IBM TRECVID'08 High-level Feature Detection

Rong Yan
IBM T. J. Watson Research Center
Hawthorne, NY 10532 USA

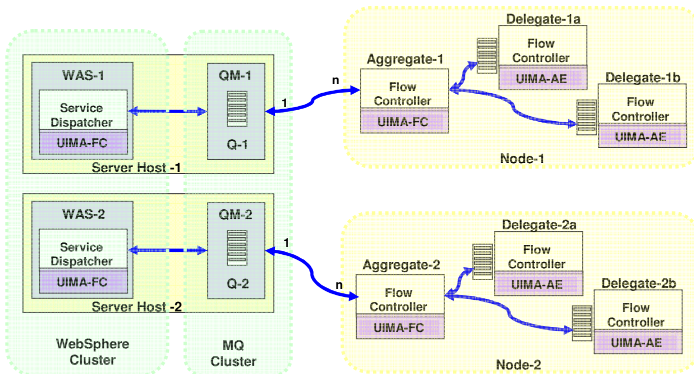
Team: Apostol Natsev, Wei Jiang, Michele Merler, John R. Smith, Jelena Tesic, Lexing Xie, Rong Yan

Main Research Highlights

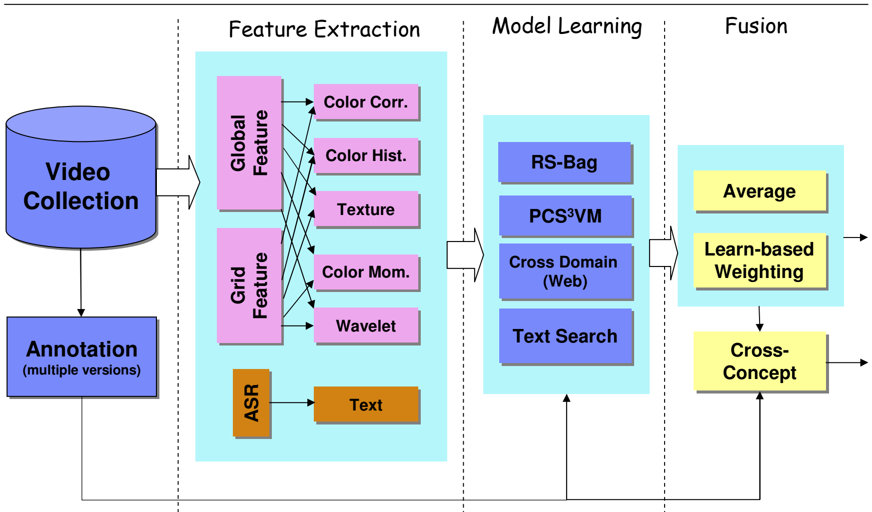
- Distributed end-to-end machine learning system
- Efficient model learning w. random subspace bagging
- Semi-supervised joint feature and concept modeling
- Cross-domain learning from web domain

Distributed End-to-End Concept Modeling Tool

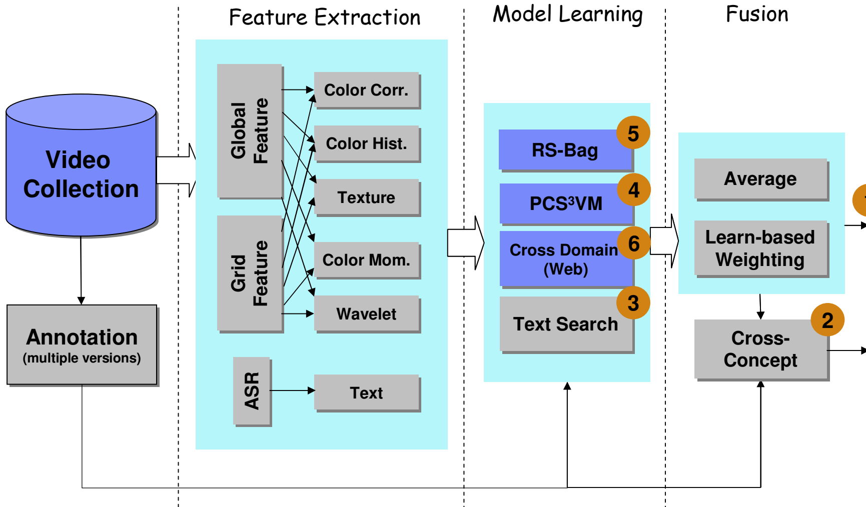
- Scalable, configurable and extensible based on UIMA and ActiveMQ
- Improve concept model throughput by 10x over last year



