Supplementary Material: Hierarchical Semantic Indexing for Large Scale Image Retrieval

Jia Deng[1,3]
Princeton University[1]

Alexander C. Berg[2]
Stony Brook University[2]

Li Fei-Fei[3]
Stanford University[3]

1. Proofs

Remark 1.1. We first provide proofs and constructions for probability vectors for non-overlapping categories (Lemma 1.4–1.12), *i.e.* $x \in \mathbb{R}^K, \sum_i x_i = 1, 0 \leq x_i \leq 1$ for $i = 1, \ldots, K$. We use Δ^{K-1} to denote the set of all such vectors. In Lemma 1.15, we show extension to the general case where $x \in \mathbb{R}^K, 0 \leq x_i \leq 1$ for $i = 1, \ldots, K$ (but does not necessarily sum to one). We use $\tilde{\Delta}^{K-1}$ to denote the set of all such vectors.

Definition 1.2. A matrix $S \in \mathbb{R}^{K \times K}$ is *hashable*, if there exists a $\lambda_S > 0$ and, for any $\epsilon > 0$, a distribution on a family $\mathcal{H}(S, \epsilon)$ of hash functions $h(\cdot; S, \epsilon)$ such that for any $x, y \in \Delta^{K-1}$,

$$0 \leq \Pr\left(h_1(x; S, \epsilon) = h_2(y; S, \epsilon)\right) \; - \; \lambda_S \cdot x^T S y \leq \epsilon$$

where h_1 and h_2 are drawn independently from $\mathcal{H}(S, \epsilon)$.

Remark 1.3. Here we relax the equality in the LSH condition $\Pr(h_1(x) = h_2(y)) = Sim(x, y)$ to equality up to ϵ. This has virtually no practical impact because in all of our constructions ϵ can be easily made negligibly small, without incurring any additional computational cost. Also note that scaling S does not affect the ranking induced by the similarity $x^T S y$.

Lemma 1.4. *If S is symmetric, element-wise non-negative and diagonally dominant, that is, $\forall i = 1, \ldots, K, \; s_{ii} \geq \sum_{j \neq i} s_{ij}$, then S is hashable.*

Proof. Define a $K \times (K+1)$ matrix $\Theta = (\theta_{ij})$, where

$$\begin{aligned}
\theta_{ij} &= \sqrt{\hat{s}_{ij}}, \; \forall i = 1, \ldots, K, \; \forall j = 1, \ldots, K, \; i \neq j. \\
\theta_{ii} &= \sqrt{\hat{s}_{ii} - \sum_{j \neq i} \hat{s}_{ij}}, \; \forall i = 1, \ldots, K. \\
\theta_{i,K+1} &= 1 - \sum_{j=1}^{K} \theta_{ij}, \; \forall i = 1, \ldots, K.
\end{aligned}$$

where $\hat{S} = \lambda_S \cdot S$ with λ_S chosen to ensure $\theta_{i,K+1} \geq 0$. Note that each row of Θ sums to one. Also note that $\theta_{ij} = \theta_{ji}, \forall i, j \leq K$ due to the symmetry of S.

Consider hash functions $h(x)$ that map a probability vector to a set of positive integers, that is, $h : \Delta^{K-1} \to 2^{\mathbb{N}}$ where $2^{\mathbb{N}}$ is all subsets of natural numbers. Note that $h(x) = h(y)$ is defined as *set equality*, that is, the ordering of elements does not matter.

To construct $\mathcal{H}(S, \epsilon)$, let $N \geq 1/\epsilon$. Then $h(x; S, \epsilon)$ is computed as follows:

1. Sample $\alpha \in \{1, \ldots, K\} \sim multi(x)$
2. Sample $\beta \in \{1, \ldots, K+1\} \sim multi(\theta_\alpha)$ where θ_α is the α^{th} row of Θ.
3. If $\beta \leq K$, return $\{\alpha, \beta\}$

4. Randomly pick γ from $\{K+1,\ldots,K+N\}$, return $\{\gamma\}$.

In implementation, h is parametrized by three uniformly drawn values $p, q \in [0,1]$ and $r \in \{1 \ldots N\}$, used respectively in the sampling process for α, β and γ.

Let x, y be probability vectors, $x, y \in \Delta^{K-1}$. Let $\alpha_x, \beta_x, \gamma_x$ be the values sampled when computing $h(x)$, and similarly for $\alpha_y, \beta_y, \gamma_y$. To compute $\Pr(h(x) = h(y))$, consider two cases below.

