Large-scale image classification

Features

- SIFT
  - Color descriptors
- Bag of Visual Words (BOV)
- Spatial Pyramid Matching (SPM)
- Improvements of BOV
  - Soft-assign the local descriptors
  - Sparse coding
    - Excellent results
    - Costly iterative optimization
  - Locality-constrained Linear Coding (LLC)
  - Nonlinear kernel
  - Intersection Kernel (IK)
- BOW with kernel kinds
- Vector of Locally Aggregated Descriptors (VLAD)
- Fisher Vector (FV)
- Learned features from deep architectures

Existing Classification Methods

Approaches

Hierarchical strategies

- Pattern
  - Tree construction
    - Top-down
    - Bottom-up
  - Node training
- Hierarchical SVM (HSVM)
  - Top-down partition be k-means
    - Train with SVM
  - Bottom-up construction using separability measure
    - Train with SVM

- Label Embedding Tree
  - OVR vs confusion matrix vs graph cut vs Tree structure

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}} \sum_{i=1}^{K} \sum_{x \in \mathcal{P}_i} \ell(x, \mathbf{w}_i)$$

  - Joint Tree loss optimization

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}, \{\mathbf{c}_i\}_{i=1}^{K}} \sum_{i=1}^{K} \sum_{x \in \mathcal{P}_i} \ell(x, \mathbf{w}_i) + \lambda \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 + \gamma \sum_{i=1}^{K} \sum_{j \in \mathcal{C}_i} \|\mathbf{c}_i - \mathbf{c}_j\|^2$$

$$\text{s.t.} \quad \mathbf{c}_i^T \mathbf{c}_j = \delta_{ij}$$

  - Speed enhancement: $O(dK) \Rightarrow O(d \log(K))$ or $O(d)$
  - Accuracy enhancement?

Table 1. Flat-versus Tree-supporting Models. We list accuracies for various flat-and tree-supporting models on various datasets. Times are in seconds for a single core on a 2.2GHz Opteron.

Dataset	Model	Acc	Acc	Acc	Time	Time	Size
ILSVRC 2010-12	LLC	0.526	0.710	0.862	1.12E+6	2.76E+6	1.26E+12
	Tree-LLC	0.517	0.699	0.854	2.25E+4	1.76E+4	2.52E+9
	Tree-LLC-Prune	0.517	0.699	0.854	2.25E+4	1.76E+4	1.26E+9
	Tree-LLC-Prune-Sparse	0.517	0.699	0.854	1.76E+4	1.76E+4	1.26E+9
ImageNet 10K	LLC	0.264	0.498	0.708	1.12E+6	2.76E+6	1.26E+12
	Tree-LLC	0.254	0.483	0.691	2.25E+4	1.76E+4	2.52E+9
	Tree-LLC-Prune	0.254	0.483	0.691	2.25E+4	1.76E+4	1.26E+9
	Tree-LLC-Prune-Sparse	0.254	0.483	0.691	1.76E+4	1.76E+4	1.26E+9

Table 2. Label Embedding and LabelRanking Tree Results.

Dataset	Model	Acc	Acc	Acc	Time	Time	Size
ILSVRC 2010-12	LLC	0.526	0.710	0.862	1.12E+6	2.76E+6	1.26E+12
	LE-Tree-Cost-Sensitive	0.519	0.703	0.857	2.25E+4	1.76E+4	2.52E+9
	LE-Tree-Cost-Sensitive	0.522	0.706	0.860	2.25E+4	1.76E+4	1.26E+9
	LE-Tree-Cost-Sensitive	0.524	0.708	0.861	1.76E+4	1.76E+4	1.26E+9
ImageNet 10K	LLC	0.264	0.498	0.708	1.12E+6	2.76E+6	1.26E+12
	LE-Tree-Cost-Sensitive	0.258	0.488	0.699	2.25E+4	1.76E+4	2.52E+9
	LE-Tree-Cost-Sensitive	0.261	0.493	0.704	2.25E+4	1.76E+4	1.26E+9
	LE-Tree-Cost-Sensitive	0.263	0.496	0.706	1.76E+4	1.76E+4	1.26E+9

- Efficient Label Tree
  - minimize expected time cost
    subject to accuracy constraints
  - or vice versa
- Faster and more balanced Hair Label Tree

Table 3. Efficient and accuracy-aware, using ILSVRC 2010-12 and ImageNet 10K, we plot the top-1 accuracy versus the average number of inner products performed per example. Two speedup is the ratio between the number of inner products the flat model and the tree model perform. Our goal is to achieve comparable or better accuracy with efficiency. With the shrinking set, computed costs for training over 20 Megapixel (ILSVRC) and 10 Gigapixel (ImageNet) examples.