System Overview



System Overview



Outline

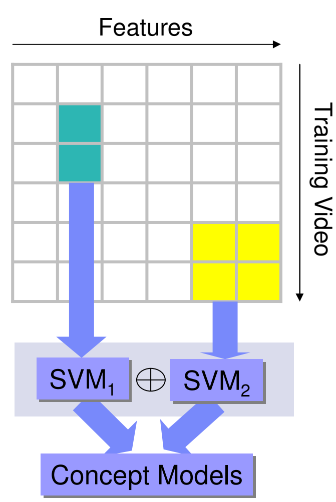
- **Baseline: Random subspace bagging**
- Principled Component Semi-Supervised SVM
- Learning from Web Domain
- Conclusion

Highlights on Feature Extraction

- Low-level features: significantly increase from 7 types to 98 types
  - Following the successful stories of previous TRECVID experiments
  - 13 different visual descriptors (e.g., color histogram, edge histogram …)
  - 8 granularities (i.e., global, center, cross, grid, horizontal parts, horizontal center, vertical parts and vertical center)
  - Future work: incorporating local features (interest point / dense sampled)
- Text search for concept detection
  - Text features officially provided by NIST
  - Juru search engine with manually expanded keywords on development set

Baseline: Random-Subspace Bagging

- Random-Subspace Bagging
  - For each concept and each low-level feature, select multiple bags of examples training set, sampled from **training examples and feature space**
  - Learn a SVM model on each bag
  - Evaluate the cross-validation performance
  - Select and fuse these models into an ensemble model based on held-out data
  - Similar to Random Forest w. decision trees
- Advantages:
  - Improve computational efficiency
  - Detection errors are theoretically bounded
  - Easy to parallelize using MapReduce w/o major changes for learning packages



Algorithm Details & Highlights

1. For f = 1 to F,
   For t = 1 to T,
   (a) Take a bootstrap sample X_+^t from positive data $\{\mathrm{x}_i\}$, $|X_+^t| = N_d$;
   (b) Take a bootstrap sample X_-^t from negative data $\{\mathrm{x}_i\}$, $|X_-^t| = N_d$;
   (c) Take a random sample F^t from the feature f with indices $\{1, ..., M\}$, $|F^t| = Mr_f$;
   (d) Learn a base model $h^{ft}(\mathrm{x})$ using X_+^t, X_-^t and F^t. SVMs with RBF kernel are used and the model parameters are chosen based on 3-fold cross validation;
2. For n = 1 to $F \cdot T$,
   (a) Select the n^{th} base model $h_n(\mathrm{x})$ with either *Greedy* or *Combined* strategy;
   (b) $F^t(\mathrm{x}) \leftarrow F^t(\mathrm{x}) + h_n(\mathrm{x})$;
   (c) Evaluate the composite classifier performance on validation data.
3. Output the best classifier on validation data.

Mapping Phase

- Learn SVM base model w. sampled data & features
- Evaluate cross-validation performance for the model

Reducing Phase

- Select n^{th} base model w. greedy/combined strategy
- Combine selected model

Recap: Random Subspace Bagging in TRECVID'07

- RSBag provides a more than 10-fold speedup on learning process and even a slightly better performance than the baseline SVM-07 approach

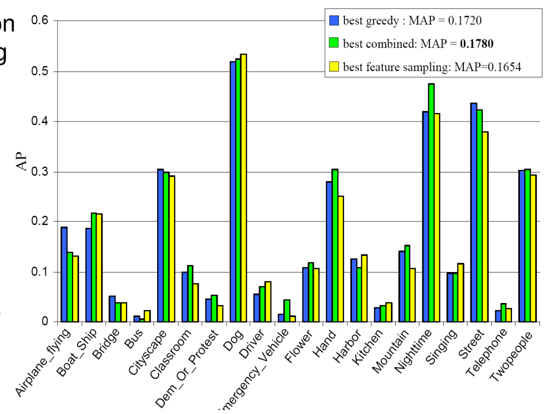
Description	Run	MAP	Time
SVM-07	-	0.0638	24602
RSBag	-	0.0667	2342
SVM-07 + SVM-Min07 + Text + HoG	-	0.0844	-
RSBag + SVM-Min07 + Text + HoG	-	0.0870	-

