Comparison of Different Novel Methods on Scene Classification Problem

Group 16

0856622 余家宏
309551067 吳子涵
309551122 曹芳驊

Index

- Motivation
- Method & Implementation
- Result
- Discussion

Motivation

- BoW (Bag of Words), simply counts the number of descriptors associated with each cluster
- Not consider much information among local descriptors of an image.
- We would like to further investigate different approaches in order to improve the image representation.
- We also train two end-to-end networks, VGG16 and ResNet34, to perform classification as a comparision.

Method & Implementation

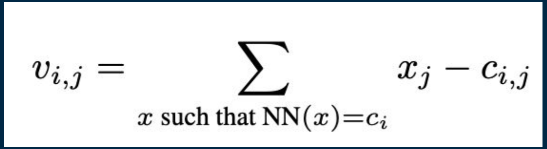
- We have tried the following method and compared the result:
  - VLAD+SVM
  - NetVLAD+SVM
  - VGG16 end-to-end
  - ResNet34 end-to-end

Method & Implementation

- VLAD

  The idea of the VLAD descriptor is to accumulate the differences x-ci of the local descriptor x assigned to ci for each cluster center. This characterizes the distribution of the vectors with respect to the center.

- the output representation would be :



$$v_{i,j} = \sum_{x \text{ such that } \mathrm{NN}(x)=c_i} x_j - c_{i,j}$$

Method & Implementation

- After the output vector V, we perform the SSR-normalization(1) and L2-normalization inorder.

$Sign(xi)\sqrt{|\,xi\,|}$ (1) $x/\sqrt{\sum_{i} xi^{2}}$ (2)

- We also perform the PCA to do reduce the dimension with output D = 128, 256, 512, 1024.

Method & Implementation

- NetVLAD is a generalized VLAD-like CNN architecture layer.
- The VLAD formula can be represented as (1), the term ak(xi) is 0,1 assignment. NetVLAD replace it as a softmax operation (2)
- It makes the whole formula differentiable and can be build with CNN

$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i)\left(x_i(j) - c_k(j)\right) \quad (1)$$

$$\bar{a}_k(\mathbf{x}_i) = \frac{e^{-\alpha\|\mathbf{x}_i - \mathbf{c}_k\|^2}}{\sum_{k'} e^{-\alpha\|\mathbf{x}_i - \mathbf{c}_{k'}\|^2}} \quad (2)$$

Method & Implementation

- We use the pre-train network on Pitts250k to generate the image representation.
- Then feed the output representation into SVM as previous part.
- Compare that using the network as the image representation or not to see the improvement of NetVLAD over VLAD.

Method & Implementation

- We also fine-tuning the VGG-16 and ResNet-34 on HW5 dataset.

- Chaning the output FC layer.
- Train the FC layer first, and then fine tune the entire network.

Results

- BoW: **54%**
- VLAD:

K \ PCA-D	128	256	512	1024
32	73.33	68.66	72.66	74
64	74.66	74.66	76	**76.66**

- NetVLAD: **86.66%**

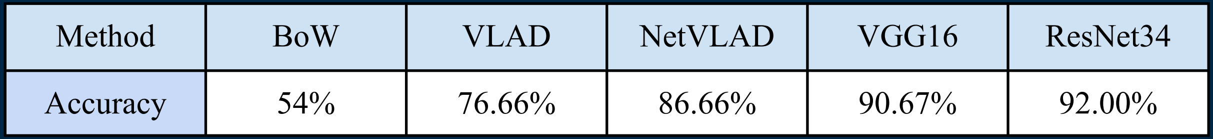
Results

- VGG
  - Batch size: 64
  - Train: 5 epochs
  - Finetune: 15 epochs
  - Result accuracy: **90.67%**

- ResNet
  - Batch size: 64
  - Train: 5 epochs
  - Finetune: 15 epochs
  - Result accuracy: **92.00%**

Discussion & Conclusion

- BoW is the basic method we used in HW 5.
- VLAD is an extension of BoW concept, and has better performance.
- NetVLAD can generate better description of images by learning the cluster centers with trainable network.
- VGG and ResNet are most widely used DL method nowaday, which have the best performance.



Method	BoW	VLAD	NetVLAD	VGG16	ResNet34
Accuracy	54%	76.66%	86.66%	90.67%	92.00%

Thanks for Listening