CMU-Informedia @ TRECVID 2014 Semantic Indexing

Lu Jiang, Xiaojun Chang, Zexi Mao, Anil Armagan, Zhengzhong Lan, Xuanchong Li, Shoou-I Yu, Yi Yang, Deyu Meng, Pinar Duygulu-Sahin, Alexander Hauptmann

Carnegie Mellon University

November 10, 2014

People

- CMU Informedia Team

Xiaojun Chang

Zexi Mao

Anil Armagan

Zhengzhong Lan

Xuanchong Li

Shoou-I Yu

Yi Yang

Deyu Meng

Pinar Duygulu-Sahin

Alexander Hauptmann

Acknowledgement

- This work has been supported in part by the National Science Foundation under Grant Number IIS-12511827, by the Department of Defense, U. S. Army Research Office (W911NF-13-1-0277) and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, NSF, ARO or the U.S. Government.
- This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the BlackLight system at the Pittsburgh Supercomputing Center (PSC).

Outline

- Submission Review
- Going Beyond 60 concepts
  - Challenges
  - Theory
  - Implementations
- Summary

Outline

- **Submission Review**
- Going Beyond 60 concepts
  - Challenges
  - Theory
  - Implementations
- Summary

Overview

- The training data is the same one used in 2013
  - IACC.1.tv10.training and IACC.1.A-C collections
- Our system includes:
  - Self-paced SVM pipeline (**discuss later in this talk**)
  - Deep Convolutional Neural Networks (DCNN)-based

Self-paced SVM Pipeline

- Individual feature performances on IACC.2.B.
  - Bow features: code book size 4,096, intersection kernel.
  - Fisher vector feature: dimension 109,056, linear kernel.
  - Intersection kernels

Raw feature	Representation	infMAP
SIFT harris-laplace	Spatial Bow	0.0866
SIFT dense-sampling	Spatial Bow	0.1096
CSIFT harris-laplace [3]	Spatial Bow	0.0842
CSIFT dense-sampling [3]	Spatial Bow	0.0988
Improved Dense Trajectory [1]	Fisher Vector (non-spatial)[2]	0.1844

[1] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," *in ICCV*, 2013.
[2] K. Chatfield, A. Lempitsky and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in BMVC, 2011.
[3] K. Sande, T. Gevers and C. Snoek, "Evaluating color descriptors for object and scene recognition," TPAMI, 2010.

Self-paced SVM Pipeline

- Individual feature performances on IACC.2.B.
  - Bow features: code book size 4,096, intersection kernel.
  - Fisher vector feature: dimension 109,056, linear kernel.
  - Intersection kernel.

Raw feature	Representation	infMAP
SIFT harris-laplace	Spatial Bow	0.0866
SIFT dense-sampling	Spatial Bow	0.1096
CSIFT harris-laplace	Spatial Bow	0.0842
CSIFT dense-sampling	Spatial Bow	0.0988
Improved Dense Trajectory	Fisher Vector (non-spatial)	0.1844

- **O1: Improved dense trajectory is the best single feature.**
- **O2: Dense-sampling seems to be better than harris-laplace.**

Feature Fusion

- Feature fusion performances on IACC.2.B.
  - CMU_Run1: heuristic fusion + related concepts propagation + junk-frame removal.

Raw feature	Comments	infMAP
SIFT	(harris + dense)	0.0963
CSIFT	(harris + dense)	0.0962
SIFT + CSIFT	Average fusion	0.1208
Improved Dense Trajectory	Fisher Vector (non-spatial)	0.1844
All Features Fusion	CMU_Run1	0.2265

- **O3: SIFT and CSIFT offers complementary info to the motion features.**

DCNN-based Pipeline

- Directly trained on keyframes.
  - Trained 347 concepts (346 + NULL)
  - Two strategies for unbalanced data:
    - Duplicate the positive training samples.
    - Not duplicate positive training samples.
    - Fusing the two result.

Raw feature	Comments	infMAP
SIFT+CISFT	Self-paced SVM	0.121
DCNN-pipeline	DCNN-based	0.134

- **O4: DCNN-pipeline yields better performance than the static features fusion in SVM-pipeline [1], but not as good as improved dense trajectory (0.184).**

[1] Z. Z. Lan, Y. Yi, N. Ballas, S. Yu, A. Hauptmann , "Resource Constrained Multimedia Event Detection, " in MMM, 2014.

