Low-rank Bilinear Pooling for Fine-Grained Classification

Shu Kong, Charless Fowlkes
Department of Computer Science
University of California, Irvine
{skong2,fowlkes}@ics.uci.edu

Fine-grained categorization aims to distinguish subordinate categories within an entry-level category, such as identifying the dog breed, bird species and aircraft models. Compared to general purpose visual categorization problems, fine-grained recognition focuses on the characteristic challenge of making subtle distinctions (low inter-class variance) despite highly variable appearance due to factors such as deformable object pose (high intra-class variance). One approach to dealing with such nuisance parameters has been to exploit strong supervision, such as detailed part-level, keypoint-level and attribute annotations [1]. However, such supervised annotations are costly to obtain, leading to efforts to utilize interactive learning [2] or partially supervised discovery of discriminative patches representing semantic parts [3]. An even more appealing possibility is to embed discriminative part learning within classification training [4] when only category labels are provided.

Recently, suprisingly good results have been produced using a simple method called *bilinear pooling* which collects second-order statistics of local features over a whole image to form a holistic representation for classification [5]. Spatial pooling introduces invariance to deformations while second-order statistics maintain selectivity. However, this yields very high-dimensional feature representations which impose substantial computational burdens and require large quantities of training data. Here we propose a simple simple strategy for utilizing bilinear features in conjunction with a low-rank classifier that addresses this concern.

Consider a linear classifier applied to bilinear pooled features (covariance) computed from data matrix $\mathbf{X}_i$. The parameters of this classifier are naturally represented as a matrix leading to a standard learning formulation as a support vector machine where the usual inner product is replaced by a trace operator:

$$\min_{\mathbf{W},b} \frac{1}{N}\sum_{i=1}^{N} \max(0, 1 - y_i \mathsf{tr}(\mathbf{W}^T\mathbf{X}_i\mathbf{X}_i^T) + b) + \frac{\lambda}{2}\|\mathbf{W}\|_F^2 \qquad (1)$$

By standard SVM duality, the optimal $\mathbf{W}$ will be a symmetric matrix which can be written as a weighted combination of the training examples $\mathbf{W}^* = \sum_{i:y_i=1} \alpha_i \mathbf{X}_i\mathbf{X}_i^T - \sum_{i:y_i=-1} \alpha_i \mathbf{X}_i\mathbf{X}_i^T$.

Our goal is to use a low-rank parameterization of $\mathbf{W}$ to reduce the number of model parameters and allow fast evaluation of $\mathsf{tr}(\mathbf{W}^T\mathbf{X}_i\mathbf{X}_i^T)$ without explicitly constructing the bilinear pooled features. We observe empircally that $\mathbf{W}^*$ learned on real datasets typically has many small magnitude eigenvalues (Fig 1) and classification accuracy does not suffer by using a low-rank approximation (2). We thus propose to parameterize $\mathbf{W}$ as a difference of positive semi-definite matrices $\mathbf{W} = \mathbf{U}_+\mathbf{U}_+^T - \mathbf{U}_-\mathbf{U}_-^T$ where $\mathbf{U}_+, \mathbf{U}_-$ are low rank (i.e., $\mathbf{U}_+, \mathbf{U}_- \in \mathcal{R}^{8\times 512}$). Not only does this reduce the degrees of freedom during parameter estimation, it also substantially reduces the computation required for evaluation since we can compute the classifier score without explicitly building the bilinear feature map. Specifically, when $\mathbf{W}$ is factored in this way we have

$$\mathsf{tr}(\mathbf{W}^T\mathbf{X}_i\mathbf{X}_i^T) = \|\mathbf{U}_+^T\mathbf{X}_i\|_F^2 - \|\mathbf{U}_-^T\mathbf{X}_i\|_F^2$$

When training a multi-class model, we can further reduce the number of parameters by first projecting the bilinear feature into a subspace shared across all classes prior to evaluating a particular classifier.

