VSB2-Net: Visual-Semantic Bi-Branch Network for Zero-Shot Hashing

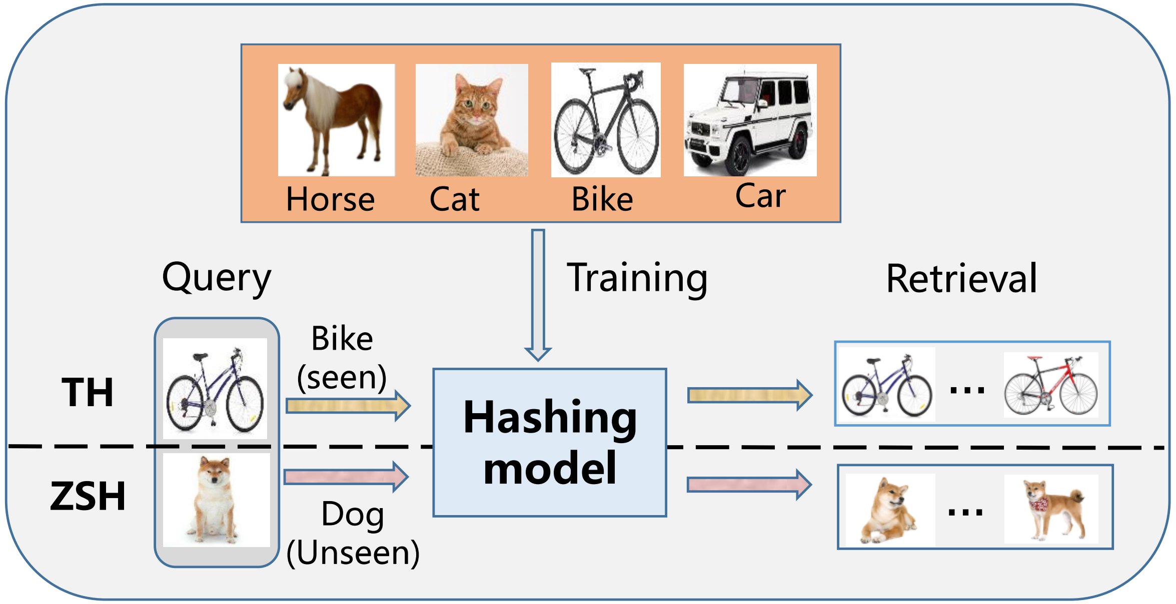
Xin Li, Xiangfeng Wang, Bo Jin, Wenjie Zhang, Jun Wang and Hongyuan Zha

