Large-scale Image Classification Using Supervised Spatial Encoder

Dmitriy Bespalov
Drexel University
db59@drexel.edu

Yanjun Qi
NEC Labs America
yanjun@nec-labs.com

Bing Bai
NEC Labs America
bbai@nec-labs.com

Ali Shokoufandeh
Drexel University
as79@drexel.edu

Abstract

Spatial pyramid matching (SPM) component is part of most state-of-art image classification methods. SPM encodes spatial distribution of image features, in an unsupervised fashion, by partitioning an image into regions at multiple scales and concatenating feature vectors for these regions. In this paper we propose to replace the unsupervised SPM procedure with a supervised two-stage feature selection that requires the image partitioned at a single scale. Experimental results show the proposed method performs statistically significantly better than the SPM baseline.

1. Introduction

In this paper we consider large-scale image classification (LIC) task. Extensive research efforts in recent years produced significant leap in classification accuracy of LIC methods. State-of-art LIC systems contain five stages or steps. 1) Extracting low-level image descriptors (LID) over a dense lattice from image $\mathbf{x}$. HoG, SURF or SIFT are popular choices of LID. We refer to Step 1 in Figure 1 as "LID extraction". 2) A "coding" step encodes each LID via a nonlinear feature mapping into a vector space $\mathbb{R}^{|\mathcal{D}|}$. This space is induced by a codebook $\mathcal{D}$ of visual-word features or codewords (see Step 2 in Figure 1). 3) A "pooling" step aggregates coding results for a sub-image into a single feature vector (FV) representing the image region (see Step 3 in Figure 1). We refer to LID extraction, coding and pooling (Steps 1–3 in Figure 1) as "FV construction" that results in "bag-of-features" (BoF) representation for an image region. 4) Partitioning $\mathbf{x}$ into regions at multiple scales (e.g., 1×1, 2×2, 4×4) and concatenating FVs for all regions to form final vector space representation of $\mathbf{x}$. This procedure known as spatial pyramid matching (SPM) was proposed by Lazebnik *et al.* [4]. Step 3 in Figure 1 illustrates a single SPM scale with 3×3 partitions. 5) A classification step that predicts image labels from vector space representation of $\mathbf{x}$ (see Step 6 in Figure 1).

The first four stages in current LIC systems are all unsupervised, in the sense that LID extraction, coding, pooling and SPM steps do not consider image label information. Image labels are only used in the final (fifth) step of the LIC system to train a classifier. Recent improvements of LIC systems are mostly attributed to refinements in the coding step: e.g., sparse-coding [9], locality-constrained linear coding [7], super-vector coding [5] and Fisher kernel-based method [6].

In this paper we propose to replace unsupervised SPM procedure with a two-stage embedding method that involves a supervised feature selection step called "Supervised Spatial Encoder" (SSE). SSE encodes spatial interactions among the regions in a latent space (see Step 4 in Figure 1). Latent embedding of individual regions are then combined to form image-level representation of $\mathbf{x}$ (see Step 5 in Figure 1). The parameters of the two-stage latent embedding are biased towards target LIC task. The experimental results show the proposed method performs statistically significantly better than the unsupervised SPM baseline.

2. Method

We formally define the SSE procedure in Section 2.1. Section 2.2 describes the two classifiers that we consider in this work. Before we proceed, a quick overview of notations is in order. Let $\mathcal{Y} = \{1, ..., C\}$ denote a set of class labels and $\mathcal{X}$ denote a collection of labeled images (i.e., training set), where $\mathcal{X} = \{(\mathbf{x}_i, y_i)_{i=1,...,L} | \mathbf{x}_i \in \mathcal{X} \ \& \ y_i \in \mathcal{Y}\}$ and $|\mathcal{X}| = L$. We denote cardinality of a set with $|.|$. Operator $\times$ denotes vector or matrix multiplication, while $\cdot$ is used to emphasize the multiplication of scalar variables.

2.1 Supervised Spatial Encoder

Given an image $\mathbf{x}$, the proposed procedure first partitions $\mathbf{x}$ into $N \times N$ regions, and BoF representation