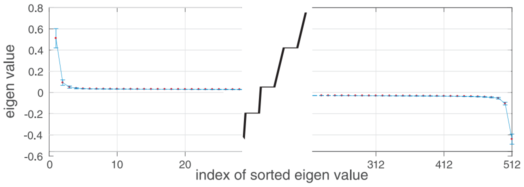

Figure 1: Spectrum of linear SVM parameters

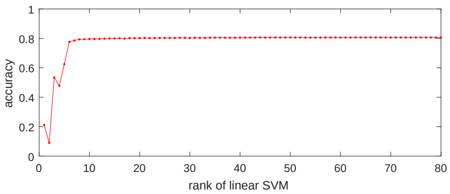

Figure 2: Accuracy of low-rank approximations

	Full Bilinear	Random Maclaurin	Tensor Sketch	Ours
Feature Dim.	c^2 [262K]	d [10K]	d [10K]	hwm [78K]
Feature Params.	0	$2cd$ [40MB]	$2c$ [4KB]	cm [200KB]
Classifier Params.	kc^2 [200MB]	kd [8MB]	kd [8MB]	krm [0.6MB]
Total Param.	kc^2 [200MB]	$2cd + kd$ [48MB]	$2c + kd$ [8MB]	$cm + krm$ [0.8MB]
Feature Eval.	$O(hwc^2)$	$O(hwcd)$	$O(hw(c + d\log d))$	$O(hwcm)$
Classifier Eval.	$O(kc^2)$	$O(kd)$	$O(kd)$	$O(khwmr)$
acc. CUB (%)	84.00	83.86	84.00	84.21

Table 1: Comparison of different bilinear models in terms of dimension, memory, and computational complexity. We compare to Gao *et al.* who propose two approaches based on polynomial kernel approximation [6]. The bilinear features are computed over feature maps of dimension $h \times w \times c$ for a k-way classification problem. d is the feature dimension in the approximate bilinear models using Random Maclaurin or Tensor Sketch. m is the reduced feature dimension of our model and r is the rank of our low-rank classifier. Numbers in brackets indicate typical value when bilinear pooling is applied after the last convolutional layer of VGG16 model with input image of size 448×448, *i.e.* $h = w = 28$, $c = 512$, $d = 10,000$, $m = 100$, $r = 8$ and applied to the CUB200-2011 bird dataset where $k = 200$. Model size only counts the parameters above the last convolutional layer. The classification accuracies ($\%$) of different methods are listed in the last row.

For class k we have $\mathbf{U}_{\pm}^{(k)} \approx \mathbf{P}\mathbf{V}_{\pm}^{(k)}$ where $\mathbf{P} \in \mathbb{R}^{c \times m}$ maps features of dimension c down to $m << c$ and $\mathbf{V}_{+}^{(k)}, \mathbf{V}_{-}^{(k)}$ are the parameters of the k^{th} low-rank classifier.

We use standard stochastic-gradient descent to train the resulting model in an end-to-end manner and achieving equivalent or better test performance to a full rank SVM. As an example, our model achieves 84.21% accuracy on Caltech-UCSD Birds 200 dataset [7], outperforming other state-of-the-art methods; moreover the parameter set learned by our model is ten times smaller than a recently proposed bilinear model [6], and hundreds times smaller than the standard bilinear CNN model [5]. Detailed comparison is shown in Table 1.

References

[1] Zhang, N., Shelhamer, E., Gao, Y, & Darrell, T. (2016), Fine-grained pose prediction, normalization, and recognition. *ICLR workshop*.

[2] Branson S., Perona P., Belongie S. (2011), Strong Supervision From Weak Annotation: Interactive Training of Deformable Part Models. *ICCV*.

[3] Wang, Y., Choi, J., Morariu, V. & Davis, L.S. (2016), Mining Discriminative Triplets of Patches for Fine-Grained Classification. *CVPR*.

[4] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015), Spatial transformer networks. *NIPS*.

[5] Lin, T. Y., RoyChowdhury, A. & Maji, S. (2015), Bilinear CNN models for fine-grained visual recognition. *CVPR*.

[6] Gao, Y., Beijbom, O., Zhang, N. & Darrell, T. (2016), Compact Bilinear Pooling. *CVPR*.

[7] Wah, C., Branson, S. Welinder, P., Perona, P. & Belongie, S. (2011), The Caltech-UCSD Birds-200. *Caltech Technical Report CNS-TR-2010-001*.