Towards Semantic Embedding in Visual Vocabulary

Session: Object Recognition II: using Language, Tuesday 15, June 2010

Rongrong Ji\textsuperscript{1}, Hongxun Yao\textsuperscript{1}, Xiaoshuai Sun\textsuperscript{1}, Bineng Zhong\textsuperscript{1}, and Wen Gao\textsuperscript{1,2}

\textsuperscript{1}Visual Intelligence Laboratory, Harbin Institute of Technology
\textsuperscript{2}School of Electronic Engineering and Computer Science, Peking University
Overview

- Problem Statement
- Building patch-labeling correspondence
- Generative semantic embedding
- Experimental comparisons
- Conclusions
Overview

- Problem Statement
- Building patch-labeling correspondence
- Generative semantic embedding
- Experimental comparisons
- Conclusions
• Building visual codebooks
  – Quantization-based approaches
    • K-Means, Vocabulary Tree, Approximate K-Means et al.
  – Feature-indexing-based approaches
    • K-D Tree, R Tree, Ball Tree et al.
• Refining visual codebooks
  – Topic model decompositions
    • pLSA, LDA, GMM et al.
  – Spatial refinement
    • Visual pattern mining, Discriminative visual phrase et al.
• With the prosperity of Web community
• Problems to achieve this goal
  – Supervision @ image level
  – Correlative semantic labels
  – Model generality

A traditional San Francisco street view with car and buildings

Flower

Rose
• Our contribution
  – Publish a ground truth path-labeling set
    http://vilab.hit.edu.cn/~rrji/index_files/SemanticEmbedding.htm
  – A generalized semantic embedding framework
    • Easy to deploy into different codebook models
  – Modeling label correlations
    • Model correlative tagging in semantic embedding
Introduction

- The proposed framework

```
• Preliminary
  - Flickr Photos Clawing
  - Local Feature Extraction

• Correspondence for Embedding
  - Building Patch-Label Correspondence
  - Correspondence Purification based on DDE
  - Semi-Supervised Clustering via Markov Random Field Modeling
  - Building Visual Vocabulary Model

• Semantic Embedding
  - Semantic Embedding as Gibbs Distribution in Hidden Field
  - Modeling WordNet-based Concept Correlations

Supervised visual codebook construction
```
Overview

- Problem Statement
- Building patch-labeling correspondence
- Generative semantic embedding
- Experimental comparisons
- Conclusions
Building Patch-Labeling Correspondence

- Collecting over 60,000 Flickr photos
  - For each photo
    - DoG detection + SIFT
- “Face” labels (for instance)
• Purify the path-labeling correspondences
  – Density-Diversity Estimation (DDE)
  – Formulation
    • For a semantic label $s_i$, its initial correspondence set is
      
      \[
      < D_i, s_i > = \{ d_1, d_2, \ldots, d_n \}, s_i >
      \]

      \[
      = \{ le_{d_1 s_i}, le_{d_2 s_i}, \ldots, le_{d_n s_i} \}
      \]

      \[
      D_i = \{ d_1, d_2, \ldots, d_n \}
      \]

      – Patches extracted from images with label $s_i$

      \[
      le_{ij}
      \]

      – A correspondence from label $s_i$ to local patch $d_j$
• Density
  – For a given $d_l$, **Density** $Den_{d_l}$ reveals its representability for $s_i$:

$$Den_{d_l} = \frac{1}{m} \sum_{j} \exp(\|d_l - a_j\|_{L2})$$

Average neighborhood distance in $m$ neighbors

$n_m$: number of images in neighborhood

$n_i$: number of total images with label $s_i$

• Diversity
  – For a given $d_l$, **Diversity** $Div_{d_l}$ reveals its unique score for $s_i$:

$$Div_{d_l} = -\frac{n_m}{n_i} \ln\left(\frac{n_m}{n_i}\right)$$
$D^\text{Purify}_i = \{d_j \mid DDE_{d_j} > T\}$

s.t. $DDE_{d_j} = Den_{d_j} \times Div_{d_j}$

High $T$: Concerns more on Precision, not Recall
• Case study: “Face” label (before DDE)
• Case study: “Face” label (after DDE)
Overview

- Problem Statement
- Building patch-labeling correspondence
- Generative semantic embedding
- Experimental comparisons
- Conclusions
Generative Hidden Markov Random Field

- A Hidden Field for semantic modeling
- An Observed Field for local patch quantization
• Hidden Field \( S = \{s_i\}_{i=1}^m \)
  – Each \( s_i \) produces correspondence links (le) to a subset of patch in the Observed Field
  – Links \( l_{ij} \) denotes semantic correlations between \( s_i \) and \( s_j \)
- **Observed Field** $D = \{d_i\}_{i=1}^n$
  - Any two nodes follow visual metric (e.g. $L2$)
  - Once there is a link $le_{ij}$ between $d_i$ and $s_j$, we constrain $d_i$ by $s_j$ from the hidden field
• In the ideal case

  – Each $d_i$ is conditionally independent given $S$:

    $$P(D \mid S) = \prod_{i \in m} \{P(d_i \mid s_j) \mid P(d_i \mid s_j) \neq 0\}$$

  Feature set $D$ is regarded as (partial) generative from the Hidden Field
– Formulize Clustering procedure

• Assign a unique cluster label $c_i$ to each $d_i$

• Hence, $D$ is quantized into a codebook $W = \{w_k\}_{k=1}^K$ with corresponding features $V = \{v_k\}_{k=1}^K$

