Supplementary Material: Hierarchical Semantic Indexing for Large Scale Image Retrieval

Jia Deng\textsuperscript{1,3} Princeton University\textsuperscript{1} Alexander C. Berg\textsuperscript{2} Stony Brook University\textsuperscript{2} Li Fei-Fei\textsuperscript{3} Stanford University\textsuperscript{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 $\Delta^{K-1}$ 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 $\hat{\Delta}^{K-1}$ 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 $H(S, \epsilon)$ of hash functions $h(\cdot; S, \epsilon)$ such that for any $x, y \in \Delta^{K-1}$,

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

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

Remark 1.3. Here we relax the equality in the LSH condition $\Pr(h_1(x) = h_2(y)) = \text{Sim}(x, y)$ to equality up to $\epsilon$. This has virtually no practical impact because in all of our constructions $\epsilon$ 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 $x^T Sy$.

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

$$\theta_{ij} = \sqrt{s_{ij}}, \forall i = 1, \ldots, K, \forall j = 1, \ldots, K, i \neq j.$$

$$\theta_{ii} = \sqrt{\hat{s}_{ii} - \sum_{j \neq i} \hat{s}_{ij}}, \forall i = 1, \ldots, K.$$

$$\theta_{i,K+1} = 1 - \sum_{j=1}^{K} \theta_{ij}, \forall i = 1, \ldots, K.$$

where $\hat{S} = \lambda_S \cdot S$ with $\lambda_S$ chosen to ensure $\theta_{i,K+1} \geq 0$. Note that each row of $\Theta$ 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, $h : \Delta^{K-1} \rightarrow 2^\mathbb{N}$ 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 $H(S, \epsilon)$, let $N \geq 1/\epsilon$. Then $h(x; S, \epsilon)$ is computed as follows:

1. Sample $\alpha \in \{1, \ldots, K\} \sim \text{multi}(x)$
2. Sample $\beta \in \{1, \ldots, K + 1\} \sim \text{multi}(\theta_\alpha)$ where $\theta_\alpha$ is the $\alpha$th row of $\Theta$.
3. If $\beta \leq K$, return $\{\alpha, \beta\}$
4. Randomly pick $\gamma$ 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 $\alpha, \beta$ and $\gamma$.

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

$$
\Pr(h(x) = h(y) \mid \alpha_x = i \land \alpha_y = j) = \Pr(\beta_x = j \land \beta_y = i \mid \alpha_x = i \land \alpha_y = j) + \Pr(\gamma_x = \gamma_y \land \beta_x = K + 1 \land \beta_y = K + 1 \mid \alpha_x = i \land \alpha_y = j) = \Pr(\beta_x = j \mid \alpha_x = i) \times \Pr(\beta_y = i \mid \alpha_y = j) + \Pr(\gamma_x = \gamma_y \mid \beta_x = K + 1, \beta_y = K + 1) \times \Pr(\beta_x = K + 1 \mid \alpha_x = i) \times \Pr(\beta_y = K + 1 \mid \alpha_y = j) = \theta_{ij} + \frac{1}{N} \theta_{i,K+1} \theta_{j,K+1}
$$

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

$$
\Pr(h(x) = h(y) \mid \alpha_x = \alpha_y = i) = \Pr(\beta_x = \beta_y \leq K \mid \alpha_x = \alpha_y = i) + \Pr(\gamma_x = \gamma_y \land \beta_x = K + 1 \land \beta_y = K + 1 \mid \alpha_x = \alpha_y = i) = \sum_{j=1}^{K} \Pr(\beta_x = \beta_y = j \mid \alpha_x = \alpha_y = i) + \Pr(\gamma_x = \gamma_y \mid \beta_x = K + 1, \beta_y = K + 1) \times \Pr(\beta_x = K + 1 \mid \alpha_x = i) \times \Pr(\beta_y = K + 1 \mid \alpha_y = j) = \sum_{j=1}^{K} \theta_{ij}^2 + \frac{1}{N} \theta_{i,K+1}^2
$$

Summing up the above conditional probabilities, we get

$$
\Pr(h(x) = h(y)) = \sum_{i \neq j} x_i y_j \Pr(h(x) = h(y) \mid \alpha_x = i \land \alpha_y = j) + \sum_i x_i y_i \Pr(h(x) = h(y) \mid \alpha_x = \alpha_y = i) = \sum_{i,j} x_i \delta_{ij} y_j + \frac{1}{N} \sum_{i \neq j} x_i y_j \theta_{i,K+1} \theta_{j,K+1} + \frac{1}{N} \sum_i x_i y_i \theta_{i,K+1}^2 = \lambda_s x^T S y + \frac{1}{N} \sum_{i,j} x_i y_j \theta_{i,K+1} \theta_{j,K+1}
$$

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 $\alpha \in \{1, \ldots, K\}$ sampled from $\text{multi}(x)$.