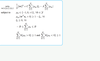
Dataset	Model	Acc	Acc	Acc	Speedup	Time	Size
ILSVRC 2010-12	LLC	0.526	0.710	0.862	1	2.76E+6	1.26E+12
	Tree-Cost-Sensitive	0.519	0.703	0.857	2.25E+4	1.76E+4	2.52E+9
	Tree-Cost-Sensitive	0.522	0.706	0.860	2.25E+4	1.76E+4	1.26E+9
	Tree-Cost-Sensitive	0.524	0.708	0.861	1.76E+4	1.76E+4	1.26E+9
ImageNet 10K	LLC	0.264	0.498	0.708	1	2.76E+6	1.26E+12
	Tree-Cost-Sensitive	0.258	0.488	0.699	2.25E+4	1.76E+4	2.52E+9
	Tree-Cost-Sensitive	0.261	0.493	0.704	2.25E+4	1.76E+4	1.26E+9
	Tree-Cost-Sensitive	0.263	0.496	0.706	1.76E+4	1.76E+4	1.26E+9

- Probabilistic Label Tree

$$p(y|\mathbf{x}) = \sum_{i=1}^{|\mathcal{P}_y|} p(y|\mathcal{P}_y^i) p(\mathcal{P}_y^i|\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^D, y \in \mathcal{Y}$$

- Relaxed Tree

- Relaxed Hierarchy
  - Original version:
    k-means and threshold η => 2l-partition => SVM training for parent/leaves partitions



  - Improved version:
    Mark each label with { 0, 1, +1 } => 1 Vs Rest SVM => binary partition

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}, \{\xi_i\}_{i=1}^{N}} \frac{1}{2} \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 + C \sum_{i=1}^{N} \xi_i$$

$$\text{subject to:} \quad \sum_{i=1}^{K} \mathbf{w}_i^T \mathbf{x}_n - \mathbf{w}_y^T \mathbf{x}_n \geq \Delta(y, i) - \xi_n, \quad \forall n, y, i$$

$$\xi_n \geq 0, \quad \forall n$$

$$\Delta(y, i) = \begin{cases} 0 & \text{if } y = i \\ 1 & \text{if } y \neq i \text{ and } y \neq 0 \\ \rho & \text{if } y \neq i \text{ and } y = 0 \end{cases}$$

Hierarchical Feature Learning

- Learning hierarchical feature extractors by sparse coding and MAX pooling
- Jointly optimizing all the layers

Optimization

Approximation

- Nearest Neighbor (NN) / Approximate Nearest Neighbor (ANN)
  - In accuracy: NN << CNN, ANN << CNN, for large dimensional data
  - Does not scale to large datasets
  - Speed and simplicity
- One-vs-All: One-vs-Rest / One-vs-One

$$\ell_{ova}(\mathbf{x}, y, \mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$$

- Multiclass SVM

$$\ell_{mcsvm}(\mathbf{x}, y, \mathbf{w}) = \sum_{y' \neq y} \max(0, \mathbf{w}_{y'}^T\mathbf{x} - \mathbf{w}_y^T\mathbf{x} + 1)$$

- Ranking SVM

$$\ell_{rank}(\mathbf{x}, y, \mathbf{w}) = \sum_{y' \neq y} \sum_{\tau=1}^{|\mathcal{Y}|-1} \mathbb{I}(\mathbf{w}_{y'}^T\mathbf{x} - \mathbf{w}_y^T\mathbf{x} + 1 > \tau) \cdot \frac{1}{\tau}$$

- Weighted approximate ranking SVM

$$\ell_{wsabie}(\mathbf{x}, y, \mathbf{w}) = \sum_{y' \neq y} L(\tau(y')) \max(0, \mathbf{w}_{y'}^T\mathbf{x} - \mathbf{w}_y^T\mathbf{x} + 1)$$