(RSBag: random subspace bagging w. 3 models, data sample ratio 0.2 and feature sample ratio 0.5)

TREC'08 Internal Validation Results for RSBag

- Setting: 70% of devel. data as training set, 30% as validation set

1. Consistently better validation performance with larger bag size, but saturate around 1000 training data per bag
2. The combined strategy consistently outperforms the greedy strategy
3. Sampling the feature space with a ratio of 50% gives slightly worse result, but much less training time



Official Evaluation Results: Baseline (RSBag), Text

- Observations
  - Fusion on multiple learning configurations / features almost always help, even when performance of individual component is not as good as baseline
  - Text features improve the performance on "Airplane", "Boat_Ship", but hurt the performance on "Nighttime", "Street"

Description	Run	Type	MAP
RSBag w. Greedy	-	A	0.0975
RSBag w. Combined	-	A	0.1010
Baseline (RSBag fused)	Baseline_5	A	0.1052
Text	-	A	0.0231
Baseline + Text	-	A	0.1240

Outline

- Baseline: Random subspace bagging
- Principled Component Semi-Supervised SVM
- Learning from Web Domain
- Conclusion

Principled Component Semi-Supervised SVM (PCS³VM)

Small sample learning difficulty: limited labeled data, high dimensionality

Existing solutions:

1. Semi-supervised learning: consider both X_L and X_U

   Pursue discrimination over X_L with constraints from both X_L and X_U, under some assumption about data distribution.

2. Dimension reduction: Reduce dimensionality of X

   Learn a feature subspace that preserves certain properties of the data distribution, e.g., PCA preserves data diversity and LDA preserves discriminant property.

PCS³VM: Joint Feature Subspace and SVM Learning

Proposed Solution:
Learn feature subspace that is discriminative over labeled data XL, and simultaneously learn large margin SVM in the feature subspace

Supervised SVM: learn classifier (**w**, **b)** pursing discrimination over X_L

$$\min_{\mathbf{w},\mathbf{b}} Q_d = \min_{\mathbf{w},\mathbf{b}} \left\{ \frac{1}{2} \| \mathbf{w} \|_2^2 + C \sum_{i=1}^{n_L} \varepsilon_i \right\},\ s.t.\ y_i(\mathbf{w}^T \phi(\mathbf{x}) + b) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0,\ \forall \mathrm{x}_i \in X_L$$

PCA: learn projection **a**, preserving the variance over entire data X

$$\max_{\mathbf{a}} Q_f = \max_{\mathbf{a}} tr\left\{ \mathbf{a}^T X X^T \mathbf{a} \right\},\ s.t.\ \mathbf{a}^T \mathbf{a} = I$$

Our Algorithm: jointly learn projection **a** and classifier (**w**, **b**)

$$\min_{\mathbf{w},\mathbf{b},\mathbf{a}} (Q_d - Q_f),\ s.t.\ \mathbf{a}^T \mathbf{a} = I,\ y_i(\mathbf{w}^T \mathbf{a}^T \phi(\mathbf{x}) + b) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0,\ \forall \mathrm{x}_i \in X_L$$

Evaluation Results: PCS³VM

- Observations
  - Although PCS³VM does not outperform baseline on average, it achieves better results on several concepts such as “nighttime” and “driver”.
  - Combined with baseline, it obtains 12% performance gain in terms of MAP.

Description	Run	Type	MAP
Baseline	Baseline_5	A	0.1052
PCS³VM	-	A	0.0750
Baseline + PCS³VM	BaseSSL_4	A	0.1182
Baseline + PCS³VM + Text	BaseSSLText_3	A	0.1268

Learning from Web Domain

- Online channels have provided a rich source of multimedia training data
  - How useful for learning TREC concepts?
- Approaches for web domain learning
  - Manually create 2-5 queries per concept
  - Download top 100-200 images from two different websites: Google & Flickr
  - Manually annotate all the images, end up with 30,000 images for 20 concepts
  - Use the baseline method to learn models

Detailed Evaluation Results on Learning Web Domain

- The success of learning from web domain depends on the generality of the target domain, e.g. “emergency vehicle” (+) and “dog” (-)
- Improve 6 out of 20 concepts, but MAP (0.052) is lower than baseline
  - The gap become much smaller (0.052 vs. 0.070) after removing 4 concepts