Main Submissions

Runs are under Type A condition (TRECVID data only)

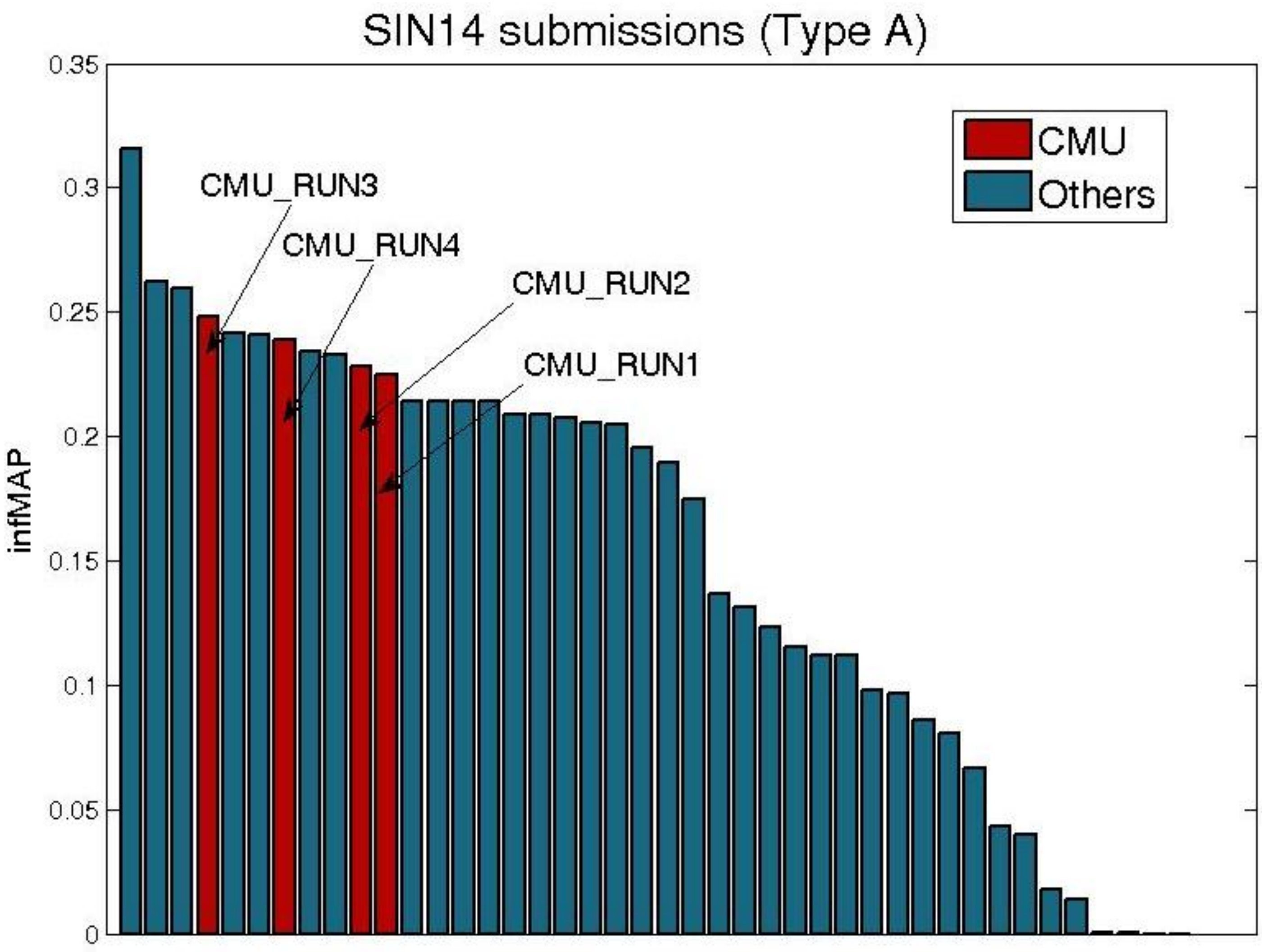
- CMU_Run1: baseline run by Self-paced SVM pipeline.
- CMU_Run2: averages CMU_Run1 with DCNN-based pipeline on 15/60 concepts.
- CMU_Run3: CMU_Run2 + MMPRF [1] by visual and metadata feature.
- CMU_Run4: CMU_Run2 + weighted fusion (learned on the results on IACC.2.A)