• Cost for codebook candidate $C$

\[
P(C|D) = \frac{P(C)P(D|C)}{P(D)}
\]
• Semantic Constraint $P(C)$
  – Define a MRF on the Observed Field

  \[ P(C) = \frac{1}{\mathcal{H}} \exp(-\mathcal{L}(C)) = \frac{1}{\mathcal{H}} \exp(-\sum_{k=1}^{K} \mathcal{L}_{N_k}(w_k)) \]

  – For a quantizer assignments $C$, its probability can be expressed as a Gibbs distribution from the Hidden Field as

  \[ \forall i \quad P(c_i|C) = P(c_i|\{c_j | le_{jx} \neq 0, le_{iy} \neq 0, x \in \mathcal{N}_y \}) \]
• That is
  – Two data points $d_i$ and $d_j$ in $w_k$ contribute to $\mathcal{L}_{N_k}(w_k)$ if and only if

  \[
P(C) = \frac{1}{\mathcal{H}} \exp \left( - \sum_{k=1}^{K} \mathcal{L}_{N_k}(w_k) \right)
  \]

  \[
  = \frac{1}{\mathcal{H}} \exp \left( - \sum_{k=1}^{K} \sum_{i \in N_k} \sum_{j \in N_k} \left\{ -l_{xy} \mid le_{xi} \neq 0 \land le_{yj} \neq 0 \right\} \right)
  \]
• Visual Constraint $P(D|C)$
  – Whether the codebook $C$ is visually consistent with current data $D$
  • Visual distortion in quantization

$$
\sum_{i=1}^{n} \{P(d_i, v_k) | c_i = w_k\} \\
\propto \exp \left( - \sum_{i=1}^{n} \{\text{Dis}(d_i, v_k) | c_i = w_k\} \right)
$$
• Overall Cost
  • Finding MAP of $P(C|D)$ can be converted into maximizing its posterior probability

$$P(C|D) \propto P(D|C)P(C) \propto \left( \sum_{i=1}^{n} \{P(d_i, v_k) | c_i = w_k\} \right) \times \left( \frac{1}{\mathcal{H}} \exp \left( \sum_{k=1}^{K} \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_k} \{l_{xy} | le_{xi} \neq 0 \land le_{yj} \neq 0\} \right) \right)$$
Generative Semantic Embedding

• EM solution
  
  - E step
    
    $$\text{Obj}(c_i|d_i) = \arg \min_k (-\text{Dis}(d_i, v_k) + \frac{1}{\mathcal{H}'} \sum_{j \in \mathcal{N}_k} \{l_{xy} | l_{e_{xi}} \neq 0 \land l_{e_{yj}} \neq 0\})$$
    
  - M step
    
    $$v_k = \frac{\sum_{d_i \in W_k} d_i}{\|W_k\|} \quad s.t. \quad W_k = \{d_i|c_i = w_k\}$$
  
  Assign local patches to the closest clusters

  Update the visual word center
Algorithm 1: Building Supervised Visual Vocabulary

1. **Input:** Visual data $D = \{d_i\}_{i=1}^n$, Semantic supervision $S = \{s_j\}_{j=1}^m$, Correspondence set $\{LE_1, ..., LE_m\}$, and Semantic correlation $l_{ij}$ for any two $s_i$ and $s_j$ calculated by WordNet::Similarity[27], Maximum iteration $N_I$.

2. **Pre-computing:** Calculate the nearest neighbors in the Hidden Field using an $o(m^2)$ sequential scanning heap. Initialize a random set of clustering centers $\mathcal{W} = \{w_k\}_{k=1}^K$.

3. **Iterative EM Steps:**
   - **while** $\mathcal{V} = \{v_k\}_{k=1}^K$ still change or the number of iteration is within $N_I$ **do**
     - **E Step:** For each $d_i$ in $D$, assign $c_i = w_k$ that satisfying the objective function in Equation 13, in which nearest neighbors in the Hidden Field are obtained from pre-computing (Step 2).
     - **M Step:** For each $w_k$ in $\mathcal{W}$, update its corresponding feature vector $v_k$ based on Equation 14.
   - **end**

4. **Output:** Supervised vocabulary $C = \{c_k\}_{k=1}^K$ with its inverted indexing structure (Indexed after EM).
Generative Semantic Embedding

• Model generation
  – To supervised codebook with label independence assumption
    \[ P(C) = \frac{1}{\mathcal{H}} \exp \left( - \sum_{N_i \in N} L_{N_i}(w_i) \right) \]
    \[ = \frac{1}{\mathcal{H}} \exp \left( - \sum_{k=1}^{K} \sum_{i \in N_k} \sum_{j \in N_k} \{ -[l_{xy}]_{le_{xi} \neq 0 \land le_{yj} \neq 0} \} \right) \]
  – To unsupervised codebooks
    • Making all \( l=0 \)
Overview

- Problem Statement
- Building patch-labeling correspondence
- Generative semantic embedding
- Experimental comparisons
- Conclusions
Case study of ratios between inter-class distance and intra-class distance with and without semantic embedding in the Flickr dataset.
MAP@1 comparisons between our GSE model to Vocabulary Tree [1] and GNP [34] in Flickr 60,000 database.
Experimental Comparisons

Confusion table comparisons on PASCAL VOC dataset with method in [24].
Overview

- Problem Statement
- Building patch-labeling correspondence
- Generative semantic embedding
- Experimental comparisons
- Conclusions
Conclusion

• Our contribution
  • Propagating Web Labeling from images to local patches to supervise codebook quantization
  • Generalized semantic embedding framework for supervised codebook building
  • Model correlative semantic labels in supervision

• Future works
  • Adapt one supervised codebook for different tasks
  • Move forward supervision into local feature detection and extraction
Thank you!