per image.



findings of Section 3 to a different set of classes. The hierarchy of performance is maintained with the PQ encoding. For CIFAR-100, where the softmax outputs 75D vectors, the performance loss due to the hashing is limited even with relatively small 32-bit codes. For ImageNet, the raw retrieval performance is lower, due to the larger number of classes (250). 64-bit unsupervised encoding looses about 2%.

Classification. Results of Protocol 2 (Table 4) show that there is a trade-off between classification error and quantization error: activations of lower layers are more general-purpose (see, e.g., [30]), so training on train25 is more effective. However, lower layers have larger activation maps, which are harder to encode, which leads to a compromise. In this example, the best transfer performance we can achieve with 4 bytes is 47.7%. **For a higher number of bytes, however, it is worth putting the quantization layer lower in the network.**

Figure 1 shows that more bytes bring the performance closer to the full-vector performance. The margin for improvement left to supervised hashing is to bring this performance closer to the 69.6% obtained without any encoding.

7. CONCLUSION

In this paper, we showed that the supervised hashing protocols currently used in the literature are flawed because the evaluation is done on the same classes as the training. In this setting, encoding in binary the output of a simple classifier provides a very strong baseline. To circumvent this issue, we introduced two new protocols that evaluate hashing functions on a disjoint set of labels. The first one evaluates the retrieval performance on a disjoint set of classes. It is very close to the classical setup of unsupervised hashing, and traditional methods seem to perform well. The second protocol evaluates the accuracy of a classifier trained on hash codes with classes never seen before.

8. REFERENCES

- [1] Mohammad Norouzi, David J Fleet, and Ruslan R Salakhutdinov, “Hamming distance metric learning,” in *NIPS*, 2012.
- [2] Jun Wang, Wei Liu, Andy X. Sun, and Yu-Gang Jiang, “Learning hash codes with listwise supervision,” in *ICCV*, 2013.
- [3] Jianfeng Wang, Jingdong Wang, Nenghai Yu, and Shipeng Li, “Order preserving hashing for approximate nearest neighbor search,” in *ACM Multimedia*, 2013.
- [4] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan, “Deep semantic ranking based hashing for multi-label image retrieval,” in *CVPR*, June 2015.
- [5] Brian Kulis and Trevor Darrell, “Learning to hash with binary reconstructive embeddings,” in *NIPS*, December 2009.
- [6] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang, “Semi-supervised hashing for scalable image retrieval,” in *CVPR*, June 2010.
- [7] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang, “Supervised hashing with kernels,” in *CVPR*, June 2012.
- [8] Ruslan Salakhutdinov and Geoffrey Hinton, “Semantic hashing,” *IJAR*, 2009.
- [9] Mohammad Norouzi and David J. Fleet, “Minimal loss hashing for compact binary codes,” in *ICML*, 2011.
- [10] Guosheng Lin, Chunhua Shen, Qinfeng Shi, Anton van den Hengel, and David Suter, “Fast supervised hashing with decision trees for high-dimensional data,” in *CVPR*, June 2014.
- [11] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen, “Supervised discrete hashing,” in *CVPR*, June 2015.
- [12] Xiaojuan Wang, Ting Zhang, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang, “Supervised quantization for similarity search,” in *CVPR*, June 2016.
- [13] Thanh-Toan Do, Anh-Zung Doan, and Ngai-Man Cheung, “Learning to hash with binary deep neural network,” *ECCV*, 2016.
- [14] Ziming Zhang, Yuting Chen, and Venkatesh Saligrama, “Efficient training of very deep neural networks for supervised hashing,” in *CVPR*, June 2016.
- [15] Thanh-Toan Do, Anh-Dzung Doan, Duc Thanh Nguyen, and Ngai-Man Cheung, “Binary hashing with semidefinite relaxation and augmented lagrangian,” *ECCV*, 2016.
- [16] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *CVPR*, 2009.
- [17] Andrew Fitzgibbon Lorenzo Torresani, Martin Szummer, “Efficient object category recognition using classemes,” in *ECCV*, 2010.
- [18] Stephen E Robertson, “The probability ranking principle in ir,” *Journal of documentation*, 1977.
- [19] Hervé Jegou, Teddy Furon, and Jean-Jacques Fuchs, “Anti-sparse coding for approximate nearest neighbor search,” in *ICASSP*, January 2012.
- [20] Alex Krizhevsky, “Learning multiple layers of features from tiny images,” Tech. Rep., University of Toronto, 2009.
- [21] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen, “Deep supervised hashing for fast image retrieval,” in *CVPR*, June 2016.
- [22] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan, “Supervised hashing for image retrieval via image representation learning,” *AAAI*, 2014.
- [23] Wei Dong, Richard Socher, Li Li-Jia, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, June 2009.
- [24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [25] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka, “Metric learning for large scale image classification: Generalizing to new classes at near-zero cost,” in *ECCV*, December 2012.
- [26] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in *CVPR*, 2011.
- [27] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky, “Neural codes for image retrieval,” in *ECCV*, September 2014.
- [28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, December 2012.
- [29] Hervé Jegou, Matthijs Douze, and Cordelia Schmid, “Product quantization for nearest neighbor search,” *IEEE Trans. PAMI*, January 2011.
- [30] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, “How transferable are features in deep neural networks?,” in *NIPS*, 2014.