E²BoWs: An End-to-End Bag-of-Words Model via Deep Convolutional Neural Network

Xiaobin Liu[1], Shiliang Zhang[1], Tiejun Huang[1], Qi Tian[2]
[1]School of Electronic Engineering and Computer Science, Peking University, Beijing, 100871, China
{ xbliu.vmc, slzhang.jdl, tjhuang }@pku.edu.cn
[2]Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249-1604, USA
qitian@cs.utsa.edu

Abstract

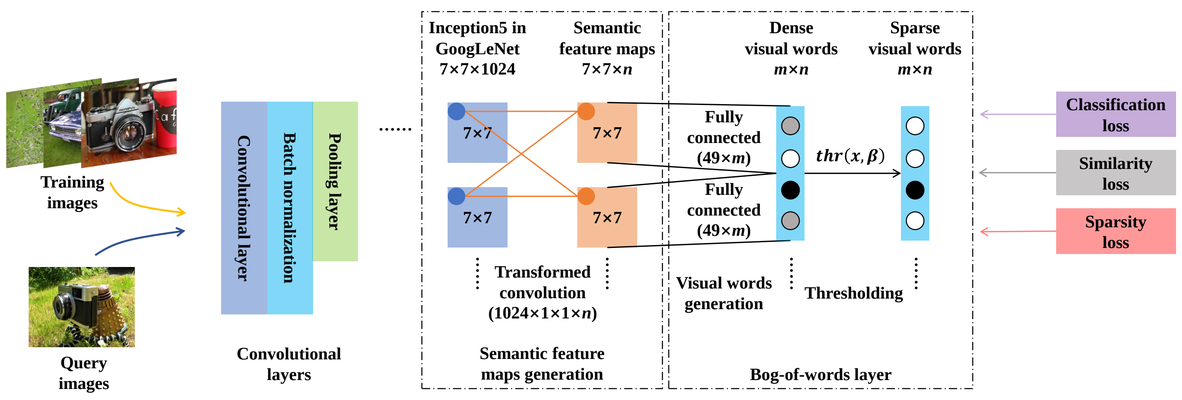
Traditional Bag-of-visual Words (BoWs) model is commonly generated with many relatively independent steps including and thus is hard to be jointly optimized. Moreover, the dependency on hand-crafted local feature makes it not effective in conveying high-level semantics. These issues largely hinder the performance of BoWs model in large-scale image applications. To conquer these issues, we propose an End-to-End BoWs (E²BoWs) model based on Deep Convolutional Neural Network (DCNN). Our model takes an image as input, then identifies and separates the semantic objects in it to generate semantic feature maps, and finally outputs the visual words with high semantic discriminative power for each feature map. We also introduce a novel learning algorithm to train the model, which further ensures the accuracy and efficiency of the retrieval system. Experimental results on several public datasets show that our method achieves promising accuracy and efficiency compared with recent deep learning based retrieval works.

Contributions

The major contributions of this work:

1. Out E2BoWs model is generated in an end-to-end manner, thus is more efficient and easier to be jointly optimized and tuned.
2. We incorporating DCNN into BoWs model, which is potential to bring higher discriminative power to semantics.
3. Visual words generated by proposed E2BoWs conveys clear semantic cues compared with DCNN based hash models.

Proposed Model

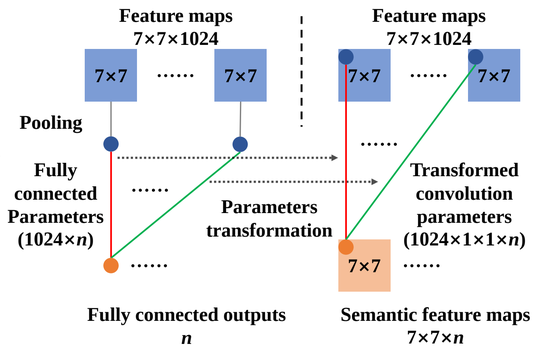


Given an input image $\mathcal{I}$, a vector of visual words v is generated directly by proposed model: $v = \mathcal{F}(\mathcal{I}, \theta)$, where $\mathcal{F}$ is the mapping function of proposed model and θ is parameters in E²BoWs model.