Optimization

- Batch methods
  - Sequential minimal optimization
    - LibSVM
    - SVM^{light}
  - Coordinate-descent with second-order acceleration
    - Liblinear
    - Shogun
- Stochastic optimization
  - Stochastic gradient descent (SGD)
    - Pegasos
    - LibSGD
    - ImageNet10K: 10000 categories, 9 million images FV-10K dimensions: 200 days
    - ImageNet10K: 10000 categories, 9 million images FV-10K dimensions: 200 days
    - 1 hour for 1 SVM => 416.7 days => 6.3 days on a 66-machine cluster
    - 2 days
    - 2 years

Experimental conclusions

- Good on small datasets => Good on large datasets

- Semantic hierarchy (WordNet) => Visual confusion between the categories

- Difference in difficulty between different datasets is independent of feature and classifier choice

- Denser datasets predict lower accuracy

Caltech categories are sparse => not representative of real-world problems

- Hierarchical cost => more informative

- Early stopping works

- Data rebalancing is important

$$\frac{1}{N} \sum_{n=1}^{N} \ell_{mcsvm}(\mathbf{x}, y, \mathbf{w}) + \frac{1}{K} \sum_{k=1}^{K} \ell_{mcsvm}(\mathbf{x}_k, y_k, \mathbf{w})$$

- Different objective functions => similar results
  OVR is robust and competitive

Pattern

- Tree construction
  - Top-down
  - Bottom-up
- Node training

Contribution

- New objective function
- Merge solutions to
  - Differentiable
  - Tree construction
  - Label partitioning
  - Feature selection
  - Feature diversity
- Forest framework

Soft-Label Forest

Motivation

- Background
  - High dimensional data
  - Large number of samples
  - Large number of classes
  - Intra-class diversity
  - Inter-class similarity
  - Feature diversity
- Flat models:
  - Loss minimization: Pattern labels
  - Robust & accurate: Slow
- Classic tree models:
  - Criteria evaluation: Pattern samples
  - Fast accurate: Flat
- Existing hierarchical models:
  - Loss minimization: Pattern labels
  - Bad Robust: Fast
- What if ... ?
  - Loss minimization: Pattern samples / Soft-partition labels
  - Robust & Accurate: Fast
- Decision Tree
  - ID3
  - C4.5
- CART

Solution

Soft-Decision Model

- To find an optimal decision boundary in the feature space to clearly separate the samples into two groups according to their labels

- Known positive and negative samples: classification

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}, \{\xi_i\}_{i=1}^{N}} \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 + C \sum_{n=1}^{N} \xi_n$$

$$\text{subject to:} \quad \forall i: \quad y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n$$

$$\xi_n \geq 0$$

- Unknown positive and negative samples: label partitioning

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}, \{\xi_i\}_{i=1}^{N}, \{y_n\}_{n=1}^{N}} \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 + C \sum_{n=1}^{N} \xi_n + \lambda \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{I}(y_n \neq i) \mathbb{I}(y_n \neq j) \|\mathbf{w}_i^T \mathbf{x}_n - \mathbf{w}_j^T \mathbf{x}_n\|^2$$

$$\text{subject to:} \quad \forall i: \quad y_n(\mathbf{w}_i^T \mathbf{x}_n + b_i) \geq 1 - \xi_n$$

$$\xi_n \geq 0$$

$$y_n \in \{1, \ldots, K\}$$

- As to feature diversity
  - $y_n = 0$: negative class
  - $y_n = 1$: positive class
  - $y_n = 0.5$: ambiguous class => little loss => divided into subclasses

Soft-Label Tree

- p_n = probability of reaching some node in the tree
- Minimize the expected loss for all nodes

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}} \sum_{n=1}^{N} \sum_{i \in \mathcal{I}} \mathbb{E}_{z \sim p_n(i)} \left[\ell(\mathbf{x}_n, y_n, \mathbf{w}_i) \right] + \sum_{i \in \mathcal{I}} \|\mathbf{w}_i\|^2$$

$$\text{where} \quad \mathbb{E}_{z \sim p_n(i)} \left[\ell(\mathbf{x}, y, \mathbf{w}_i) \right] = \mathbb{I}(y \in \mathcal{Y}_i^1) \ell(\mathbf{x}, y, \mathbf{w}_i) + \mathbb{I}(y \in \mathcal{Y}_i^2) \ell(\mathbf{x}, y, \mathbf{w}_i)$$

