Supervised Hierarchical Cross-Modal Hashing

Changchang Sun†, Xuemeng Song†, Fuli Feng‡, Wayne Xin Zhao§, Hao Zhang*, Liqiang Nie†

†School of Computer Science and Technology, Shandong University
‡School of Computing, National University of Singapore
§School of Information, Renmin University of China
*Mercari, Inc, Japan
Unprecedented growth of multimedia data on the Internet.

Application: cross-modal retrieval.

Solution: supervised cross-modal hashing.

Background

- Mini-skirt
- Long Skirt
- Wide-leg Jeans

Hamming Space

UNIQLO Women Cotton Mini Skirt.
Chicwish Endless Blooming Rose Max Skirt.
Chloé Frayed High-rise Wide-leg Jeans.

Labels

Image

Text
Related Work

- Define cross-modal similarity matrix

Related Work

- Learn semantic information from multiple labels

Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. In CVPR, 2018
Motivation

- Explore the rich semantic information conveyed by the label hierarchy.

![Label Hierarchy Diagram]

- Finest-grained layer
  \[ I_1 \quad I_3 \quad \text{Dissimilar} \]

- Less finer-grained layer
  \[ I_1 \quad I_3 \quad \text{Similar} \]

Figure 1: Illustration of the label hierarchy.
Challenges

- How to employ the *label hierarchy* to guide the cross-modal hashing and preserve the *underlying correlations* from original space to hamming space.
Challenges

- How to enhance the **hierarchical discriminative** power of hash codes.

![Diagram showing hash codes and clothing items]

- Skirt → Hash Code → Skirt (Mini-Skirt)
- Jeans → Hash Code → Jeans (Wide-leg Jeans)
Challenges

- The lack of benchmark dataset, whose data points should involve **multiple modalities** and are **hierarchically labeled**.

<table>
<thead>
<tr>
<th>Super-class</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flowers</td>
<td>Rose, Sunflower, Lily...</td>
</tr>
<tr>
<td>Fish</td>
<td>Goldfish, Shark, Dolphin...</td>
</tr>
<tr>
<td>Insect</td>
<td>Bee, Butterfly, Caterpillar...</td>
</tr>
<tr>
<td>Fruit</td>
<td>Apple, Peach, Pear...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1: Hierarchical labels of benchmark dataset CIFAR-100.
Figure 2: Illustration of the proposed scheme, HiCHNet.
Framework

- Regularized Cross-modal Hashing
  - Layer-wise Hash Representation

$K$ Fully Connected Networks $\rightarrow K$ Layers

\[
\begin{align*}
  h_{v_i}^k &= s(W_v^k \sim v_i + g_v^k), k = 1, \ldots, K \\
  h_{t_j}^k &= s(W_t^j \sim t_j + g_t^k), k = 1, \ldots, K \\
  b_{v_i}^k &= \text{sign}(h_{v_i}^k), k = 1, \ldots, K \\
  b_{t_j}^k &= \text{sign}(h_{t_j}^k), k = 1, \ldots, K
\end{align*}
\]

$h_{v_i}^k(h_{t_j}^k)$: layer-wise hash representation

$b_{v_i}^k(b_{t_j}^k)$: layer-wise binary hash codes
Regularized Cross-modal Hashing

Layer-wise Semantic Similarity Preserving

Objective function (negative log likelihood):

\[ \Gamma_1 = - \sum_{k=1}^{K} \tau_k \sum_{i,j=1}^{N} \frac{S^k_{ij}}{S^k_{ij}} - \log(1 + e^{\phi^k_{ij}}) \]

- Ground Truth
  - \( S^k_{ij} = 1 \) Same label at the k-th layer
  - \( S^k_{ij} = 0 \) Different label at the k-th layer

Layer Confidence

Semantic Similarity

\[ \phi^k_{ij} = \frac{1}{2} (h^k_v)^T h^k_{ij} \]
Framework

- Regularized Cross-modal Hashing
- Binarization Difference Penalizing

To derive the optimal continuous surrogates of the hash codes

\[
B^k_v = \text{sgn}(H^k_v)
\]

\[
B^k_t = \text{sgn}(H^k_t)
\]

\[
\Gamma_2 = \sum_{K=1}^{k} \alpha (\|B^k_v - H^k_v\|_F^2 + \|B^k_t - H^k_t\|_F^2) + \beta (\|H^k_v a\|_2^2 + \|H^k_t a\|_2^2)
\]