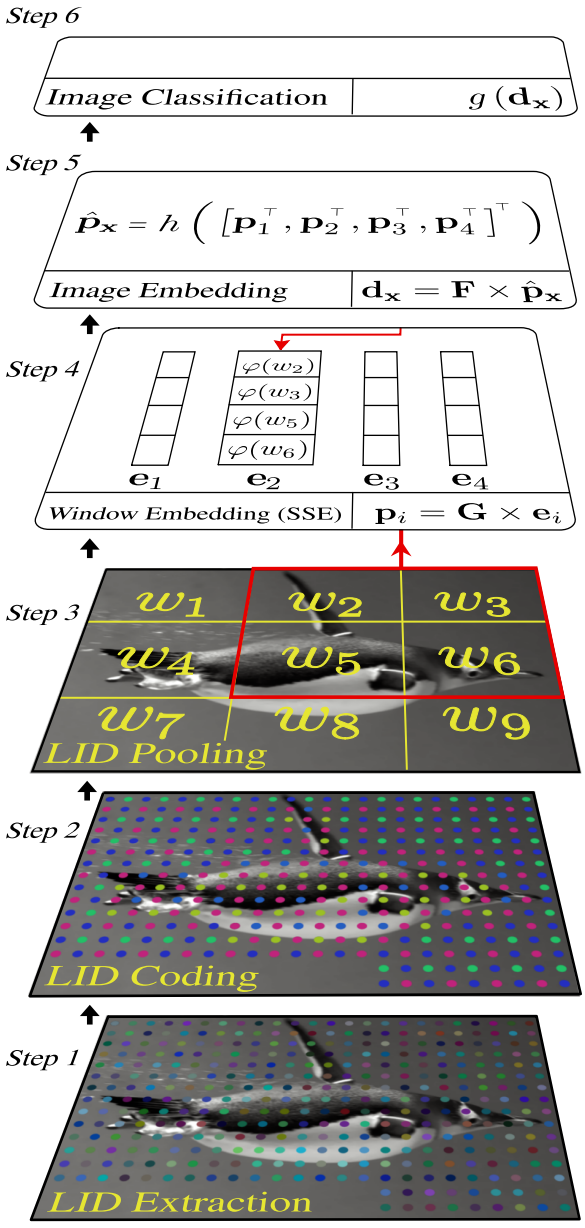

Figure 1. Image classification using SSE.
BEST VIEWED IN COLOR.

is computed for each region (Steps 1–3 in Figure 1). In Step 4 SSE encodes all sliding windows of $n \times n$ regions in $\hat{M}$-dimensional latent space. Figure 1 provides an overview of the proposed method with $N = 3$ and $n = 2$. For an i-th sliding window consisting of $k = n^2$ regions $\{w_{j_1}, w_{j_2}, \dots w_{j_k}\}$, vector $\mathbf{e}_i$ denotes concatenation of its FVs as:

$$\mathbf{e}_i = [\varphi(w_{j_1})^\top, \varphi(w_{j_2})^\top, \dots \varphi(w_{j_k})^\top]^\top, \quad (1)$$

where $\varphi(w) \in \mathbb{R}^{|\mathcal{D}|}$ is the BoF representation of region w. The $\hat{M}$-dimensional latent embedding $\mathbf{p}_i$ of the i-th window is obtained from

$$\mathbf{p}_i = \mathbf{G} \times \mathbf{e}_i, \quad (2)$$

where $\mathbf{G} \in \mathbb{R}^{\hat{M} \times n^2 \cdot |\mathcal{D}|}$. For brevity we drop bias terms in the definitions of latent projections. We note that projection (2) maintains n^2 independent embedding parameters for each region w_j, based on its position within the i-th sliding window.

Step 5 concatenates latent embedding of all sliding windows into a single vector $\hat{\mathbf{p}}_\mathbf{x}$ defined as:

$$\hat{\mathbf{p}}_\mathbf{x} = h\left(\left[\mathbf{p}_1^\top, \mathbf{p}_2^\top, \dots \mathbf{p}_{N^2}^\top\right]^\top\right), \quad (3)$$

where $h(\cdot) = \tanh(\cdot)$[1]. Then the final image-level latent representation $\mathbf{d}_\mathbf{x} \in \mathbb{R}^M$ is the result of the second projection step:

$$\phi(\mathbf{x}) \equiv \mathbf{d}_\mathbf{x} = \mathbf{F} \times \hat{\mathbf{p}}_\mathbf{x}, \quad (4)$$

where $\mathbf{F} \in \mathbb{R}^{M \times N^2 \cdot \hat{M}}$.

In summary, SSE projects all sliding windows each containing n^2 regions into a latent space using projection (2). The latent representation of all sliding windows are then concatenated to form vector $\mathbf{p}_\mathbf{x}$ in (3). The second projection layer is used to obtain image-level latent representation $\mathbf{d}_\mathbf{x}$ in (4). In other words, the proposed method resembles construction of SPM in a latent space. In contrast to SPM procedure that constructs image-level representation in an unsupervised bottom-up fashion, Step 4 in the proposed method performs a supervised feature selection at an intermediate level of spatial pyramid, while Step 5 computes image-level representation of $\mathbf{x}$ in the supervised latent space. It is this latter encoding, $\mathbf{d}_\mathbf{x}$, of image $\mathbf{x}$ that will be used as the input of the classifier (see Section 2.2).