Concept	Baseline	Web
Emergency Vehicle	0.0076	0.0620 (rank 4th in 200 runs)
Airplane	0.0919	0.1339
Boat Ship	0.1568	0.1411
Dog	0.2508	0

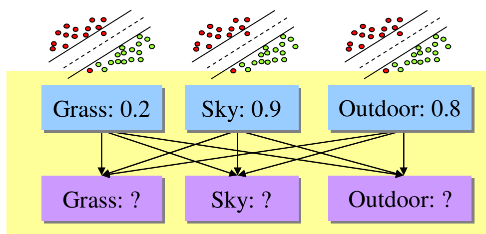
TRECVID **Web**

Cross-concept Detection to Combine Web Concepts

- Concept relational models to improve detection performance
  - Naïve Bayes algorithm by estimating pairwise conditional probabilities via maximum likelihood

$$L_c = \log \frac{P(y_c = 1 | y_{1:M})}{P(y_c = 0 | y_{1:M})} = \log \frac{P(y_c = 1)}{P(y_c = 0)} + \sum_{i=1:M} \log \frac{P(y_i | y_c = 1)}{P(y_i | y_c = 0)}$$

  - The conditional probabilities are smoothed by prior probabilities
  - Use 20 Type-A concepts and 200+ concepts learned from web domain



Evaluation Results: Web Domain & Cross-Concept

- Observations
  - Fusing web training data with baseline using multi-concept learning approaches, we can improve MAP by another 2%
  - Improve a number of concepts, e.g., "Airplane flying" and "Kitchen"

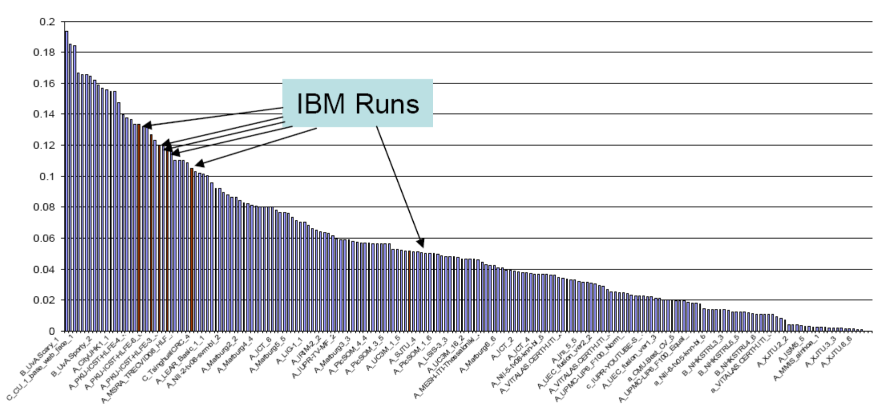
Description	Run	Type	MAP
Baseline	Baseline_5	A	0.1052
Web Domain Learning	CrossDomain_6	C	0.0519
Baseline + PCS^{3}VM	BaseSSL_4	A	0.1182
Cross-Concept: Baseline, Web, PCS^{3}VM	BNet_2	C	0.1200

Summarization of Submitted Runs

- Submitted: 5 runs + 1 Best Overall Run (~0.134 MAP)
- Almost all components help, even individual MAP is worse than baseline
- Best run improve over visual baseline by 30%

Description	Run	Type	MAP
Web Domain	CrossDomain_6	C	0.0519
Baseline	Baseline_5	A	0.1052
Baseline + PCS^{3}VM	BaseSSL_4	A	0.1182
Baseline + PCS^{3}VM + Text	BaseSSLText_3	A	0.1268
Cross-Concept: Baseline, Web, PCS^{3}VM	BNet_2	C	0.1200
Best Overall Run	BOR_1	C	0.1340

Overall Performance



Concluding Remarks

- Large-scale distribute machine learning system, improve computational throughput by 10x over last year
- Random subspace bagging considerably improve computational efficiency w/o hurting performance, also easy to parallelize
- PCS^{3}VM jointly learns feature space and large-margin SVMs, provide better performance after combined with baseline
- Web domain training data can be leveraged for generic concepts or infrequent concepts on target domain
- Future directions
  - More scalable distributed algorithm, local features, non-visual meta-data…