Efficiently Retrieving Images that We Perceived as Similar

Abstract

Despite growing interest in using sparse coding based methods for image classification and retrieval, progress in this direction has been limited by the high computational cost for generating each image's sparse representation. To overcome this problem, we leverage sparsity-based dictionary learning and hash-based feature selection to build a novel unsupervised way to efficiently pick out a query image's most important high-level features that can determine to which group we would visually perceived as similar. The preliminary results based on L1 feature map show the method's efficiency and high accuracy from the visual cognitive perspective. Finally, we consider a more general problem of how to make the pre-learned dictionary to adaptively refine the features contained according to past queries.

Motivation and Introduction

As the amount of digital data grows in unprecedented speed, new opportunities come with new challenges. The real value of big data lies in the ability to extract from it meaningful, even insightful information, rather than the "big" itself. Furthermore, many applications also require information to be retrieved *fast*. *Efficient* similar image retrieval thus becomes an important problem in the field of artificial intelligence with many real-world applications. The task is closely related to the nature of *human cognition* since any definition of *similarity* is meaningful only when it coincides with human feeling. Though the similar-or-not decision comes intuitively in no time for human, to find a well-defined decision guideline for computers is extremely hard. To resolve this stark discrepency that can inhibit human-machine cooperation, in this paper, we propose a novel method that emulates several important aspects of actual neurophysiological mechanisms, including *sparse coding* in primary visual cortex (V1), *synaptic plasticity*, and *mutual inhibition* between neurons. The mechanisms are not chosen randomly, instead, they complement each other's weak points for overall improvements without sacrificing efficiency.

Given unlabeled data, *sparse coding* provides a class of algorithms capable of extracting higher-level features that are actually more cognitively effective than hand-picked ones by emulating partial activity of neurons. The features can be regarded as the most representative building blocks by which the input data can be reconstructed most *efficiently* – highest accuracy with fewest elements used. The features form the *overcomplete* bases that resemble the *receptive fields* of neurons of in the visual cortex, making sparse coding a more appropriate medium than other widely-used computer vision features such as SIFT (Lowe 1999), GIST (Oliva et al. 2001), HOG (Dalal et al. 2005) etc., to bridge human cognition and algorithmic way of learning.

Unfortunately, though there are works (Ge et al. 2013) that tackle the similar image search problem by first representing query images as their sparse codes, the high computational cost involved in sparse code calculation renders it infeasible for (near) realtime retrieval tasks. Thus, we proposed in this paper a novel approach to overcome the performance bottleneck by a change of perspective: equipped with already-learned features, can we filter out the dominant features *directly* from an image, without resorting to the complete set of its sparse code? By regarding those selected features collectively as the image's *hash value*, the method actually transforms the sparse coding based approach to one more similar to efficient hash-based methods.

Considering the fact that sparse representation is based on an *overcomplete* basis (Olshausen and Field 1996), we use concepts similar to that of inner-products and orthogonal decompositions to aid our feature selection. By combining the inhibition-like decomposition scheme with an additional mechanism that emulates synaptic plasticity, we are able to come up with a retrieval system with hash-based-like efficiency, and performance comparable to GIST-based approach. Experiments based on *SUN397 scene benchmark* confirmed the effectiveness and efficiency of our method.

Proposed Retrieval System

System framework

Given a query natural image, we firstly decorrelate the image to equalize the variance which is also employed in preprocessing for dictionary due to potential factual and corrupted and this also roughly simulate spatial-frequency reponse characteristic of retinal ganglion cells proposed by (Olshausen 1997) in our cognitive system. We then uniformly select several image patches to extract a certain pat-

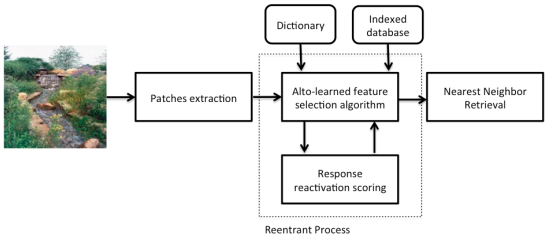


Figure 1: Proposed system.

tern of the image. We feed all extracted vectors into our auto-learned feature selection algorithms to encode the data. Finally, we use L2 distance as default metric to compute similarity score. The system diagram is shown in Fig. 1.

Offline Unsupervised Dictionary Learning

Sparse-based dictionary learning has proven been effective in natural images which are mostly scene image. Given input unlabeled scene images, the effective sparse coding proposed by (Lee et al. 2007) captures succinct feature with higher meanings and generate a dictionary with overcomplete bases which are effective to represent the image in data set given the corresponding sparse code. The basic descriptions such as edges and line segments are efficiently encoded into atoms of dictionary so we will pre-trained the dictionary as our dimension reduction projection bases.

