Compact Bilinear Pooling

Lina Zhu

June 6 ,2018

Abstract

Bilinear models have been shown to perform impressively on a wide range of visual tasks such as semantic subdivision, fine-grained recognition, and facial recognition recognition. However, bilinear features are high-dimensional, often on the order of hundreds of thousands to hundreds of millions, which makes them unsuitable for subsequent analysis. We propose that two compact bilinear representations have the same discriminative power as a full bilinear representation but only a few thousand dimensions. Our compact representation allows for backward propagation of classification, and the error of faults makes end-to-end optimization possible for visual recognition systems. The compact bilinear representation is a bilinear collection derived from a new core analysis that provides insight into the power of the bilinear pool of discrimination, as well as the study of further platform compact pooling methods.

1. Introduction

The coding and collection of visual features is an indispensable part of the semantic image analysis method.Rediscovered pioneered models and related efforts that typically involve a series of separate steps: function extraction, coding, merging, and classification; each is thoroughly investigated as a bag in many publications. Visual Words (BoVW) framework. Notable contributions include HOG[1] and SIFT descriptors, fisher coding, bilinear aggregation , and spatial pyramid, each of which significantly improves the accuracy of recognition. Recent results show that the gradient in the end-to-end reverse propagation convolutional neural network (CNN) can achieve joint optimization of the entire pipeline, thereby significantly improving the recognition accuracy. While the difference between these steps of CNN is not obvious from the BoVW pipeline, you can view the first few convolutional layers as feature extractors while the latter completely connects the layers as a pool and encoding mechanism. This has been recently explored in the method of combining feature extraction the architecture of the CNN paradigm encodes the steps from the BoVW paradigm. It is worth noting that Lin et al. recently achieved significant improvements in fine-grained visual recognition by completely replacing the fully connected layer bilinear aggregation. First, we propose two compact bilinear convergence methods that can reduce the performance of the two dimensions of the feature dimension with little loss to the complete bilinear convergence. Second, we show that back-propagation through compact bilinear aggregation can be efficiently computed, enabling an end-to-end optimized identification network. Third, we provide a new core point of view of the bilinearized pool that merely inspired the compact method proposed, but also provided theoretical insights into bilinear pooling.

2. Configurations of compact pooling

Both RM and TS pooling have a user defined projection dimension d, and a set of projection parameters, W. To investigate the parameters of the proposed compact bilinear methods, we conduced extensive experiments on the CUB200 dataset which contains 11,788 images of 200 bird species, with a fixed training and testing set split. We evaluate in the mode where part annotations are not provided at neither training nor testing time, and use VGG-M for all experiments in this section.Fig. 1 summarizes our results. As the projection dimension d increases, the two compact bilinear methods reach the performance of the full bilinear pooling. When not finetuned, the error of TS with d = 16K is 1.7% less than that of bilinear feature, while only using 6.1% of the original number of dimensions. When fine tuned, the performance gap disappears: TS with d = 16K has an error rate of 22.66%, compared to 22.44% of bilinear pooling.

The gap between the PCA-reduced bilinear feature and TS feature is large especially when the feature dimension is small and network not fine tuned(Table 1). When fine tuned, the gap shrinks but the PCA-Bilinear approach is not good at utilizing larger dimensions. For example, the PCA approach reaches a 23.8% error rate at 16K dimensions, which is larger than the 23.2% error rate of TS at 4K dimensions.

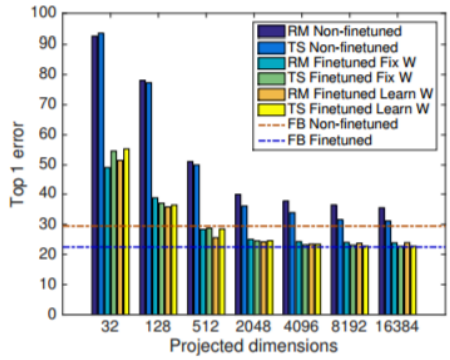

Figure 1. Classification error on the CUB dataset. Comparison of Random Maclaurin (RM) and Tensor Sketch (TS) for various combinations of projection dimensions and finetuning options.

Table 1. Comparison between PCA reduced feature and TS. Numbers refer to Top 1 error rates without and with fine tuning respectively.

dim.	256	1024	4096	16384
PCA	72.5/42.9	49.7/28.9	41.3/25.3	36.2/23.8
TS	62.6/32.2	41.6/25.5	33.9/23.2	31.1/22.5

3. Conclusion

We have modeled bilinear pooling in a kernelized framework and suggested two compact representations, both of which allow back-propagation of gradients for end-to-end optimization of the classification pipeline. Our key experimental results is that an 8K dimensional TS feature has the same performance as a 262K bilinear feature, enabling a remarkable 96.5% compression. TS is also more compact than fisher encoding, and achieves stronger results. We believe TS could be useful for image retrieval, where storage and indexing are central issues or in situations which require further processing: e.g. part-based models [2], conditional random fields, multi-scale analysis, spatial pyramid pooling or hidden Markov models; however these studies are left to future work. Further, TS reduces network and classification parameters memory significantly which can be critical e.g. for deployment on embedded systems. Finally, after having shown how bilinear pooling uses a pairwise polynomial kernel to compare local descriptors, it would be interesting to explore how alternative kernels can be incorporated in deep visual recognition systems.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 2