Case 1: Suppose $\alpha_x = i \in \{1,\ldots,K\}$, $\alpha_y = j \in \{1,\ldots,K\}$, $i \neq j$. Then

$$\begin{aligned}
\Pr(h(x) = h(y) \mid \alpha_x = i \wedge \alpha_y = j) &= \Pr(\beta_x = j \wedge \beta_y = i \mid \alpha_x = i \wedge \alpha_y = j) + \\
&\quad \Pr(\gamma_x = \gamma_y \wedge \beta_x = K+1 \wedge \beta_y = K+1 \mid \alpha_x = i \wedge \alpha_y = j) \\
&= \Pr(\beta_x = j \mid \alpha_x = i) \times \Pr(\beta_y = i \mid \alpha_y = j) + \\
&\quad \Pr(\gamma_x = \gamma_y \mid \beta_x = K+1, \beta_y = K+1) \times \\
&\quad \Pr(\beta_x = K+1 \mid \alpha_x = i) \times \Pr(\beta_y = K+1 \mid \alpha_y = j) \\
&= \theta_{ij}\theta_{ji} + \frac{1}{N}\,\theta_{i,K+1}\theta_{j,K+1} \\
&= \hat{s}_{ij} + \frac{1}{N}\,\theta_{i,K+1}\theta_{j,K+1}
\end{aligned}$$

Case 2: Suppose $\alpha_x = \alpha_y = i \in \{1,\ldots,K\}$. Then

$$\begin{aligned}
\Pr(h(x) = h(y) \mid \alpha_x = \alpha_y = i) &= \Pr(\beta_x = \beta_y \leq K \mid \alpha_x = \alpha_y = i) + \\
&\quad \Pr(\gamma_x = \gamma_y \wedge \beta_x = K+1 \wedge \beta_y = K+1 \mid \alpha_x = \alpha_y = i) \\
&= \sum_{j=1}^{K} \Pr(\beta_x = \beta_y = j \mid \alpha_x = \alpha_y = i) + \\
&\quad \Pr(\gamma_x = \gamma_y \mid \beta_x = K+1, \beta_y = K+1) \times \\
&\quad \Pr(\beta_x = K+1 \mid \alpha_x = i) \times \Pr(\beta_y = K+1 \mid \alpha_y = j) \\
&= \sum_{j=1}^{K} \theta_{ij}^2 + \frac{1}{N}\,\theta_{i,K+1}^2 \\
&= \hat{s}_{ii} + \frac{1}{N}\,\theta_{i,K+1}^2
\end{aligned}$$

Summing up the above conditional probabilities, we get

$$\begin{aligned}
\Pr(h(x) = h(y)) &= \sum_{i \neq j} x_i y_j \Pr(h(x) = h(y) | \alpha_x = i \wedge \alpha_y = j) + \\
&\quad \sum_i x_i y_i \Pr(h(x) = h(y) | \alpha_x = \alpha_y = i) \\
&= \sum_{i,j} x_i \hat{s}_{ij} y_j + \frac{1}{N} \sum_{i \neq j} x_i y_j \theta_{i,K+1}\theta_{j,K+1} + \frac{1}{N} \sum_i x_i y_i \theta_{i,K+1}^2 \\
&= \lambda_S x^T S y + \frac{1}{N} \sum_{i,j} x_i y_j \theta_{i,K+1}\theta_{j,K+1}
\end{aligned}$$

To conclude the proof, observe that

$$0 \leq \frac{1}{N} \sum_{i,j} x_i y_j \theta_{i,K+1}\theta_{j,K+1} \leq \frac{1}{N} \left(\sum_i x_i \theta_{i,K+1} \right) \left(\sum_j x_j \theta_{j,K+1} \right) \leq \epsilon$$

□

Remark 1.5. For the special case where S is the identity matrix, $h(x;S)$ reduces to $h(x;I)$, which returns an $\alpha \in \{1,\dots,K\}$ sampled from $multi(x)$.

Lemma 1.6. *If S is a matrix of all ones, then S is hashable.*

Proof. Note that $x^T S y = 1$ in this case since $x, y \in \Delta^{K-1}$. Simply let $\mathcal{H}$ consist of one constant function. $\square$

Definition 1.7. A matrix $Q \in \mathbb{R}^{m\times m}$ is a *zero padded extension* of $S \in \mathbb{R}^{n\times n}$ if there exists an one-to-one function f that maps the indices $\tau = \{1\dots n\}$ to $\{1\dots m\}$ such that $Q_{i,j} = S_{f^{-1}(i),f^{-1}(j)}$ for any $i,j \in f(\tau)$ and $Q_{i,j} = 0$ otherwise.