Run ID	infMAP	infNDCG	P@10	P@100
CMU_Run1	0.2265	0.4660	0.6700	0.5583
CMU_Run2	0.2297	0.4710	0.6900	0.5683
CMU_Run3	**0.2480**	**0.4975**	**0.7000**	**0.5900**
CMU_Run4	0.2403	0.4844	0.6900	0.5730

- **O5: MMPRF (MultiModal Pseudo Relevance Feedback) offers reasonable improvements (relative 8.0%, 1.8% absolute).**
- **O6: Weighted fusion yields reasonable improvements (relative 4.6%, absolute 1.1%).**

[1] L. Jiang, T. Mitamura, S.-I. Yu, A. Hauptmann. Zero-Example Event Search using MultiModal Pseudo Relevance Feedback. In ICMR, 2014

Main Submissions



No Annotation Submissions

- SVM models trained on web images retrieved by Bing.
- Maximum 1000 images for a concept.
- SIFT Feature + SVM RBF kernel.

Run ID	Pipeline	infMAP	infNDCG	P@10	P@100
CMU_Run5	no-annotation	0.0118	0.1099	0.1100	0.0757
CMU_Run6	no-annotation	0.0085	0.0956	0.0967	0.0680

- **O7: Domain difference between still images and video shots is huge!**

Observations

- **O1:** Improved dense trajectory is the best single feature.
- **O2:** Dense-sampling seems to be better than harris-laplace.
- **O3:** SIFT&CSIFT offers complementary info to the motion features.
- **O4:** DCNN-pipeline yields better performance than the static features fusion in SVM-pipeline, but not as good as improved dense trajectory.
- **O5:** MMPRF offers reasonable improvements.
- **O6:** Weighted fusion yields reasonable improvements.
- **O7:** Domain difference between still images and videos is huge!

Outline

- Submission Review
- Going Beyond 60 concepts
  - **Motivation and Challenges**
  - Theory
  - Implementations
- Summary

Motivation and Challenges

- SIN 14 task: 60 concepts on 200k shots.
- SIN Full: 346 concepts on 500k shots.
- What if we go beyond: 1,000 concepts on 1 million shots.

- **Many larger shot-based datasets are out there:**
  - Yahoo **YFCC100M (0.8 million videos)** with tags & descriptions.
  - **Google Sports (1.1 million videos)** with automatically generated labels.
  - Data are noisy and no clean ground-truth labels are available in both datasets.

- **Everybody knows that more concepts are better.**
  - Recognize more objects/scenes/actions in videos.
  - Usually lead to improvement on search and retrieval.

[1] Yahoo YFCC http://labs.yahoo.com/news/yfcc100m/
[2] Google Sports https://code.google.com/p/sports-1m-dataset/