Binarization Difference Regularization

Information Maximization
Framework

Hierarchical Discriminative Learning

- Objective function (negative log likelihood):

\[
p_{v_i}^k = \text{soft max}(U_v^k h_{v_i}^k + q_v^k), k = 1, \ldots, K
\]

\[
p_{t_j}^k = \text{soft max}(U_t^j h_{t_j}^k + g_t^k), k = 1, \ldots, K
\]

\[
\psi_h = - \sum_{k=1}^{K} \rho_k \sum_{i=1}^{N} [(y_i^k)^T \log(p_{v_i}^k) + (y_{i}^k)^T \log(p_{t_j}^k)]
\]

Layer Confidence

Ground-truth
Framework

Final Objective Function

$$\min_{B^k, \Phi_v, \Phi_t} \gamma \psi_r + (1 - \gamma) \psi_h$$

- Regularized Cross-modal Hashing
- Hierarchical Discriminative Learning

Non-negative Tradeoff Parameter
Experiment

Dataset

- **Two datasets**: FashionVC (public) and Ssense (created by ourselves).
- **Raw data**: 25,974 image-text instances with hierarchical labels.
- **Preprocessing**: Removed the noisy instances that involve multiple items. Filtered out the categories with less than 70 instances.
Experiment

Dataset

- **Two datasets**: FashionVC (public) and Ssense (created by ourselves).

<table>
<thead>
<tr>
<th></th>
<th>FashionVC</th>
<th>Ssense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>16,862</td>
<td>13,696</td>
</tr>
<tr>
<td>Retrieval Set</td>
<td>16,862</td>
<td>13,696</td>
</tr>
<tr>
<td>Query Set</td>
<td>3,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Total Labels</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>The First Layer Labels</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>The Second Layer Labels</td>
<td>27</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1: Statistics of our datasets.
Experiment

Dataset

- FashionVC Label Hierarchy: 35 categories with two layers
Experiment

Dataset

- Ssense Label Hierarchy: 32 categories with two layers
Experiment

- **Experiment Setting**

- **Task**
  - Image to Text
  - Text to Image

- **Protocol**: Mean Average Precision

- **Baselines**
  - Shallow Learning: CCA, SCM-Or, SCM-Se, DCH
  - Deep Learning: CDQ, SSAH, DCMH

- 500-D SIFT Features and 4096-D Deep Features
### Experiment

#### On Model Comparison

**Table 2:** The MAP scores of different methods on two datasets. The shallow learning baselines use the SIFT features.

<table>
<thead>
<tr>
<th>Method</th>
<th>FashionVC</th>
<th></th>
<th></th>
<th>Ssense</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image→Text</td>
<td>Text→Image</td>
<td></td>
<td></td>
<td>Image→Text</td>
<td>Text→Image</td>
</tr>
<tr>
<td></td>
<td>16bits</td>
<td>32bits</td>
<td>64bits</td>
<td>128bits</td>
<td>16bits</td>
<td>32bits</td>
</tr>
<tr>
<td>CCA</td>
<td>0.150</td>
<td>0.130</td>
<td>0.114</td>
<td>0.103</td>
<td>0.141</td>
<td>0.126</td>
</tr>
<tr>
<td>SCM-Or</td>
<td>0.176</td>
<td>0.128</td>
<td>0.109</td>
<td>0.095</td>
<td>0.159</td>
<td>0.121</td>
</tr>
<tr>
<td>SCM-Se</td>
<td>0.353</td>
<td>0.328</td>
<td>0.355</td>
<td>0.217</td>
<td>0.254</td>
<td>0.288</td>
</tr>
<tr>
<td>DCH</td>
<td>0.224</td>
<td>0.246</td>
<td>0.266</td>
<td>0.312</td>
<td>0.209</td>
<td>0.254</td>
</tr>
<tr>
<td>CDQ</td>
<td>0.456</td>
<td>0.583</td>
<td>0.559</td>
<td>0.621</td>
<td>0.445</td>
<td>0.593</td>
</tr>
<tr>
<td>SSAH</td>
<td>0.609</td>
<td>0.661</td>
<td>0.703</td>
<td>0.391</td>
<td>0.724</td>
<td>0.794</td>
</tr>
<tr>
<td>DCMH</td>
<td>0.552</td>
<td>0.579</td>
<td>0.603</td>
<td>0.623</td>
<td>0.589</td>
<td>0.638</td>
</tr>
<tr>
<td>HiCHNet</td>
<td>0.611</td>
<td>0.688</td>
<td>0.721</td>
<td>0.715</td>
<td>0.818</td>
<td>0.871</td>
</tr>
</tbody>
</table>