Remark 1.8. In other words, Q is obtained by symmetrically inserting rows and columns of zeros into S.

Lemma 1.9. *If Q is a zero padded extension of S and S is hashable, then Q is hashable.*

Proof. Let $\epsilon > 0$, and let $x, y \in \Delta^{K-1}$. Define $x_{f(\tau)} \in \mathbb{R}^n$ such that its i^{th} element is $x_{f^{-1}(i)}$. We the define $g(x;Q,\epsilon)$ as follows:

1. Sample $\alpha \in \{1,\dots,m\} \sim multi(x)$
2. If $\alpha \in f(\tau)$, return $\left(0, h(\frac{x_{f(\tau)}}{||x_{f(\tau)}||_1}; S, \frac{\epsilon}{2})\right)$, where $h \in \mathcal{H}(S, \frac{\epsilon}{2})$ as in Definition 1.2
   Else return $\beta \in \{1,\dots,N\}$ uniformly drawn, where $N = \lceil 2/\epsilon \rceil$.

We now show Q is hashable.

$$\begin{aligned}
\Pr\left(g(x;Q,\epsilon) = g(y;Q,\epsilon)\right) &= \Pr\left(\alpha_x \in f(\tau) \wedge \alpha_y \in f(\tau) \wedge h\left(\frac{x_{f(\tau)}}{||x_{f(\tau)}||_1}; S, \frac{\epsilon}{2}\right) = h\left(\frac{y_{f(\tau)}}{||y_{f(\tau)}||_1}; S, \frac{\epsilon}{2}\right)\right) \\
&\quad + \Pr\left(\beta_x = \beta_y\right) \\
&= ||x_{f(\tau)}||_1 \cdot ||y_{f(\tau)}||_1 \cdot \left(\lambda_S \frac{x^T_{f(\tau)}}{||x_{f(\tau)}||_1} S \frac{y_{f(\tau)}}{||y_{f(\tau)}||_1} + \delta\right) \\
&\quad + \frac{1}{N}(1 - ||x_{f(\tau)}||_1)(1 - ||y_{f(\tau)}||_1) \\
&= \lambda_S \cdot x^T Q y + ||x_{f(\tau)}||_1 \cdot ||y_{f(\tau)}||_1 \cdot \delta + \frac{1}{N}(1 - ||x_{f(\tau)}||_1)(1 - ||y_{f(\tau)}||_1)
\end{aligned}$$

where $0 \le \delta \le \epsilon/2$ by the choice of h. Note that

$$||x_{f(\tau)}||_1 \cdot ||y_{f(\tau)}||_1 \cdot \delta + \frac{1}{N}(1 - ||x_{f(\tau)}||_1)(1 - ||y_{f(\tau)}||_1) \le \epsilon/2 + \epsilon/2 = \epsilon$$

$\square$

Lemma 1.10. *If S is hashable, then aS is hashable for any $a > 0$.*

Proof. This follows directly from Definition 1.2 (by using $\lambda_{aS} = \frac{1}{a}\lambda_S$).

$\square$

Lemma 1.11. *If $Q = \sum_{l=1}^{L} S_l$ and S_l is hashable for $l = 1,\dots,L$, then Q is hashable.*

Proof. Suppose the hash function for S_l is h_l and the scalar is λ_{S_l}, for $l = 1,\dots,L$.
Let $z = \sum_{l=1}^{L} \frac{1}{\sqrt{\lambda_{S_l}}}$ and $\theta \in \mathbb{R}^L$ where $\theta_l = \frac{1}{z} \cdot \frac{1}{\sqrt{\lambda_{S_l}}}$.
We construct hash function $g(x;Q,\epsilon)$ as follows:

1. Sample $\alpha \in \{1,\dots,L\} \sim multi(\theta)$.
2. return $(\alpha, h_\alpha(x; S_l, \epsilon/L))$.

Then

$$
\begin{aligned}
\Pr\left(g(x;Q,\epsilon)=g(y;Q,\epsilon)\right) &= \sum_{l=1}^{L}\Pr\left(\alpha_x=\alpha_y=l\wedge h_l(x;S_l,\epsilon/L)=h_l(y;S_l,\epsilon/L)\right)\\
&= \sum_{l=1}^{L}\theta_l^2(\lambda_{S_l}x^TS_ly+\delta_l)\\
&= \frac{1}{z^2}x^TQy+\sum_{l=1}^{L}\theta_l^2\delta_l
\end{aligned}
$$

where $0\le\delta_l\le\epsilon/L$. Note that $0\le\sum_l^L\theta_l^2\delta_l\le\epsilon$ and thus Q is hashable.