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 $H$ consist of one constant function. \hfill \Box

Definition 1.7. A matrix $Q \in \mathbb{R}^{m \times n}$ 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 $\tau = \{1 \ldots n\}$ to $\{1 \ldots m\}$ 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 $\alpha \in \{1, \ldots, m\} \sim \text{multi}(x)$

2. If $\alpha \in f(\tau)$, return $(0, h(\frac{x_{f(\tau)}}{||x_{f(\tau)}||_1}; S, \frac{\epsilon}{2}))$, where $h \in H(S, \frac{\epsilon}{2})$ as in Definition 1.2.

Else return $\beta \in \{1, \ldots, N\}$ uniformly drawn, where $N = \lceil 2/\epsilon \rceil$.

We now show $Q$ is hashable.

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

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

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

\hfill \Box

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 $\lambda_{aS} = \frac{1}{a} \lambda_S$).

\hfill \Box

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

Proof. Suppose the hash function for $S_l$ is $h_l$ and the scalar is $\lambda_{S_l}$, for $l = 1, \ldots, L$.

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 $\alpha \in \{1, \ldots, L\} \sim \text{multi}(\theta)$.

2. return $(\alpha, h_\alpha(x; S_l, \epsilon/L))$. 

Then
\[
\Pr (g(x; Q, \epsilon) = g(y; Q, \epsilon)) = \sum_{l=1}^{L} \Pr (\alpha_x = \alpha_y = l \land h_l(x; S_l, \epsilon/L) = h_l(y; S_l, \epsilon/L)) = \sum_{l=1}^{L} \theta_l^2 (\lambda S_l x^T S_l y + \delta_l)
\]
\[
= \frac{1}{z^2} x^T Q y + \sum_{l=1}^{L} \theta_l^2 \delta_l
\]
where $0 \leq \delta_l \leq \epsilon/L$. Note that $0 \leq \sum_{l=1}^{L} \theta_l^2 \delta_l \leq \epsilon$ 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 $\xi$, 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, $\xi(q) \geq 0$ for any $q \in V_r$ and $\xi(q) \geq \xi(p)$ 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 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

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

(1)

and

\[
f_r(\Omega_c) \bigcap f_r(\Omega_d) = \emptyset, \text{ for any } c, d \in \sigma(r) \text{ and } c \neq d.
\]

(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

\[
S^{(r, \xi)}_{f_r(\Omega_c), f_r(\Omega_d)} = \xi(\pi_{\Omega_c, \Omega_d}) = \xi(r) \cdot \mathbf{1}, \text{ for any } c, d \in \sigma(r) \text{ and } c \neq d.
\]

(3)

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

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

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

(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 $\xi$ is downward non-decreasing, we conclude that $\xi'(d) \geq 0$ 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 $\xi'$ is hashable.
Now we show that $Q^{(e)}$ 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 $\Omega_c$ to indices of $S^{(c,\xi)}$. Recall that $f_r$ maps nodes in $\Omega_r$ (including $\Omega_c$) to indices in $S^{(r,\xi)}$.

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

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

\[
Q^{(e)}_{ij} = S^{(r,\xi)}_{ij} - \xi(r) = \xi'((\pi f^{-1}_r(i), f^{-1}_r(j))) - \xi(r) \quad \text{(By definition of } f_r) \\
= \xi'((\pi f^{-1}_r(i), f^{-1}_r(j))) \quad \text{(By definition of } \xi') \\
= S^{(c,\xi')}_{f_r f^{-1}_r(i), f_r f^{-1}_r(j)} \quad \text{(By definition of } f_c) \\
= S^{(c,\xi')}_{f_r f^{-1}_r(i), f_r f^{-1}_r(j)}
\]

By Definition 1.7, $Q^{(e)}$ 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. \hfill \square

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 $x, y \in \Delta^{K-1}$,

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

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 $x, y \in \Delta^{K-1}$, 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 $\lambda_S$ and for any $\epsilon > 0$, a distribution on a family of functions $\hat{\mathcal{H}}$ such that

\[
0 \leq \Pr_{\hat{h}_1, \hat{h}_2 \in \hat{\mathcal{H}}} (\hat{h}_1(\hat{x}) = \hat{h}_2(\hat{y})) - \hat{x}^T \hat{S} \hat{y} \leq \epsilon.
\]

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

\[
0 \leq \Pr_{\hat{h}_1, \hat{h}_2 \in \hat{\mathcal{H}}} (\hat{h}_1(\hat{x}) = \hat{h}_2(\hat{y})) - x^T y \leq \epsilon.
\]

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

\[
0 \leq \Pr(h_1(x) = h_2(y)) - x^T y \leq \epsilon.
\]

By Definition 1.14, $S$ is generally hashable. \hfill \square