- Data rebalancing

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}} \sum_{n=1}^{N} p_n \ell(\mathbf{x}_n, \quad \mathbf{w}) + \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 + \lambda \sum_{i=1}^{K} \sum_{j=1}^{K} \left(\frac{|\mathcal{D}_i^1|}{|\mathcal{D}_i^2|} - \frac{|\mathcal{D}_j^1|}{|\mathcal{D}_j^2|} \right)^2$$

- Linear model with hinge loss

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}, \{\xi_i\}_{i=1}^{N}} \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 + C \sum_{n=1}^{N} \xi_n$$

$$\text{subject to:} \quad \forall i: \quad \begin{cases} \mathbf{w}_i^T \mathbf{x}_n \geq 1 - \xi_n, & \text{if } y_n \in \mathcal{Y}_i^1 \\ \mathbf{w}_i^T \mathbf{x}_n \leq -1 + \xi_n, & \text{if } y_n \in \mathcal{Y}_i^2 \\ \xi_n \geq 0 \end{cases}$$

Soft-Label Forest

- Bootstrapping
- Down-sampling
- Feature selection: L1 regularization
- Final objective

$$\min_{\{\mathbf{w}_i\}_{i=1}^{K}} \sum_{n=1}^{N} \sum_{i \in \mathcal{I}} \mathbb{E}_{z \sim p_n(i)} \left[\ell(\mathbf{x}_n, y_n, \mathbf{w}_i) \right] + \sum_{i \in \mathcal{I}} \|\mathbf{w}_i\|^2$$

$$\text{where} \quad \mathbb{E}_{z \sim p_n(i)} \left[\ell(\mathbf{x}, y, \mathbf{w}_i) \right] = \mathbb{I}(y \in \mathcal{Y}_i^1) \ell(\mathbf{x}, y, \mathbf{w}_i) + \mathbb{I}(y \in \mathcal{Y}_i^2) \ell(\mathbf{x}, y, \mathbf{w}_i)$$

$$\text{subject to:} \quad \forall i: \quad \begin{cases} \mathbf{w}_i^T \mathbf{x}_n \geq 1 - \xi_n, & \text{if } y_n \in \mathcal{Y}_i^1 \\ \mathbf{w}_i^T \mathbf{x}_n \leq -1 + \xi_n, & \text{if } y_n \in \mathcal{Y}_i^2 \\ \xi_n \geq 0 \end{cases}$$

Experimentation

- Caltech-256

Methods	Acc(%)	Cost	Depth
One-vs-All HSVM	34.10	350070.13	8
One-vs-All SVM	34.10	1530526.92	1
Time=1.5 - 6.0	34.10	918042.1	15.5
Time=6.0 - 12.0	34.10	1530526.92	1
Relaxed Hierarchy	32.44	74691.71	12.6

Table 1. Key performance on Caltech-256. (+) in the cost parameter you in units of multiplication operations. (F is the number of feature dimensions, N is the number of samples, an average of quarter samples for root and number of nodes for other nodes, and 2 for sparse inner product.)

Figure 2. Diagram of different approaches on Caltech-256. The accuracy is plotted against estimated complexity with respect to Time (s). All (Time=6.0) accuracy is 1 and its cost is 1530526.92 (Time=12.0). Relaxed Hierarchy, Soft-Label Forest maintain baseline's accuracy (0.341) at only a fraction of the cost. Soft-Label Forest's 0.341 accuracy outperforms all other recent models.

- ILSVRC2010

Methods	Acc(%)	Cost	Depth
One-vs-All	0.526	1261970.5	1
Time=1.0	0.519	1000703.8	21.6
Time=1.5 - 6.0	0.526	1261970.5	18.4
Time=6.0 - 12.0	0.526	1261970.5	1

Table 2. Performance on ILSVRC2012. (F is the number of feature dim and N is the number of samples, an average of quarter samples for root and number of nodes for other nodes, and 2 for sparse inner product.) Soft-Label Forest shows superior accuracy to Label Tree while remaining competitive against One-vs-All.

Figure 3. Diagram of models on ILSVRC2012. The accuracy is plotted against estimated complexity with respect to Time (s). All (Time=6.0) accuracy is 1 and its accuracy is 0.526 (Time=12.0). Soft-Label Forest's 0.519 can save up to 93.0% of the computational cost of One-vs-All (Time=1.0). Soft-Label Forest's 0.526 is 0.007 higher than LLC, but its accuracy is less acceptable.