Outline

- Introduction
- Motivation
- Framework
- Experiments

Introduction

- The definition between traditional hashing and zero-shot hashing

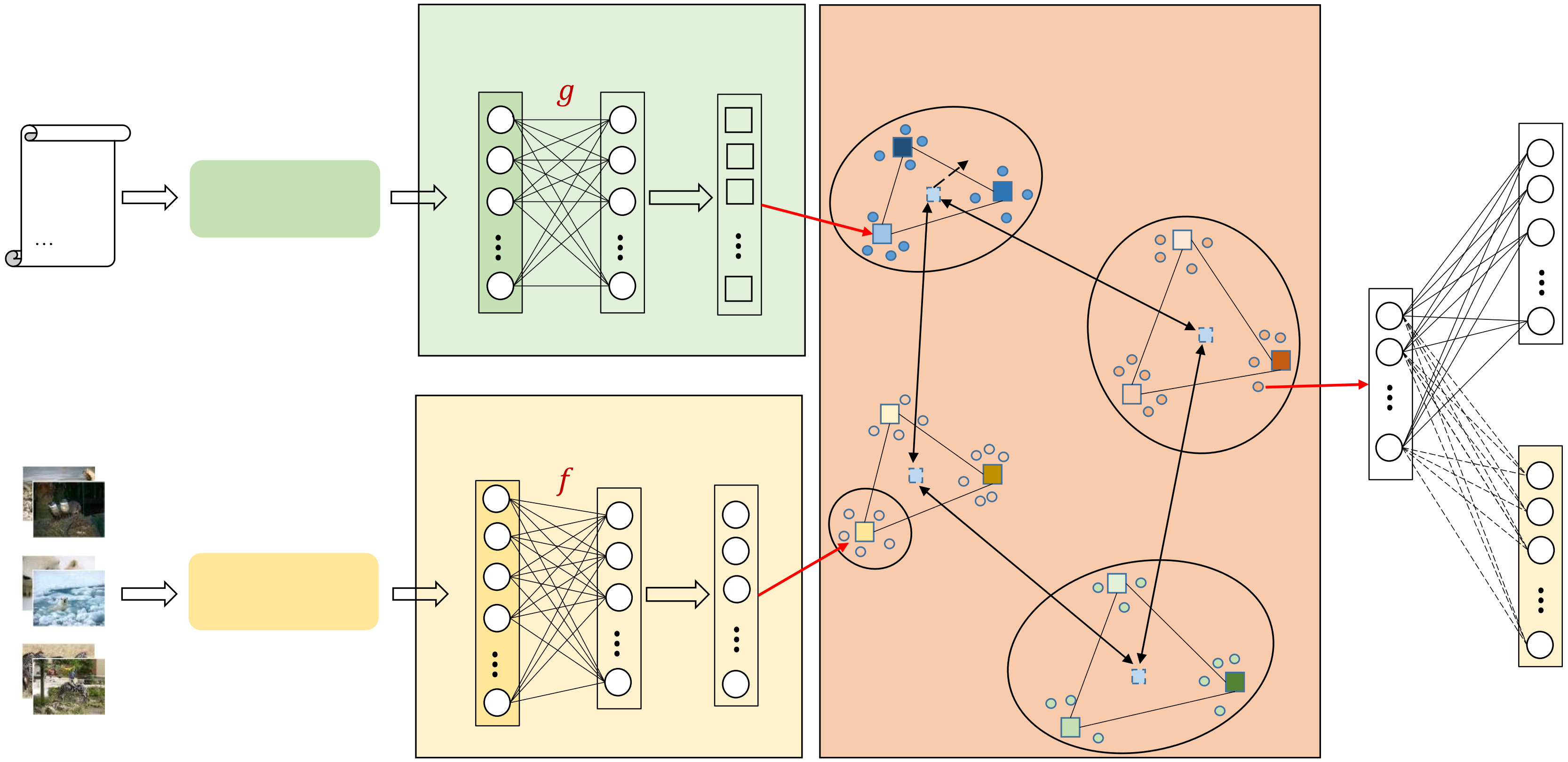


Motivation

- The existing methods mainly focused on optimizing the mapping between hash codes and semantic space, but ignored the core of the hash problem that generates discriminative hash codes.

- The hash codes are binary-like, and have lower representative ability compared with the high-dimensional visual feature vectors, which makes it difficulty to distinguish similar classes in hamming space.

Framework



- The architecture is a bi-branch network, which includes the semantic similarity branch and the visual feature transfer branch.
- The reconstruction module and classification module are directly employed to enhance the generalization and transfer abilities on unseen classes.

Framework

- Semantic Similarity Branch Network
  - Computing cosine similarity between semantic word vectors
  - Applying K nearest neighbors technique to separate all L training word vectors into T groups
  - utilizing a barycenter-based fisher criteria to maximize the distance between two word vector groups and maintain the similarity relationships within each group
- Loss function is designed as follows

$$
\begin{aligned}
\mathcal{L}_s = {} & \alpha \sum_{t=1}^{T} \sum_{w^{\ell_i}, w^{\ell_j} \in \mathcal{N}_t} \left\| \mathrm{Sim}(w^{\ell_i}, w^{\ell_j}) - \mathrm{Sim}(s^{\ell_i}, s^{\ell_j}) \right\|^2 \\
& + \beta \sum_{t_1=1}^{T} \sum_{t_2=1, t_2 \neq t_1}^{T} \max\left(0, \lambda - \left\| \mathcal{BC}(\mathcal{N}_{t_1}) - \mathcal{BC}(\mathcal{N}_{t_2}) \right\|^2 \right) \\
& + \gamma \sum_{i=1}^{T} \left\| |s^{\ell_i}| - e \right\|_1 ,
\end{aligned}
$$

Framework

- Visual Feature Transfer Branch Network
  - employing dot product metric between hash code and the target semantic vector to characterize the similarity relationships
- Loss function is designed as follows

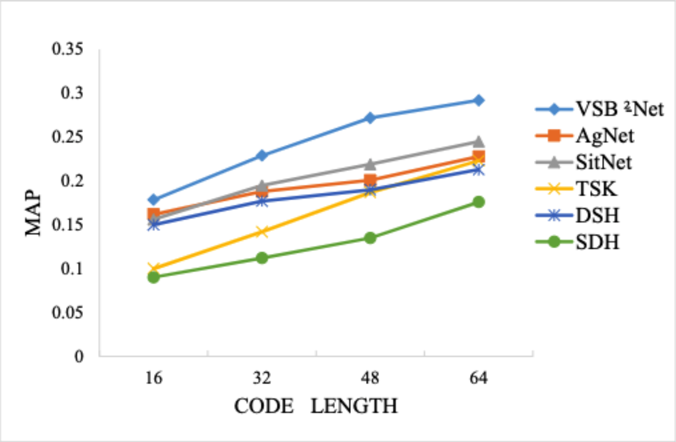
$$\mathcal{L}_t = \sum_{i=1}^{N} \max\left(0, m - b_i^{\mathrm{T}} * s^{\ell_i} + \max_{\ell_j \neq \ell_i} b_i^{\mathrm{T}} * s^{\ell_j}\right)$$

