Large-scale Image Classification Using Supervised Spatial Encoder

Dmitriy Bespalov†, Yanjun Qi‡, Bing Bai‡, Ali Shokoufandeh†
†Department of Computer Science, Drexel University, Philadelphia, PA, USA
‡Machine Learning Group, NEC Labs America, Princeton, NJ, USA

Motivation

- Bag-of-features (BoF) model is a popular quantization procedure for low-level image descriptors (e.g., SIFT, HoG, SURF, RBG histogram, local binary pattern, etc). However, the BoF model ignores useful information in locations of the image descriptors.

- Spatial Pyramid Matching (SPM) [1] is an extension to the BoF representation that encodes spatial distribution of the quantized image descriptors. However, SPM is a fixed scheme and can not adapt to different dataset when needed.

- The proposed Supervised Spatial Encoder (SSE) is a supervised alternative to the SPM model. SSE encodes spatial distribution of the quantized image descriptors in a low-dimensional latent space from a single image partition.
- The SSE model was originally proposed to encode test phrases, which amounted to capturing the spatial distribution of single words in documents [2]. A text document can be seen as well-formed sequences of sparse vectors, like the quantized SIFT in images.
- Similar ideas of using location information can be found in convolutional neural networks (CNN). However, CNN builds hierarchical representation from features in a dense vector form. SSE works on a well-formed sequences of sparse vectors, like the quantized SIFT in images.

Baseline model: BoF representation with SPM

1. Given an image \( x \), the low-level image descriptors (LID) are extracted from a dense lattice over \( x \).
2. Each LID is coded with a sparse vector in \( \mathbb{R}^{D} \), which is induced by the codebook \( D \).
3-5. SPM partitions the image \( x \) into \( 4 \times 4 \), \( 2 \times 2 \), and \( 1 \times 1 \) segments. The coded LIDs are pooled to form a BoF vector in \( \mathbb{R}^{D} \) that describes each segment.
6. BoF vectors for all segments are concatenated to obtain a final representation for \( x \) in \( \mathbb{R}^{D \times (D+1)} \).
7. Image classification is carried out with a multinomial logistic regression (MLR) [3], which is a popular neural network classifier that estimates probabilities for \( C \) classes using the map \( g: \mathbb{R}^{D} \rightarrow \mathbb{R}^{C} \). MLR classifier is implemented with a linear projection layer and the LogSoftMax layer.

Proposed model: supervised spatial encoder

1. An image \( x \) is partitioned into \( 3 \times 3 \) segments. \( e_{w} \in \mathbb{R}^{D} \) denotes a BoF vector for the segment \( w \in x \).
2. The SSE layer \( \varphi_{w}: \mathbb{R}^{E} \rightarrow \mathbb{R}^{D} \) projects \( w \), a sliding window of \( 2 \times 2 \) segments, into an \( M \)-dimensional space.
3. Linear projection \( \varphi_{M}: \mathbb{R}^{D} \rightarrow \mathbb{R}^{M} \) computes image-level embedding of \( x \).
4. Image classification is carried out with an MLR classifier. In our implementation of the SSE system we set \( M = C \), so the MLR classifier \( g(d_{w}) \) applies only the LogSoftMax layer to \( d_{w} \).

Experiments: large-scale image classification

- ILSVRC2011 dataset [4] was used to benchmark SPM and SSE.
- The dataset contains 1.2 million training images which are organized into \( C = 1000 \) classes. Each class is a synset in WordNet. ILSVRC2011 contains 50,000 validation and 100,000 testing images.
- Locality-constrained linear coding (LLC) [5] was used to quantize SIFT descriptors into \( |D| \)-dimensional sparse BoF vectors, where \( |D| = 4 \). SSE used a \( 4 \times 4 \) image partition to embed sliding windows of \( 3 \times 3 \) segments into \( \mathbb{R}^{D} \), where \( M = 100 \).
- Each image was considered as correctly classified if its ground-truth label was among the top \( n \) predicted labels. Top-5 and Top-1 denote the cases when \( n = 5 \) and \( n = 1 \).

Conclusions

- The experimental results suggest that SSE achieves higher image classification accuracy than the unsupervised SPM. In addition, SSE requires a single image partition and avoids saving BoF vectors at multiple scales.
- Using multiple types of LIDs in BoF is known to improve image classification. The use of SIFT, local binary pattern, and normalized RBG histogram in our BoF vectors lowered the Top-5 and Top-1 classification errors for SSE to 39.2% and 61.5%, respectively. The best performing method with LLC for ILSVRC2011 had the Top-5 classification error of 35.9%.
- Lower performance of our system can be attributed to the moderately chosen parameters of our LLC-based BoF model. For instance, the winner of ILSVRC2010 used LLC with \( |D| = 20 \), 490 codewords, scaled images to 500 pixels (we scaled to 300) and used 20 closest codewords to describe each LID (we used 5).

Bibliography