Supplemental Material

1 More experimental results

Apart from the evaluation based on Hamming distance metric, we also provide the Precision@top100 indicator. Table 1 and Table 2 show the comparison results in Precision@top100 on the three datasets. Compared with other methods, DLDAH achieves superior performance. To be specific, there are three points we should pay attention to. First, LDAH performs much worse than conventional supervised methods, even some unsupervised ones, e.g., ITQ. This is because the simple linear projection of LDA is not entirely suitable to deal with such highly complex deep features. However, the proposed DLDAH enjoys extremely noticeable improvement of about 50-70 points. Such performance comes from the efficient LDA supervision for training the whole network in the end-to-end fashion. Second, the deep model with triplet loss (i.e., DTSH) performs better than the pairwise one (i.e., DHN) in MAP, but worse in Precision@top100. Such phenomenon could be because the triplet loss pays more attention in distinguishing different categories but fails to shrink the intra-category covariance, while the pair loss takes the opposite way. But DLDAH takes both advantages and shows the best scores on both metrics. Third, DSDH almost shows decent performance, as it combines the pairwise supervision and the original one-hot supervision. In fact, the pairwise label comes from the one-hot label, which means the two-stream supervision is redundancy, meanwhile, such network requires complex alternating optimization. In contrast, DLDAH can be integrally trained by off-the-shelf SGD optimizer, therefore shows better performance.

Table 1: The comparison results of different hashing methods in Precision@top100 on MNIST and ImageNet.

Dataset	MNIST						ImgaeNet			
#bits	8	16	24	32	64	128	8	16	32	48
LSH	0.1798	0.3555	0.4143	0.4840	0.6910	0.7915	0.0289	0.0623	0.1371	0.2293
SH	0.4382	0.6419	0.6942	0.7285	0.7666	0.7938	0.0728	0.2090	0.3321	0.3974
ITQ	0.5671	0.7177	0.7805	0.8190	0.8571	0.8824	0.0876	0.2767	0.4375	0.4993
LDAH	0.5917	0.5905	0.5488	0.5053	0.4278	0.4107	0.0591	0.1767	0.3406	0.4352
SDH	0.4853	0.6963	0.7603	0.8061	0.8539	0.9080	0.1223	0.4280	0.5353	0.5717
FSDH	N/A	0.7372	0.7113	0.7372	0.7113	0.7372	N/A	N/A	N/A	N/A
DHN	0.7932	0.9320	0.9623	0.9662	0.9754	0.9850	0.1062	0.4355	0.5693	0.5996
HashNet	0.6873	0.9398	0.9689	0.9736	0.9791	0.9852	0.1042	0.4538	0.5767	0.6158
DTSH	0.8615	0.8997	0.9014	0.9224	0.9568	0.9808	0.1287	0.3817	0.5230	0.5548
DSDH	0.9082	0.9478	0.9309	0.9454	0.9810	0.9834	0.1362	0.3917	0.5238	0.5818
DLDAH	**0.9314**	**0.9714**	**0.9800**	**0.9826**	**0.9867**	**0.9860**	**0.1592**	**0.4548**	**0.5964**	**0.6635**

Preprint. Work in progress.

Table 2: The comparison results of different hashing methods in Precision@top100 on CIFAR-10.

Dataset	CIFAR-10					
Code #bits	8	16	24	32	64	128
LSH	0.1201	0.1964	0.2114	0.2702	0.3076	0.4018
SH	0.2300	0.3366	0.3571	0.3699	0.4021	0.4392
ITQ	0.2649	0.3939	0.4161	0.4385	0.5031	0.5349
LDAH	0.1926	0.2094	0.1937	0.1789	0.1695	0.1608
SDH	0.2461	0.2864	0.3212	0.3466	0.3867	0.5327
FSDH	N/A	0.3072	0.3060	0.3072	0.3072	0.3072
DHN	0.2341	0.5187	0.6813	0.7330	0.7552	0.7785
HashNet	0.2482	0.5270	0.6909	0.7322	0.7695	0.7873
DTSH	0.4076	0.6217	0.6772	0.6953	0.6792	0.4573
DSDH	0.4044	0.6305	0.6643	0.6870	0.6849	0.7197
DLDAH	**0.4632**	**0.6329**	**0.7167**	**0.7580**	**0.7740**	**0.8252**