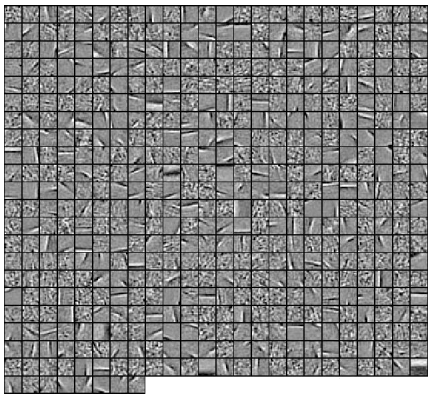


Figure 2: The Learned Dictionary

Auto-learned feature selection algorithms

Since our retrieval framework encode the image pattern of natural images into sparsity-based dictionary, we are motivated to select effective feature, especially those have high response to patches of natural images. Inspired by localitive sensitive hashing proposed by (Andoni et al. 2008), where high dimensional data can be projected to lower dimensional space with similarity preserving promise, we propose our novel algorithm to find out the atom of feature pattern in the dictionary to perform our hash-based dimensional reduction.

Firstly, we project our patches vector onto the atom of dictionary to get the highest values of the result for each vector of patches and have another zero array with the same size. We call those atoms strong responsive to the corresponding patches vector. Then, we set the value of each patches vector at corresponding atom of dictionary to be one.

Secondly, we substract the strong responsive atom from the corresponding patches vector in order to select second strong responsive atom with respect to the corresponding patches vector.

Iteratively, we will rank out the top n strong responsive atoms as our output for each patch vector. By this way, we can encode the raw data directly by the ranking of the response of corresponding atom based on sparsity based dictionary and we will show that the result has some effects consistent with our visual system.

```
Projection = abs(Data*Dictionary)
loop{
    id_column = max_column {projection}
    for i in all patches{
        outputCode(i, id_colmun) = 1
        maxValue = projection(i, id_column)
        reduction(:, i) = maxValue*Dictionary(:, id_column)
    }
    newData = newData – reduction';
    projection = abs(newData*Dictionary)

}
```

Figure 3: Auto-learned feature selection algorithm (ALFSA)

Response Reactivation Scoring

```
loop{
    scoring = zeros( length of total number );
    for i = 1 : reentrantNumber{
        index = index of closet element
        scoring(index) = scoring(index)*1.2
    }
    scoring = scoring*0.9
}
```

Figure 4: Response Reactivation Scoring

Experimental results

We evaluate our approach on a subset of scene images which is a version of MIT SUN dataset, SUN397 scene benchmark, from (!!!). Our subdataset consist of 3,583 scene images that have been grouped into 10 different classes: bamboo_forest, beach, botanical_garden, corridor, cottage_garden, hayfield, mountain_snowy, waterfallBlock, wheat_field, wine_cellearBarrelStorage. Original images in the dataset have different sizes so we resize them into the size of 200x200 pixels. Rather than represent them with

the state of the art manual-turned feature, we extract small 14x14 pixels image patches directly by uniform random selection. We call our auto-learned feature selection method ALFS and we will evaluate our method in two parts, the improvement, from naive method to our design alogrithm showing the progress,under our sparsity dictionary and the comparison with the well known human-turned global feature, GIST.

Improvement under sparsity dictioanry framework

Under our sparsity dictioanry framework inspired from neuron activity, we implemented two encoding method: one is our novel ALFS algorithm, which is called Neuron_ALFS in figures. Another one is inspired from localitive sensitive hashing method, projecting the raw images patches onto the learned dictionary, which is our first method to explore the effectiveness of image retrieval under such a novel sparsity dictionary framework. Although LSH-based method requires Gaussian random distribution, it also works fair to be discriminative under our sparse coding framework by simple hash projection on learned dictionary. While we apply this method on our learned basis vector with normal distribution, certain latent similar feature seems to be preserved after the projection to retrieve similar images. We call this naive idea inspired from traditional LSH as Neuron_LSH in figures.

Comparison with Human-tuned feature methods

Due to our scene dataset, to be fair, we employ GIST features to do the evaluation. For comparison, we extract GIST features proposed by (!!!) directly from 200x200 pixels images Due to the most state of the art working under different framework from us, we compare our ALFS method with LSH-based scheme under our spetial sparse coding framework as the baseline and we will show how much we have improved under this novel framework for image retrieval as an example.