Motivation and Challenges

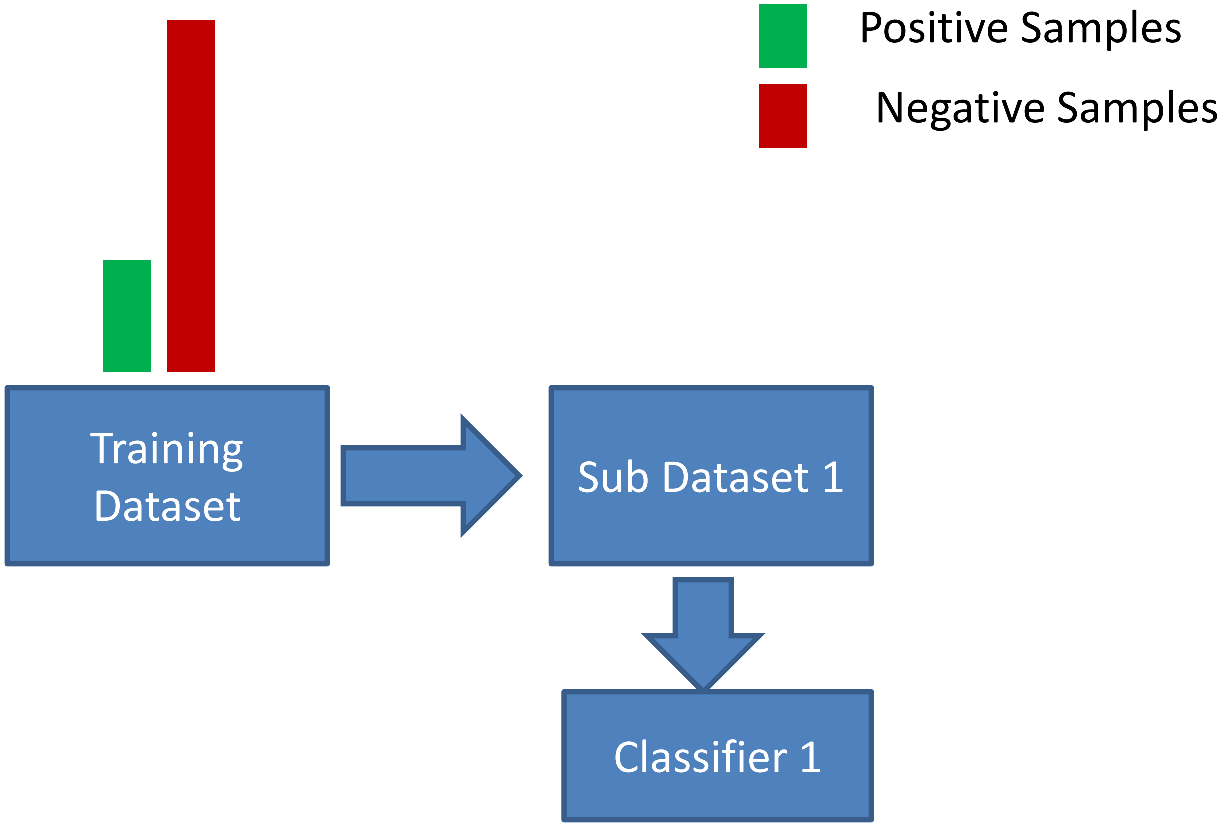
Carnegie Mellon University

- Large-scale concept training is challenging:
  - How to train robust models on **millions of shots efficiently**?
  - How to handle the **noisy big data** (no clean labels)?
- Existing approaches:
  - Augmented CascadeSVM – CMU Informedia [1]
  - Cascade SVMs – MediaCCNY [2]
  - Negative Bootstrapping – MediaMill [3,4]
  - Unit Models – IBM [5]

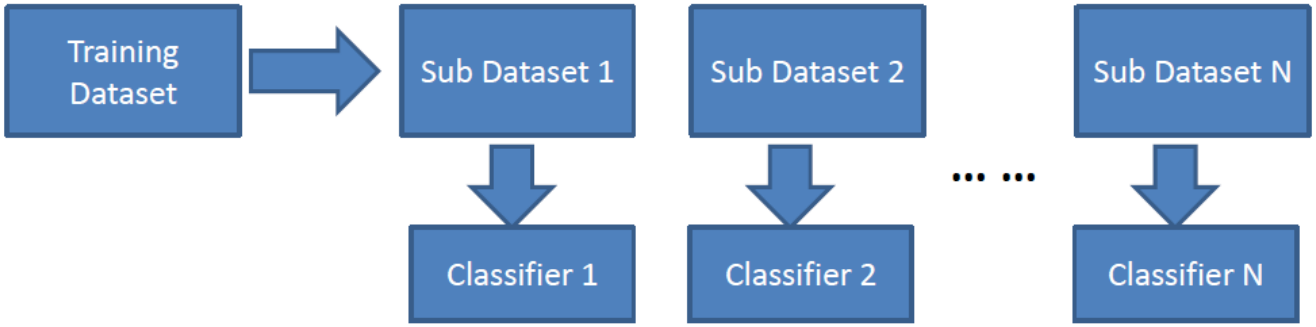
[1] Bao, Lei, et al. "Informedia@ trecvid 2011." *TRECVID2011, NIST* (2011).
[2] Yang, Xiaodong, et al. "MediaCCNY at TRECVID 2012: Surveillance event detection." *NIST TRECVID, Workshop*. 2012.
[3] Li, Xirong, et al. "Bootstrapping visual categorization with relevant negatives."*IEEE Transactions on Multimedia* 15.4 (2013): 933-945.
[4] Snoek, C. G. M., et al. "MediaMill at TRECVID 2013: Searching concepts, objects, instances and events in video." *NIST TRECVID Workshop*. 2013.
[5] Cao, Liangliang, et al. "IBM research and columbia university trecvid-2012 multimedia event detection (med), multimedia event recounting (mer), and semantic indexing (sin) systems." *Proc. TRECVID 2012 workshop. Gaithersburg, MD, USA*. 2012.

Tackling the highly imbalanced data

- Augmented Cascade SVM.
- Select negative samples in a sequential manner based on the learned model.