Framework

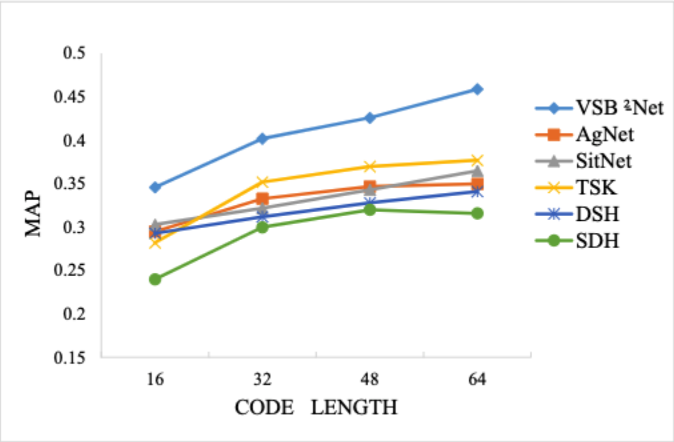
- Task-driven Regularization
  - reconstruction module: push hash codes reconstructing the visual features
  - classification module: drive the error hash codes to near the target semantic vector in order to reduce ambiguity
- Loss function is designed as follows

$$\mathcal{L}_t = \sum_{i=1}^{N} \max\left(0, m - b_i^{\mathrm{T}} * s^{\ell_i} + \max_{\ell_j \neq \ell_i} b_i^{\mathrm{T}} * s^{\ell_j}\right)$$