2 Proof of Theorem 1

Theorem 1. *Maximizing the LDA objective over deep features is equivalent to minimizing the linear least square error, i.e.,*

$$\begin{aligned} &\max_{f} Tr\left(\left(S_t+\mu I\right)^{-1}S_b\right) \\ &\Leftrightarrow \min_{f,W,b}\left\|X^TW+1b^T-\widetilde{Y}\right\|_F^2+\mu\left\|W\right\|_F^2, \end{aligned} \tag{1}$$

where f stands for the deep nonlinear network and $\widetilde{Y}=Y\left(Y^TY\right)^{-1/2}$.

Proof. Let

$$g\left(W,b\right)=\left\|X^TW+1b^T-\widetilde{Y}\right\|_F^2+\mu\left\|W\right\|_F^2. \tag{2}$$

According to the Lagrange multipliers method, setting the partial derivative w.r.t. b to zeros,

$$\frac{\partial g\left(W,b\right)}{\partial b}=\left(X^TW+1b^T-\widetilde{Y}\right)^T\cdot 1=0 \tag{3}$$

Then

$$b=\frac{1}{n}\left(\widetilde{Y}^T-W^TX\right)\cdot 1 \tag{4}$$

Taking b into Eq.2, then Eq.2 becomes,

$$\begin{aligned} &g\left(W\right) \\ &=\left\|X^TW+\tfrac{1}{n}1\cdot 1^T\left(X^TW-\widetilde{Y}\right)-\widetilde{Y}\right\|_F^2+\mu\left\|W\right\|_F^2 \\ &=\left\|HX^TW-H\widetilde{Y}\right\|_F^2+\mu\left\|W\right\|_F^2 \end{aligned} \tag{5}$$

Similarly, setting the partial derivative w.r.t. W to zeros,

$$\frac{\partial g\left(W\right)}{\partial W}=2XH\left(HX^TW-H\widetilde{Y}\right)+2\mu W=0 \tag{6}$$

Then,

$$W=\left(XHX^T+\mu I\right)^{-1}XH\tilde{Y} \tag{7}$$

By substituting Eq.7 into Eq.5, we can have

$$
\begin{aligned}
&\left\|HX^TW - H\widetilde{Y}\right\|_F^2 + \mu \left\|W\right\|_F^2 \\
&= Tr\left[W^TXHX^TW - W^TXH\widetilde{Y} - \widetilde{Y}^THX^TW + \widetilde{Y}H\widetilde{Y}\right] \\
&\quad + \mu Tr\left(W^TW\right) \\
&= Tr\left[W^T\left(XHX^T + \mu I\right)W - W^TXH\widetilde{Y} - \widetilde{Y}^THX^TW\right] \\
&\quad + Tr\left(\widetilde{Y}H\widetilde{Y}\right) \\
&= Tr\left(\widetilde{Y}H\widetilde{Y}\right) - Tr\left[\widetilde{Y}^THX^TW\right] \\
&= Tr\left(\widetilde{Y}H\widetilde{Y}\right) - Tr\left[\left(XHX^T + \mu I\right)^{-1}XH\widetilde{Y}\widetilde{Y}^THX^T\right]
\end{aligned}
\tag{8}
$$

As the term of $Tr\left(\widetilde{Y}H\widetilde{Y}\right)$ is constant, the original minimization problem is transformed into a maximization objective of LDA,

$$
\max_f Tr\left(\left(S_t + \mu I\right)^{-1}S_b\right). \tag{9}
$$

This completes our proof. □