Conclusion

Rather than traditional human-turned feature extraction our cognitive system based on sparse coding successfully combine proposed novel auto-learned feature selection algorithm with sparsity-based dictionary to create ouw own discriminative code to retrieve natural images with high performance. The sparsity-based dictionary which capture basic elements consisting a natural image is a well learned structure to encode images. Although it needs more powerful algorithm and research in large-scale image retrieval or other big data, this is the promising direction of relative application.

Discussion

How to work with big data?

When the world is filled with big data, effective approach is needed to deal with such a challenge. Large-scale image with effective and reliable performance is one of examples. Recently, we are attempting to address an open question if there is new approach based our framework to handle this

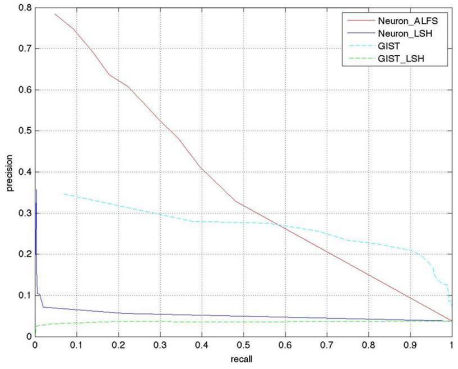
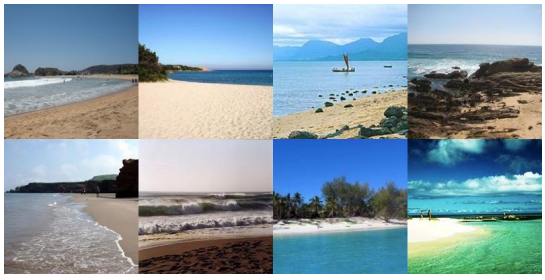


Figure 5: bamboo forest



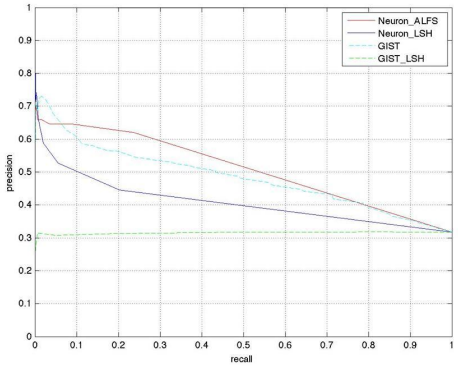


Figure 6: beach

old but not well-solved problem. Our work lies in how we design the connection between visual neuron encoding simulation and image retrieval problem and how we investigate an effective large-sale image retrieval new candidate.

Dictionary is trained off-line, and can take full advantage

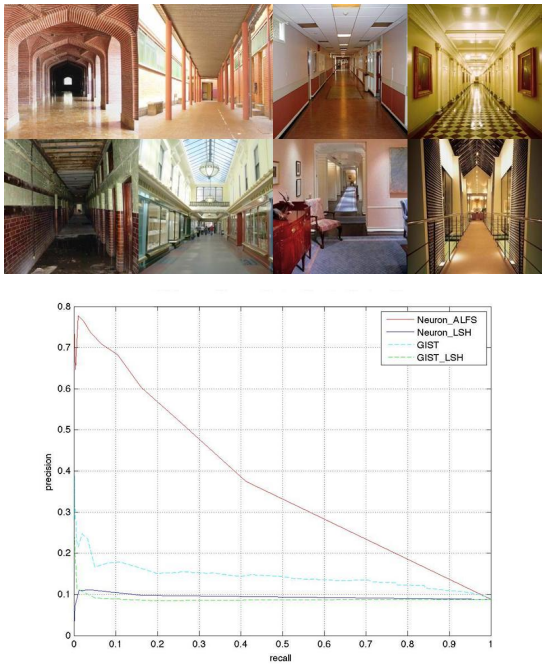


Figure 7: jpg

of the large amount of data.

References

Ge, Tiezheng, Qifa Ke, and Jian Sun. "Sparse-Coded Features for Image Retrieval." (2013).

Wright, John, et al. "Sparse representation for computer vision and pattern recognition." Proceedings of the IEEE 98.6 (2010): 1031-1044.

Sivaram, Garimella SVS, et al. "Sparse coding for speech recognition."Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

Olshausen, Bruno A., and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?." Vision research 37.23 (1997): 3311-3325.

Lee, Honglak, et al. "Efficient sparse coding algorithms." Advances in neural information processing systems 19 (2007): 801.

Lowe, David G. "Object recognition from local scale-invariant features."Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, 1999.

Oliva, Aude, and Antonio Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope." International journal of computer vision42.3 (2001): 145-175.

Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

CACM survey of LSH (2008): "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions" (by Alexandr Andoni and Piotr Indyk).Communications of the ACM, vol. 51, no. 1, 2008, pp. 117-122. directly from CACM