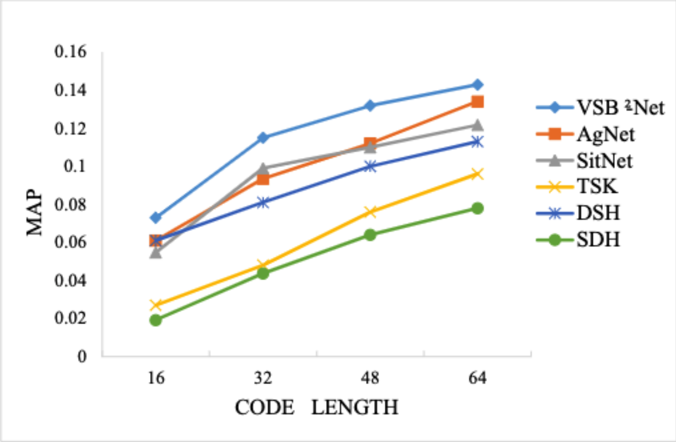
Experiments



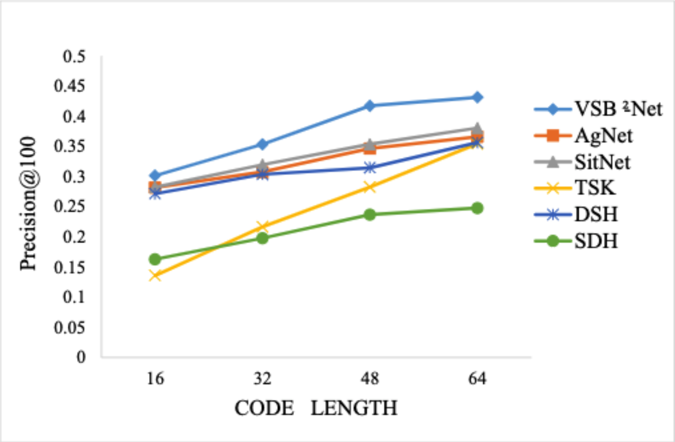
(a) AwA



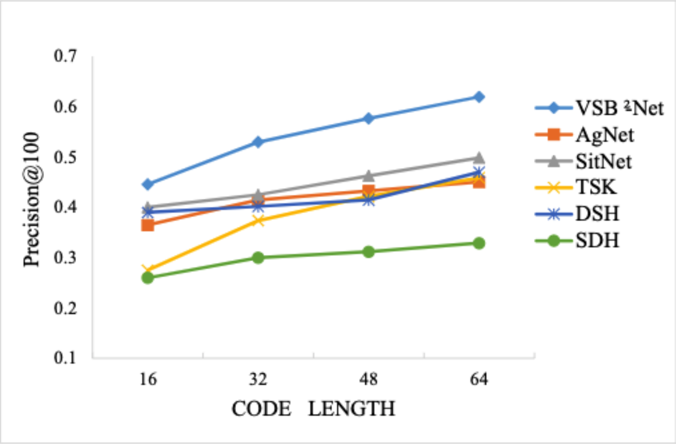
(b) APY



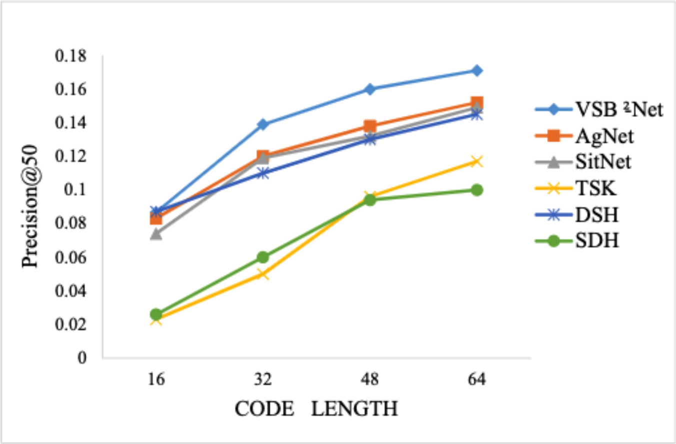
(c) CUB



(a) AwA

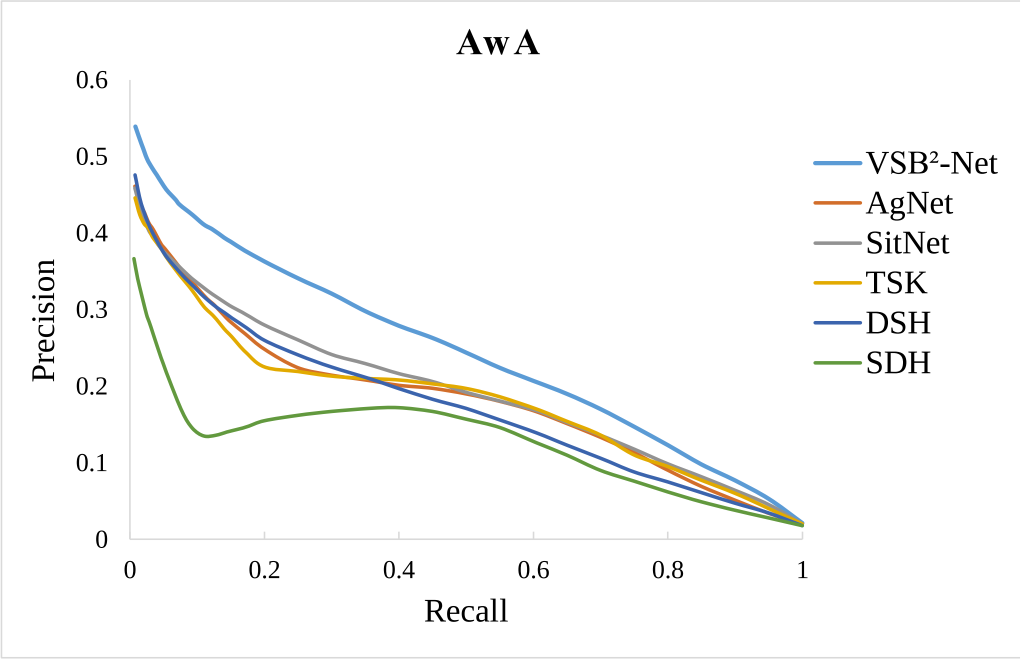


(b) APY



(c) CUB

Experiments



Precision-recall curves(64bits) of VSB2-Net on AwA dataset

Effects of different number of class for training and testing on ImageNet dataset.

Experiments

- MAP of two mapping modes w.r.t different numbers of bits on AwA.

Method	AwA(MAP)			
	16 bits	32 bits	48 bits	64bits
Mini-SitNet	0.128	0.150	0.175	0.192
Mini-VSB^2-Net	0.151	0.178	0.220	0.241

- MAP of two reconstructive space methods w.r.t different numbers of bits on AwA.

Method	AwA(MAP)			
	16 bits	32 bits	48 bits	64bits
VSB^2-G	0.155	0.198	0.217	0.270
VSB^2-Net	0.175	0.230	0.272	0.293

- Parameter verification with 64-bit hash codes on AwA.

MAP β / α	1	0.1	0.01
1	0.280	0.285	0.274
0.1	0.286	**0.292**	0.264
0.01	0.274	0.281	0.266

Thank you!