Scalable Multitask Representation Learning for Scene Classification

Maksim Lapin, Bernt Schiele
Max Planck Institute for Informatics

Matthias Hein
Saarland University

Abstract

We propose a multitask learning approach to jointly train a low-dimensional representation and the corresponding classifiers which scales to high-dimensional image descriptors, such as the Fisher Vector, and consistently outperforms the current state of the art on the SUN397 scene classification benchmark with varying amounts of training data.

1. Introduction

The underlying idea of multitask learning is that learning tasks jointly is better than learning each task individually. In particular, if only a few training examples are available for each task, sharing a jointly trained representation improves classification performance. While there has been significant progress in the area of multitask learning [1, 4, 7], most of the proposed methods do not scale well to very large feature dimensions. In this work we propose a new scalable formulation of multitask representation learning. It jointly learns a linear mapping into a lower dimensional subspace which is then used to build the classifiers for each class. The resulting optimization problem is solved via an adaptation of the recently developed stochastic dual coordinate ascent (SDCA) algorithm [9]. The proposed method is evaluated on the SUN397 scene classification benchmark [11] and consistently outperforms the current state of the art with varying amounts of training data and varying K when classification performance is measured via top-K accuracy.

2. Multitask Representation Learning

This section briefly introduces the proposed framework, which is described in more detail in [5]. We begin with some notation. Let $\{(x_i, y_{ti}) : 1 \le t \le T, 1 \le i \le n\}$ be the input/output pairs of the multitask learning problem, where $x_i \in \mathbb{R}^d$, $y_{ti} \in \{\pm 1\}$, T is the number of tasks, and n is the number of training examples per task. The setting we have in mind is that the feature space is high dimensional, which is quite common in computer vision problems, *e.g.* one has $d \ge 10^5$ with the Fisher Vector encoding [8].

In the proposed multitask representation learning framework we learn a matrix U in $\mathbb{R}^{d \times k}$ with $k \ll d$ which is used to map the original features x_i into a lower dimensional subspace via $U^\top x_i$. The linear predictors w_t are also trained in the subspace $\mathbb{R}^k$ and operate on the lower dimensional representation. The resulting problem is given below:

$$\min_{U \in \mathbb{R}^{d \times k}} \frac{1}{T} \sum_{t=1}^{T} \min_{w_t \in \mathbb{R}^k} P_{U,t}(w_t) + \frac{\mu}{2} \|U\|_F^2 , \qquad (1)$$

where the objective for task t given a fixed U is

$$P_{U,t}(w_t) = \frac{1}{n} \sum_{i=1}^{n} \left[1 - y_{ti} \left\langle w_t, U^\top x_i \right\rangle\right]_+ + \frac{\lambda}{2} \|w_t\|_2^2 ,$$

and $\lambda > 0$, $\mu > 0$ are the regularization parameters.

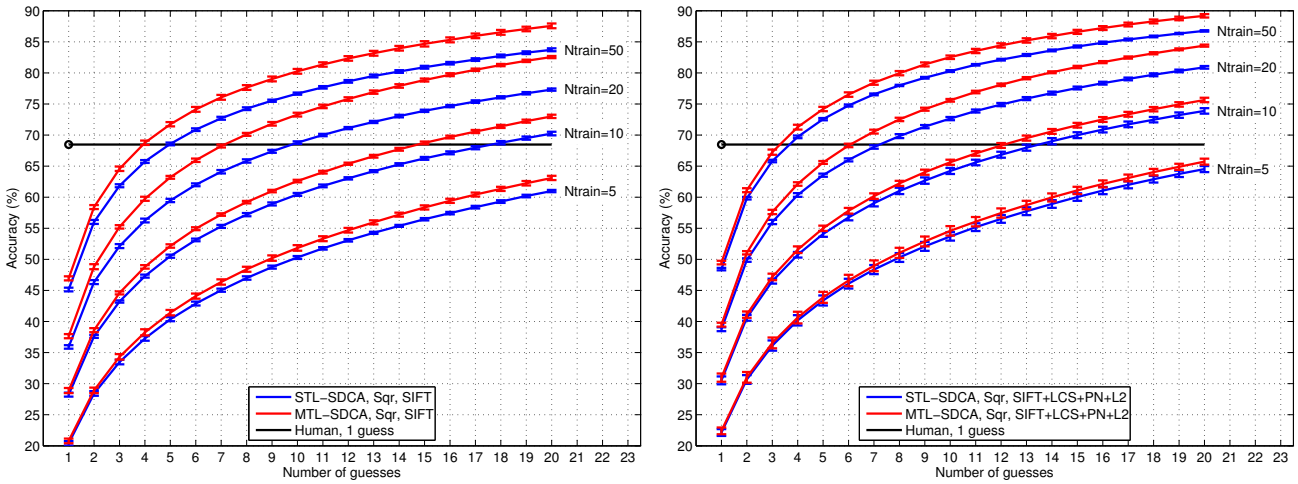
The inner problems are standard independent one-vs-all SVMs trained in a lower dimensional subspace which is determined by the matrix U. The latter is learned jointly for all tasks which facilitates knowledge transfer and is of particular interest when the amount of training examples per task is limited and at least some of the tasks are related.