Tackling the highly imbalanced data



Pros:

- Reasonable solution for handling large dataset.

Cons:

- Most are **heuristic** approaches (random sampling).
- **Ad-hoc strategies** for selecting samples.

Outline

- Submission Review
- Going Beyond 60 concepts
  - Motivation and Challenges
  - **Theory**
  - Implementations
- Summary

Self-paced Learning

- Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al 2010) is a recently proposed learning paradigm that is inspired by the learning process of humans and animals.
- The samples are not learned randomly but organized in a meaningful order which illustrates from **easy** to gradually more **complex** ones.

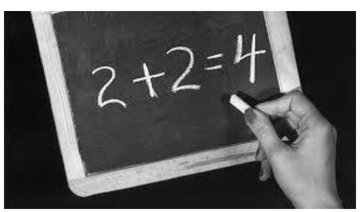
Prof. Bengio

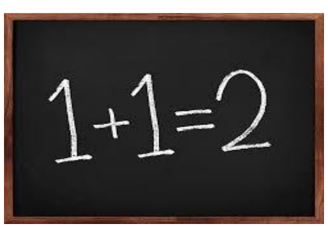
Prof. Koller

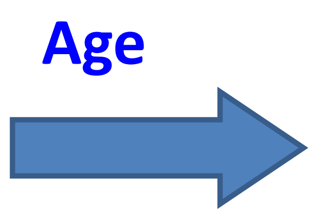
Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML, 2009.*
M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In NIPS, pages 1189–1197, 2010.

Self-paced Learning

- Easy samples to complex samples.
  - Easy sample → smaller loss to the already learned model.
  - Complex sample → bigger loss to the already learned model.







$$
\begin{aligned}
\frac{1}{g-kv}\frac{dv}{dt} &= 1 \\
\int_0^T \frac{1}{g-kv}\frac{dv}{dt}\,dt &= \int_0^T dt \\
\int_{v_0}^{v(T)} \frac{1}{g-kv}\,dv &= T \\
-\frac{1}{k}\ln|g-kv|\,\Big|_{v_0}^{v(T)} &= T \\
\ln\left|\frac{g-kv(T)}{g-kv_0}\right| &= -kT \\
\frac{g-kv(T)}{g-kv_0} &= e^{-kT}
\end{aligned}
$$

Self-paced Learning

- Easy samples to complex samples.
  - Easy sample → smaller loss to the already learned model.
  - Complex sample → bigger loss to the already learned model.

Easy samples of "bus"

Age

Complex samples of "bus"

Easy and Complex samples in Pascal VOC dataset

Easy training samples of “Chair” in **Pascal VOC** dataset

Complex training samples of “Chair” in **Pascal VOC** dataset

Similar observations are also found by the others (Lapedriza et al. 2013)

A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba. Are all training examples equally valuable?CoRR abs/1311.6510, 2013.

Easy and Complex samples in Google Image Search

Easy training samples of “Dog” returned by **Google Image**

Difficult training samples of “Dog” returned by **Google Image**

Self-paced Learning

- In self-paced learning, we optimize the following function:

Loss

$$\arg\min_{\mathbf{w},\mathbf{v}} \sum_{i=1}^{n} \boxed{v_i L_i} + f(\mathbf{v}, \lambda)$$

L_i : the loss for the ith sample. Can be any loss in off-the-shelf model, e.g. SVMs neural networks.

$v_i \in [0, 1]$: the weight for the ith sample.

The loss is discounted by a sample weight.

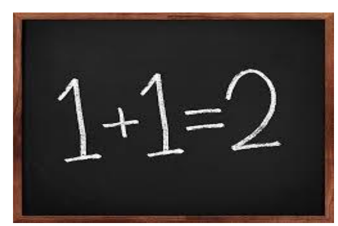
Self-paced Learning

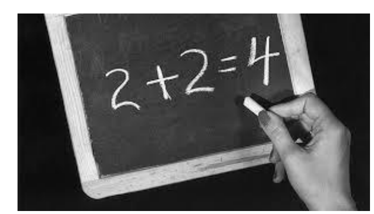
- In self-paced learning, we optimize the following function:

Self-paced function

$$\arg\min_{\mathbf{w},\mathbf{v}} \sum_{i=1}^{n} v_i L_i + \boxed{f(\mathbf{v}, \lambda)} \qquad \mathbf{v} = [v_1, \ldots v_n]$$