Relationship to CNN The proposed procedure is closely related to convolutional neural networks (CNN). However, they do have two major differences. Firstly, CNN models spatial interactions among dense FVs (pixels), while the proposed model is designed to handle sparse FVs. Secondly, CNNs have multiple layers of projections, thus are more expensive to train. To the best of our knowledge, image classification methods with CNNs are generally restricted to small or medium-scale datasets e.g., Caltech-101/256, PASCAL07. In contrast, our image classification system requires a single SSE layer to encode spatial interactions of sparse FVs computed for a partitioned image.

[1]The non-linear element-wise operator $\tanh(\cdot)$ that converts the unbounded range of the input into $[-1, 1]$.

2.2 Classifiers

We examine the proposed latent image encoding in conjunction with two different vector space classifiers. The empirical evidence suggests that irrespective of classifier, SSE-based method outperforms the baseline that relies on the SPM procedure to capture spatial distribution of image features. We implement the proposed system as a multi-layer feed-forward perceptron model [3], which is trained using stochastic gradient descent [2].

The first classifier we consider is the so-called Multinomial Logistic Regression (MLR). In the case of MLR classifier, we learn $M = C$ dimensional latent representation of image $\mathbf{x}$. In other words, $\mathbf{d}_{\mathbf{x}} \in \mathbb{R}^C$ maintains a coefficient weight for every candidate class (i.e., $M = C$). And the predicted class label for $\mathbf{x}$ is calculated as follows:

$$g_1(\mathbf{d}_{\mathbf{x}}) = \underset{i \in \{1..C\}}{\arg\max} \frac{\exp(\mathbf{d}_{\mathbf{x}}[i])}{1 + \sum_{k \in \{1..C\}} \exp(\mathbf{d}_{\mathbf{x}}[k])}, \quad (5)$$

where $\mathbf{d}_{\mathbf{x}}[k]$ denotes the k^{th} element in vector $\mathbf{d}_{\mathbf{x}}$. Given a training set $\mathcal{X}$, MLR classifier is computed by minimizing the loss function:

$$\mathcal{L}_1(\mathcal{X}) = -\sum_{i \in \{1..|\mathcal{X}|\}} \log \frac{\exp(\mathbf{d}_{\mathbf{x}_i}[y_i])}{1 + \sum_{j \in \{1..C\}} \exp(\mathbf{d}_{\mathbf{x}_i}[j])},$$

This latter loss is called "negative log likelihood" in literature.

The second classifier we consider in this work is based on the WARP Loss proposed by Weston *et al.* [8]. This classifier is based on margin-penalized ranking of pairwise (label-image) similarity values. In addition to learning latent representation $\phi(\cdot)$ for images, the WARP classifier computes M-dimensional latent representation for each label $y_i \in \mathcal{Y}$. Let $\mathbf{V} \in \mathbb{R}^{M \times C}$ denote the parameters for latent embedding for C labels, and $\mathbf{V}_{y_i}$ denote the latent embedding of y_i. The classifier is then computed by minimizing the WARP loss:

$$\mathcal{L}_2(\mathcal{X}) = \sum_{i \in \{1..|\mathcal{X}|\}} \Lambda\left(\sum_{y_j \neq y_i} I\left[\xi(\mathbf{x}_i, y_j) > \xi(\mathbf{x}_i, y_i)\right]\right),$$

where $\Lambda(k) = \sum_{j=1}^{k} \frac{1}{j}$, $\xi(\mathbf{x}, y) = \frac{\mathbf{V}_y^\top \times \mathbf{d}_{\mathbf{x}}}{\|\mathbf{V}_y\| \cdot \|\mathbf{d}_{\mathbf{x}}\|}$ and $I(\cdot)$ is the indicator function. The prediction for image $\mathbf{x}$ using the classifier optimized with the WARP loss is obtained using:

$$g_2(\mathbf{d}_{\mathbf{x}}) = \underset{i \in \{1..C\}}{\arg\max}\, \xi(\mathbf{x}, y_i). \quad (6)$$

3. Experiments

Metric We evaluate the proposed method using dataset from the Image-Net Large Scale Visual Recognition Challenge 2011 (ILSVRC2011) [1]. The dataset contains images of 1000 categories of objects, where each category corresponds to a synset (set of synonymous nouns) in WordNet. The categories are organized as leaf nodes into a hierarchy that corresponds to a subset of WordNet synset hierarchy. The competing methods in ILSVRC2011 were evaluated using two cost measures. The first, called *flat cost* (FL), measured classification hit rate among the top five label predictions produced by each method. The *hierarchical cost* (HI) used the minimum height of the lowest common ancestor of ground truth label and one of five predicted labels. To conserve space, we refer the reader to the ILSVRC2011 webpage [1] for a formal definition of the cost measures used in the competition. We report classification results obtained using one (**Top-1**) and five (**Top-5**) label predictions per image.