Specifically, we generate visual words of input images in following two steps:

1. Semantic feature map generation

We transform parameters in FC layer with n-way output into a convolutional layer. So that n semantic feature maps are generated for each input image corresponding to n training category.



2. Visual words generation

m visual words are generated from each semantic feature map, resulting in $m \times n$ visual words.

3. Thresholding

Visual words with small response values are discarded to further ensure retrieval efficiency. The procedure is formulated as follows with parameter β to be learned:

$$thr(x, \beta) = \begin{cases} x, & x > \beta \\ 0, & otherwise \end{cases}$$

Model training

◆ We expect the proposed model and generated visual words should have follows properties:

1. Training procedure converges fast.
2. Visual words preserve the similarity relationship among images for accuracy.
3. Visual words are sparse for efficiency.

Thus we design the overall objective function as follows:

$$L(\theta) = \ell_{cls} + \lambda_1 \ell_{tri} + \lambda_2 \ell_{spa}$$

ℓ_{cls}, ℓ_{tri}, ℓ_{spa} denote the loss of classification, triplet similarity and sparsity, respectively.

Triplet similarity loss: $\ell_{tri}(v_a, v_p, v_n) = max\{0, sim_{v_a}^{v_n} - sim_{v_a}^{v_p} + \alpha\}$

Sparsity loss: $\ell_{spa} = \hat{\rho} \log \frac{\hat{\rho}}{\rho} + (1 - \hat{\rho}) \log \frac{1 - \hat{\rho}}{1 - \rho}$

◆ Generalization ability improvement:

We change the threshold parameter in triplet similarity loss w.r.t each pair of images based on similarity among categories $S(c_1, c_2)$ as follows:

$$\alpha' = \frac{\alpha}{(1 + S(c_1, c_2))^2}$$

Experiments

◆ Performance on *CIFAR-10* and *CIFAR-100* (mAP)

Method	CIFAR-10	CIFAR-100
ITQ [39]	0.175	—
ITQ-CCA [39]	0.295	—
KSH [40]	0.315	—
SH [41]	0.132	—
MLH [42]	0.211	—
BRE [43]	0.196	—
CNNH [22]	0.522	—
CNNH+ [22]	0.532	—
DNNH [44]	0.581	—
DSH [21]	0.676	—
BHC [23]	0.897	0.650*
E²BoWs	**0.909**	**0.689**
E²BoWs-B	0.908	0.624

◆ Performance on *MIRFLICKR-25K* (NDCG@100)

ITQ-CCA [39]	KSH [40]	BHC [23]	E²BoWs	E²BoWs-b
0.402	0.350	0.510*	0.492	**0.526**

◆ Evaluation on generalization ability on *NUS-WIDE* (mAP)

Feature	GN_{1024}	GN_{1000}	GN_{1024}^{BN}	GN_{1000}^{BN}	E²BoWs
mAP	0.552	0.594	0.551	0.591	**0.599**
Feature	GN_{1024}-B	GN_{1000}-B	GN_{1024}^{BN}-B	GN_{1000}^{BN}-B	E²BoWs-B
mAP	0.388	0.549	0.326	0.543	**0.563**

◆ Retrieval efficiency on different datasets

Method		CIFAR-10	CIFAR-100	*MIRFLICKR-25K*
BHC [23]	ANO	480,000	480,000	406,944
E²BoWs	ANV	8.64	10.6	43.6
	ANI	960	110	975
	ANO	8,294	1,166	42,510

Acknowledgments

This work is supported by National Science Foundation of China under Grant No. 61572050, 1538111, 61620106009, 61429201, and the National 1000 Youth Talents Plan, in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

Code Link