To solve the optimization problem (1), we alternate between optimizing each w_t given a fixed U and then optimizing U given fixed w_t's. Each of these two subproblems is convex and is solved via an appropriate adaptation of SDCA. Since the overall problem (1) is not convex, we use, as the initial matrix $U^{(0)}$, a matrix of stacked predictors $\tilde{w}_t$ trained on the original features x_i in the standard single task learning regime. The latter also serves as a baseline (STL-SDCA) for the proposed method (MTL-SDCA).

3. SUN397 Experiments

This section reports our main experimental results on SUN397 [11] which is a challenging scene classification benchmark containing over 100K images of 397 categories. All source code including the scripts for running experiments as well as our implementation of the STL-SDCA and MTL-SDCA solvers can be found at our website.

We follow the protocol of Xiao *et al.* [11] and use 5, 10, 20, and 50 images per class for training and 50 images per class for testing. Our feature extraction pipeline follows the one described in [8] and uses SIFT [6] as well as Local Color Statistic (LCS) [2] as the low-level image descriptors.



(a) SIFT (b) SIFT + Color

Figure 1: Mean top-K accuracy (%) and standard deviation across 10 splits on the SUN397 dataset. The number of guesses K is varied between 1 and 20. Human performance is based on predictions of AMT workers provided by [11].

The features are further fine tuned (different PCA processing, ℓ_2- and power-normalization, see [5]) which allows our STL-SDCA baseline to outperform the results of [8].

Method		Ntrain=10	Ntrain=20	Ntrain=50
Xiao *et al*. [11]		20.9	28.1	38.0
Su and Jurie [10]				35.6 (0.4)
Donahue *et al*. [3]				40.9 (0.3)
Sánchez *et al*. [8]	SIFT	26.6 (0.4)	34.2 (0.3)	43.3 (0.2)
STL-SDCA	SIFT	28.2 (0.3)	35.9 (0.3)	45.1 (0.3)
MTL-SDCA	SIFT	**28.9 (0.4)**	**37.6 (0.3)**	**46.9 (0.3)**
Sánchez *et al*. [8]	+ Color	29.1 (0.3)	37.4 (0.3)	47.2 (0.2)
STL-SDCA	+ Color	30.5 (0.6)	38.8 (0.3)	48.4 (0.2)
MTL-SDCA	+ Color	**31.0 (0.7)**	**39.5 (0.3)**	**49.5 (0.3)**

Table 1: Mean accuracy (%) and standard deviation across 10 splits on the SUN397 dataset. See [5] for details.

Table 1 indicates superiority of a learned representation that is shared across multiple classes. MTL-SDCA is consistently better for every training subset and both choices of image descriptors. Furthermore, the Fisher Vector has striking performance even without color cues (SIFT only).

Because there are intrinsically ambiguous classes we extend our evaluation by reporting mean top-K accuracy in Figure 1. Again we observe that MTL-SDCA consistently improves classification performance not only for every image descriptor and every training subset, but also for every number of allowed guesses K. Moreover, the improvement is more significant at $K \geq 3$, *e.g.* using SIFT only and Ntrain=20 examples per class, MTL-SDCA improves top-5 accuracy by 3.7% and top-15 by 5%.

4. Conclusion

We proposed a novel multitask representation learning scheme that scales to high-dimensional image descriptors. The principle idea is to jointly learn a low dimensional representation that is shared across all classes and thus allows to leverage existing inter-class correlations. The proposed multitask learning method is not tied to a particular choice of features and can be applied with other image descriptors than the ones used in this work.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008. 1

[2] S. Clinchant, G. Csurka, F. Perronnin, and J.-M. Renders. XRCE's participation to ImagEval. In *ImageEval workshop at CVIR*, 2007. 1

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *arXiv:1310.1531*, 2013. 2

[4] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. 1

[5] M. Lapin, B. Schiele, and M. Hein. Scalable multitask representation learning for scene classification. In *CVPR*, 2014. 1, 2

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[7] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML*, 2013. 1

[8] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: theory and practice. *IJCV*, pages 1–24, 2013. 1, 2

[9] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *JMLR*, 14:567–599, 2013. 1

[10] Y. Su and F. Jurie. Improving image classification using semantic attributes. *IJCV*, 100(1):59–77, 2012. 2

[11] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 2