BoF Representation We implement Steps 1–3 in Figure 1 using VLFeat[2] computer vision library. For LID extraction, images are rescaled with the largest dimension set to 300 pixels. We use grayscale SIFT image descriptors extracted every 8 pixels using square regions at scales 8, 16, and 24 pixels. We cluster 30 million randomly chosen SIFT descriptors into 4,096 leaf clusters (i.e., codewords) using hierarchical k-means. For LID coding we closely follow the implementation of locality-constrained linear coding (LLC) method [7], and max-pooling is used during FV construction.

Baseline The proposed **SSE** method is a supervised alternative to the **SPM** procedure. In our experiments, the SPM baseline is computed using three scales: 1×1, 2×2 and 4×4. The latent embedding of SPM representation ($\mathbf{d}_{\mathbf{x}} \in \mathbb{R}^M$) is modeled using a single linear projection $\phi : \mathbb{R}^{21 \cdot |\mathcal{D}|} \rightarrow \mathbb{R}^M$.

Parameters and Training In the case of SSE method, every image is partitioned into 4×4 regions and 3×3 sliding windows are encoded in $\hat{M} = 100$ dimensional latent space. As was mentioned earlier, for MLR classifier we set $M = C = 1000$. For WARP classifier we use $M = 300$ for both SSE and SPM methods. We split the 1.2 million training images into 30 slices where images for each synset are divided evenly among the slices. The training proceeds in distributed fashion with

[2] http://www.vlfeat.org/

Table 1. Classification Error. FL and HI denote flat and hierarchical cost, respectively. The FL numbers marked with † are statistically significantly better than SPM baseline with $p < 10^{-4}$. Classification results are obtained using one (Top-1) and five (Top-5) label predictions per image.

		MLR		WARP	
Cost	# Lbls	**SPM**	**SSE**	**SPM**	**SSE**
FL	Top-5	60.7	56.5†	60.9	59.6†
FL	Top-1	77.5	75.5†	80.3	79.4†
HI	Top-5	29.0	26.6	28.8	28.3
HI	Top-1	46.2	43.7	46.3	45.3

ten work nodes, each updating the model parameters using 35,000 samples from a randomly selected slice. The parameters of models computed by work nodes are then averaged together and saved as a latest trained model. These models are regularly evaluated using validation images and the best performing model is retained for each method. After the classification performance stops improving, we use the best performing model to obtain label predictions for test images and submit the predictions to the ILSVRC2011 evaluation server [3]. For both SSE and SPM methods, the training took less than a week to complete. The FL and HI classification costs can be found in Table 3.

4 Discussion and Future Work

The SSE method results in statistically significant improvement over SPM with $p < 10^{-4}$ for both classifiers we considered in this work. It worth noting that our method relies on construction of FVs at a single scale and performs feature selection via two-stage latent projection. The baseline SPM method on the other hand, requires extraction of FVs at three scales. We also note that the classification performance of both SPM and SSE methods correspond to the tail of ranked list of participants in ILSVRC2011. Moderately chosen parameters of BoF pipeline allowed us to train the models in a timely fashion but significantly degraded classification performance. For example, a variant of LLC [5] resulted in the state-of-art performance on ILSVRC2010. The method scaled images to at most 500 pixels, used 20,480 codewords, two types of image descriptors (HoG and LBP), and 20 closest codewords to describe each LID (we used 5). In our future work we plan to improve the LIC performance of our method by using additional low-level image features: color histograms and local binary pattern [5]. In addition, we believe that using SSE to encode image partitions computed at different scales should improve the LIC performance even further.

[3] http://www.image-net.org/challenges/LSVRC/2011/test_server

References

[1] A. Berg, J. Deng, S. Satheesh, H. Su, and F.-F. Li. Image-net large scale visual recognition challenge 2011 (ilsvrc). www.image-net.org/challenges/LSVRC/2011.

[2] L. Bottou. Stochastic learning. In O. Bousquet and U. von Luxburg, editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence, LNAI 3176, pages 146–168. Springer Verlag, Berlin, 2004.

[3] R. Collobert and S. Bengio. Links between perceptrons, mlps and svms. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 23–, New York, NY, USA, 2004. ACM.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006*, volume 2, pages 2169–2178. IEEE, 2006.

[5] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and svm training. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1689 –1696, june 2011.

[6] J. Sanchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1665 –1672, june 2011.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360 –3367, june 2010.

[8] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, pages 2764–2770, 2011.

[9] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794 – 1801, june 2009.