□

Lemma 1.12. *Let $T=G(V,E)$ be a rooted tree and define $\pi_{m,n}$ to be the lowest common ancestor between node m and n for any $m,n\in V$. Let $V_r\subseteq V$ be subtree rooted at r (i.e., the set of all nodes descending from node $r\in V$ including r itself). Let $\Omega_r\subseteq V_r$ be all the leaf nodes of r and let $K_r=|\Omega_r|$. Let $f_r:\Omega_r\to\{1,\ldots,K_r\}$ be a one-to-one correspondence of the leaf nodes of r to a set of integers. Let $\xi(\cdot):V\to\mathbb{R}$ be any function defined on V. Let $S^{(r,\xi)}\in\mathbb{R}^{K_r\times K_r}$ be a similarity matrix induced by r and ξ, where $S^{(r,\xi)}_{ij}=\xi(\pi_{f_r^{-1}(i),f_r^{-1}(j)}),\forall i=1,\ldots,K_r,j=1,\ldots,K_r$.*

For any $r\in V$, if $\xi(\cdot)$ is non-negative and downward non-decreasing in the subtree of r, that is, $\xi(q)\ge0$ for any $q\in V_r$ and $\xi(q)\ge\xi(p)$ for any $p,q\in V_r$ such that q is a child of p, then $S^{(r,\xi)}$ is hashable.

Proof. Let $r\in V$. Suppose $\xi(\cdot)$ is non-negative and downward non-decreasing in the subtree of r. We prove the claim by induction on the tree.

If r is a leaf node, then $S^{(r,\xi)}$ is a scalar and thus hashable.

Now we consider the case when r is an internal node. Let $\sigma(r)$ be the set of direct children of r. Our inductive hypothesis is that given any $c\in\sigma(r)$, the similarity matrix $S^{(c,\xi')}$ induced by c and any $\xi':V_c\to\mathbb{R}$, which is non-negative and downward non-decreasing, is hashable.

For a given $c\in\sigma(r)$, let $f_r(\Omega_c)$ be the set of indices of the leaf nodes of c in $S^{(r,\xi)}$. The tree structure implies

$$\bigcup_{c\in\sigma(r)}f_r(\Omega_c)=\{1,\ldots,K_r\} \tag{1}$$

and

$$f_r(\Omega_c)\bigcap f_r(\Omega_d)=\emptyset,\text{for any } c,d\in\sigma(r) \text{ and } c\ne d\;. \tag{2}$$

That is, the columns and rows of $S^{(r,\xi)}$ can be partitioned by the direct children of r.

Also, if c and d are different direct children of r, then the lowest common ancestor between the descendant nodes of c and those of d must be r. Thus

$$S^{(r,\xi)}_{f_r(\Omega_c),f_r(\Omega_d)}=\xi(\pi_{\Omega_c,\Omega_d})=\xi(r)\cdot\mathbf{1},\text{for any } c,d\in\sigma(r)\text{ and } c\ne d\;. \tag{3}$$

where $\mathbf{1}$ is a matrix of all ones.

For a given $c\in\sigma(r)$, define $Q^{(c)}\in\mathbb{R}^{K_r\times K_r}$ such that

$$Q^{(c)}_{ij}=\begin{cases}S^{(r,\xi)}_{ij}-\xi(r) & \text{if } i,j\in f_r(\Omega_c)\\ 0 & \text{otherwise.}\end{cases}$$

It follows from (1), (2) and (3) that

$$S^{(r,\xi)}=\xi(r)\cdot\mathbf{1}+\sum_{c\in\sigma(r)}Q^{(c)} \tag{4}$$

Define $\xi'(\cdot)=\xi(\cdot)-\xi(r)$. Since the lowest common ancestor of the leaf nodes of r cannot be higher than r and ξ is downward non-decreasing, we conclude that $\xi'(d)\ge0$ for any $d\in V_r$ and $\xi'(d)$ is downward non-decreasing.

By the inductive hypothesis, given any $c\in\sigma(r)$, the similarity matrix $S^{(c,\xi')}$ induced by c and ξ' is hashable.

Now we show that $Q^{(c)}$ is a zero padded extension of $S^{(c,\xi')}$.

