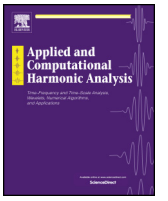
Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha



Case Studies

Scale-invariant learning and convolutional networks

Soumith Chintala, Marc'Aurelio Ranzato, Arthur Szlam, Yuandong Tian, Mark Tygert *, Wojciech Zaremba

Facebook, 1 Facebook Way, Menlo Park, CA 94025, United States

ARTICLE INFO

Article history:
Received 30 April 2016
Accepted 19 June 2016
Available online xxxx
Communicated by Charles K. Chui

Keywords:
Spectrum
Wavelets
Wavelet packets
Signal processing
Image processing
Classification
Equivariant
Invariant
Machine learning
Representation

ABSTRACT

Multinomial logistic regression and other classification schemes used in conjunction with convolutional networks (convnets) were designed largely before the rise of the now standard coupling with convnets, stochastic gradient descent, and backpropagation. In the specific application to supervised learning for convnets, a simple scale-invariant classification stage is more robust than multinomial logistic regression, appears to result in somewhat lower errors on several standard test sets, has similar computational costs, and features precise control over the actual rate of learning. "Scale-invariant" means that multiplying the input values by any nonzero real number leaves the output unchanged.

1. Introduction

Classification of a vector of real numbers (called "feature activations") into one of several discrete categories is well established and well studied, with solutions such as the ubiquitous multinomial logistic regression reviewed, for example, by [2]. However, conventional classifiers may not couple best with generation of the feature activations via convolutional networks (convnets) trained using stochastic gradient descent, as discussed, for example, by [7]. As discussed by [13], complex-valued convnets are essentially equivalent to tools familiar in harmonic analysis — data-driven multiscale windowed spectra, data-driven multiwavelet absolute values, or (in their most general configuration) data-driven nonlinear multiwavelet packets. "Data-driven" refers to fitting (also known as learning or training) the combination of the convnet and the classification stage by minimizing the cost/loss/objective function associated with the classification.

* Corresponding author.
E-mail addresses: soumith@fb.com (S. Chintala), ranzato@fb.com (M. Ranzato), aszlam@fb.com (A. Szlam), yuandong@fb.com (Y. Tian), tygert@fb.com (M. Tygert), wojciech@fb.com (W. Zaremba).

http://dx.doi.org/10.1016/j.acha.2016.06.005
1063-5203/

Classical classification stages neglect that many convnets are "equivariant" to scalar multiplication — multiplying the input values by any real number multiplies the output by the same factor; the present paper leverages this equivariance via a "scale-invariant" classification stage — a stage for which multiplying the input values by any nonzero real number leaves the output unchanged. The scale-invariant classification stage turns out to be more robust to outliers (including obviously mislabeled data), fits/learns/trains precisely at the rate that the user specifies, and apparently results in slightly lower errors on several standard test sets when used in conjunction with some typical convnets for generating the feature activations. The computational costs are comparable to those of multinomial logistic regression. Similar classification has been introduced earlier in other contexts by [3,6,10,12,14,15] and others. Complementary normalization includes the work of [1,4] and the associated references. The key to effective learning is rescaling, as described in Section 3 below (see especially the last paragraph there).

The remainder of the present paper has the following structure: Section 2 sets the notation. Section 3 introduces the scale-invariant classification stage. Section 4 analyzes its robustness. Section 5 illustrates the performance of the classification on several standard data sets. Section 6 draws several conclusions. The two appendices, Appendix A and Appendix B, provide more detailed derivations.

2. Notational conventions

All numbers used in the classification stage will be real valued (though the numbers used for generating the inputs to the stage may in general be complex valued). We follow the recommendations of [8]: all vectors are column vectors (aside from gradients of a scalar with respect to a column vector, which are row vectors), and we use $\|v\|$ to denote the Euclidean norm of a vector v; that is, $\|v\|$ is the square root of the sum of the squares of the entries of v. We use $\|A\|$ to denote the spectral norm of a matrix A; that is, $\|A\|$ is the greatest singular value of A, which is also the maximum of $\|Av\|$ over every vector v such that $\|v\| = 1$. The terminology "Frobenius norm" of A refers to the square root of the sum of the squares of the entries of A. The spectral norm of a vector viewed as a matrix having only one column or one row is the same as the Euclidean norm of the vector; the Euclidean norm of a matrix viewed as a vector is the same as the Frobenius norm of the matrix.

3. A scale-invariant classification stage

We study a linear classification stage that assigns one of k classes to each real-valued vector x of feature activations (together with a measure of confidence in its classification), with the assignment being independent of the Euclidean norm of x; the Euclidean norm of x is its "scale." We associate to the k classes target vectors $t_1, t_2, \ldots, t_k$ that are the vertices of either a standard simplex or a regular simplex embedded in a Euclidean space of dimension $m \geq k$ — the dimension of the embedding space being strictly greater than the minimum $(k-1)$ required to contain the simplex will give extra space to help facilitate learning; [3,6,10,12,14,15] (amongst others) discuss these simplices and their applications to classification. For the standard simplex, the targets are just the standard basis vectors, each of which consists of zeros for all but one entry. For both the regular and standard simplices,

$$\|t_1\| = \|t_2\| = \cdots = \|t_k\| = 1. \tag{1}$$

Given an input vector x of feature activations, we identify the target vector t_j that is nearest in the Euclidean distance to

$$z = \frac{y}{\|y\|}, \tag{2}$$

Please cite this article in press as: S. Chintala et al., Scale-invariant learning and convolutional networks, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.005