↑ 0.2% 2.7% 1.8% 9.2% 9.4% 7.7% 7.0% 20.6%

**Table 3:** The MAP scores of different methods on two datasets. The shallow learning baselines use the VGG-F features.

<table>
<thead>
<tr>
<th>Method</th>
<th>FashionVC</th>
<th></th>
<th></th>
<th>Ssense</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image→Text</td>
<td>Text→Image</td>
<td></td>
<td></td>
<td>Image→Text</td>
<td>Text→Image</td>
</tr>
<tr>
<td></td>
<td>16bits</td>
<td>32bits</td>
<td>64bits</td>
<td>128bits</td>
<td>16bits</td>
<td>32bits</td>
</tr>
<tr>
<td>CCA</td>
<td>0.217</td>
<td>0.197</td>
<td>0.182</td>
<td>0.162</td>
<td>0.243</td>
<td>0.224</td>
</tr>
<tr>
<td>SCM-Or</td>
<td>0.256</td>
<td>0.176</td>
<td>0.141</td>
<td>0.121</td>
<td>0.278</td>
<td>0.185</td>
</tr>
<tr>
<td>SCM-Se</td>
<td>0.429</td>
<td>0.462</td>
<td>0.373</td>
<td>0.486</td>
<td>0.522</td>
<td>0.564</td>
</tr>
<tr>
<td>DCH</td>
<td>0.356</td>
<td>0.525</td>
<td>0.602</td>
<td>0.602</td>
<td>0.420</td>
<td>0.586</td>
</tr>
<tr>
<td>CDQ</td>
<td>0.456</td>
<td>0.583</td>
<td>0.559</td>
<td>0.621</td>
<td>0.445</td>
<td>0.593</td>
</tr>
<tr>
<td>SSAH</td>
<td>0.609</td>
<td>0.661</td>
<td>0.703</td>
<td>0.391</td>
<td>0.724</td>
<td>0.794</td>
</tr>
<tr>
<td>DCMH</td>
<td>0.502</td>
<td>0.579</td>
<td>0.603</td>
<td>0.623</td>
<td>0.589</td>
<td>0.638</td>
</tr>
<tr>
<td>HiCHNet</td>
<td>0.611</td>
<td>0.688</td>
<td>0.721</td>
<td>0.715</td>
<td>0.818</td>
<td>0.871</td>
</tr>
</tbody>
</table>

↑ 0.2% 2.7% 1.8% 9.2% 9.4% 7.7% 7.0% 18.0%
Figure 3: **HiCHNet-flat**: One derivative of our HiCHNet model.
实验

关于标签层次结构

![Performance of HiCHNet and HiCHNet-flat on FashionVC.](attachment:image.png)

图4: HiCHNet和HiCHNet-flat在FashionVC上的性能。
On Case Study 1

- Retrieve from the whole retrieval set

<table>
<thead>
<tr>
<th>Text Query</th>
<th>HiCHNet</th>
<th>DCMH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black Matches Loafers.</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
<tr>
<td>Black Large Jaw Backpack.</td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>Hotel Diamond Tag Keychain.</td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure 5: Illustration of ranking results from the whole retrieval set. The irrelevant images are highlighted in red boxes.
On Case Study 2

- Retrieve from the constrained subset of 10 images of different categories.

Figure 6: Illustration of ranking results from the constrained retrieval set.
Conclusion

- We first validate the benefits of utilizing the category hierarchy in cross-modal.
- We propose a novel supervised hierarchical cross-modal hashing framework.
- We build a large-scale benchmark dataset from the global fashion platform Ssense. Extensive experiments demonstrate the superiority of HiCHNet over the state-of-the-art methods.
Thanks for the travel grant from SIGIR.

Email: sunchangchang123@gmail.com
Back Up
On Category Analysis

Figure 7: Performance of HiCHNet and DCMH on different categories of FashionVC and Ssense in the task of “Text→Image”.
On Component Analysis

Figure 8: Sensitivity analysis of the hyper-parameters.