Let $K_c = |\Omega_c|$ and f_c be the function that maps the nodes in Ω_c to indices of $S^{(c,\xi')}$. Recall that f_r maps nodes in Ω_r (including Ω_c) to indices in $S^{(r,\xi)}$.

Let $f : \{1,\ldots,K_c\} \to \{1,\ldots,K_r\}$, where $f = f_r \cdot f_c^{-1}$. Let $\tau = \{1,\ldots,K_c\}$. It follows that $f(\tau) = f_r(\Omega_c)$.

For any $i,j \in f(\tau)$, that is, $\forall i,j \in f_r(\Omega_c)$,

$$\begin{aligned}
Q^{(c)}_{ij} &= S^{(r,\xi)}_{ij} - \xi(r) \\
&= \xi(\pi_{f_r^{-1}(i),f_r^{-1}(j)}) - \xi(r)(\text{By definition of } f_r) \\
&= \xi'(\pi_{f_r^{-1}(i),f_r^{-1}(j)})(\text{By definition of } \xi') \\
&= S^{(c,\xi')}_{f_c\cdot f_r^{-1}(i),f_c\cdot f_r^{-1}(j)}(\text{By definition of } f_c) \\
&= S^{(c,\xi')}_{f^{-1}(i),f^{-1}(j)}
\end{aligned}$$

By Definition 1.7, $Q^{(c)}$ is a zero padded extension of $S^{(c,\xi')}$ and is therefore hashable by Lemma 1.9. It follows from Lemma 1.6, Lemma 1.10, Lemma 1.11 and from (4) that $S^{(r,\xi)}$ is hashable. □

Remark 1.13. Note that a similarity matrix derived from a hierarchy, as in Lemma 1.12, is not necessarily diagonally dominant. For example, if a leaf node has many siblings, the sum of its similarities with its siblings can easily be more than its self similarity.

Definition 1.14. A matrix $S \in \mathbb{R}^{K\times K}$ is *generally hashable*, if there exists a $\lambda_S > 0$ and, for any $\epsilon > 0$, a distribution on a family $\mathcal{H}(S,\epsilon)$ of hash functions $h(\cdot; S,\epsilon)$ such that for any $x, y \in \tilde{\Delta}^{K-1}$,

$$0 \le \Pr\left(h_1(x;S,\epsilon) = h_2(y;S,\epsilon)\right) \;-\; \lambda_S \cdot x^T S y \le \epsilon$$

where h_1 and h_2 are drawn independently from $\mathcal{H}(S,\epsilon)$.

Lemma 1.15. Hashing for the general case. *Any hashable matrix $S \in \mathbb{R}^{K\times K}$ is generally hashable.*

Proof. For any $x, y \in \tilde{\Delta}^{K-1}$, let $\hat{x} = (x/K, 1 - \sum_i x_i/K) \in \mathbb{R}^{K+1}$ and $\hat{y} = (y/K, 1 - \sum_i y_i/K) \in \mathbb{R}^{K+1}$. Observe that $\hat{x}$ and $\hat{y} \in \Delta^K$. Let

$$\hat{S} \in \mathbb{R}^{(K+1)\times(K+1)}, \hat{S} = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}$$

$\hat{S}$ is a zero padded extension of S and is therefore hashable by Lemma 1.9. That is, there exists a $\lambda_{\hat{S}}$ and for any $\epsilon > 0$, a distribution on a family of functions $\hat{\mathcal{H}}$ such that

$$0 \le \Pr{}_{\hat{h}_1,\hat{h}_2\in\hat{\mathcal{H}}}(\hat{h}_1(\hat{x}) = \hat{h}_2(\hat{y})) - \hat{x}^T\hat{S}\hat{y} \le \epsilon.$$

Observe that $\hat{x}^T\hat{S}\hat{y} = x^T S y$. Therefore

$$0 \le \Pr{}_{\hat{h}_1,\hat{h}_2\in\hat{\mathcal{H}}}(\hat{h}_1(\hat{x}) = \hat{h}_2(\hat{y})) - x^T S y \le \epsilon.$$

Let $h(z) = \hat{h}(\hat{z})$, for any $z \in \tilde{\Delta}^{K-1}$. Observe that $\Pr(h_1(x) = h_2(y)) = \Pr(\hat{h}_1(\hat{x}) = \hat{h}_2(\hat{y}))$. Therefore,

$$0 \le \Pr(h_1(x) = h_2(y)) - x^T S y \le \epsilon.$$

By Definition 1.14, S is generally hashable. □