- The self-paced function determines a learning scheme on how models learn new samples.
- **Physically it corresponds to learning schemes that human use to learn different tasks.**



More Self-paced Functions

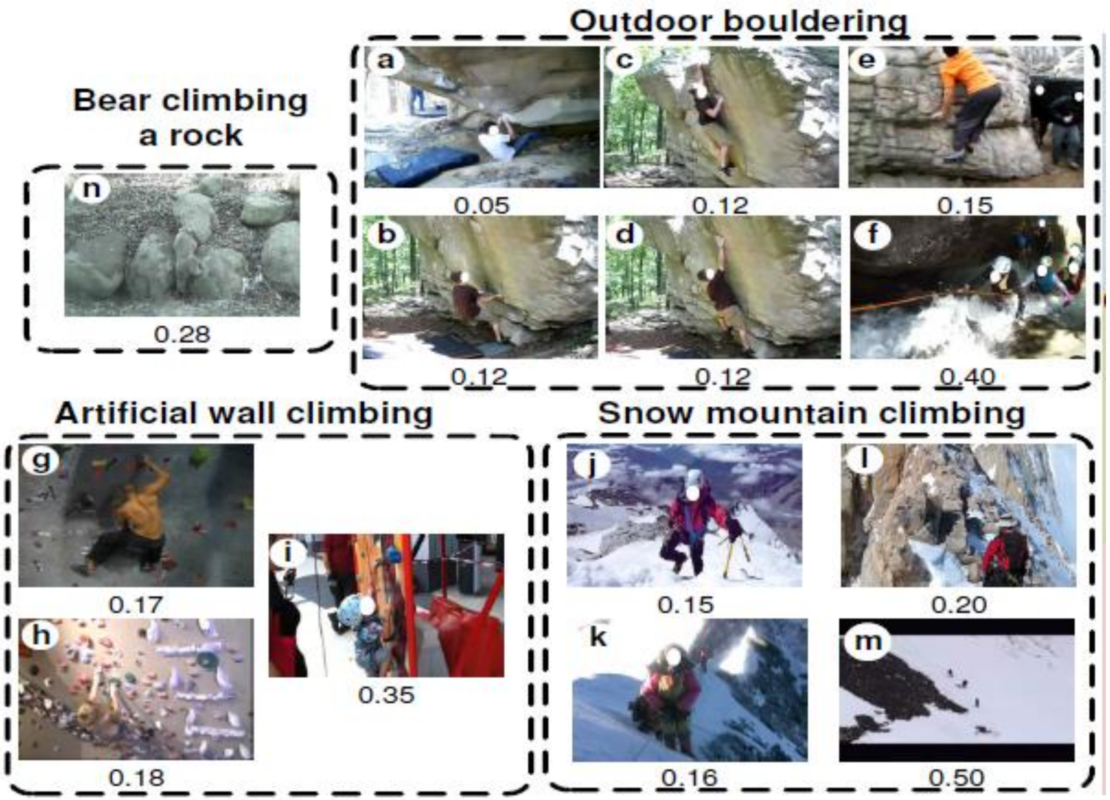
- *Binary:* $$f(\mathbf{v};\lambda) = -\lambda\|\mathbf{v}\|_1$$
- *Linear:* $$f(\mathbf{v};\lambda) = \lambda(\frac{1}{2}\|\mathbf{v}\|_2^2 - \sum_{i=1}^{n} v_i)$$
- *Logarithmic:* $$f(\mathbf{v};\lambda) = \sum_{i=1}^{n} \zeta v_i - \frac{\zeta^{v_i}}{\log \zeta} \qquad \zeta = 1-\lambda, (0<\lambda<1)$$
- *Mixture:* $$f(\mathbf{v};\lambda,\gamma) = -\zeta \sum_{i=1}^{n} \log(v_i + \frac{\zeta}{\lambda}) \qquad \zeta = \frac{\gamma\lambda}{\lambda-\gamma}, (\lambda>\gamma>0)$$
- ***Diversity*:*** $$f(\mathbf{v};\lambda,\gamma) = -\lambda\|\mathbf{v}\|_1 - \gamma\|\mathbf{v}\|_{2,1}$$

[1]L. Jiang, D. Meng, T. Mitamura and A. Hauptmann."Easy Samples First: Self-paced Reranking for Zero-Example Multimedia Search." *ACM International Conference on Multimedia*. ACM, 2014.

[1] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In NIPS, pages 1189–1197, 2010.

*Function is non-convex but still can find optimal.

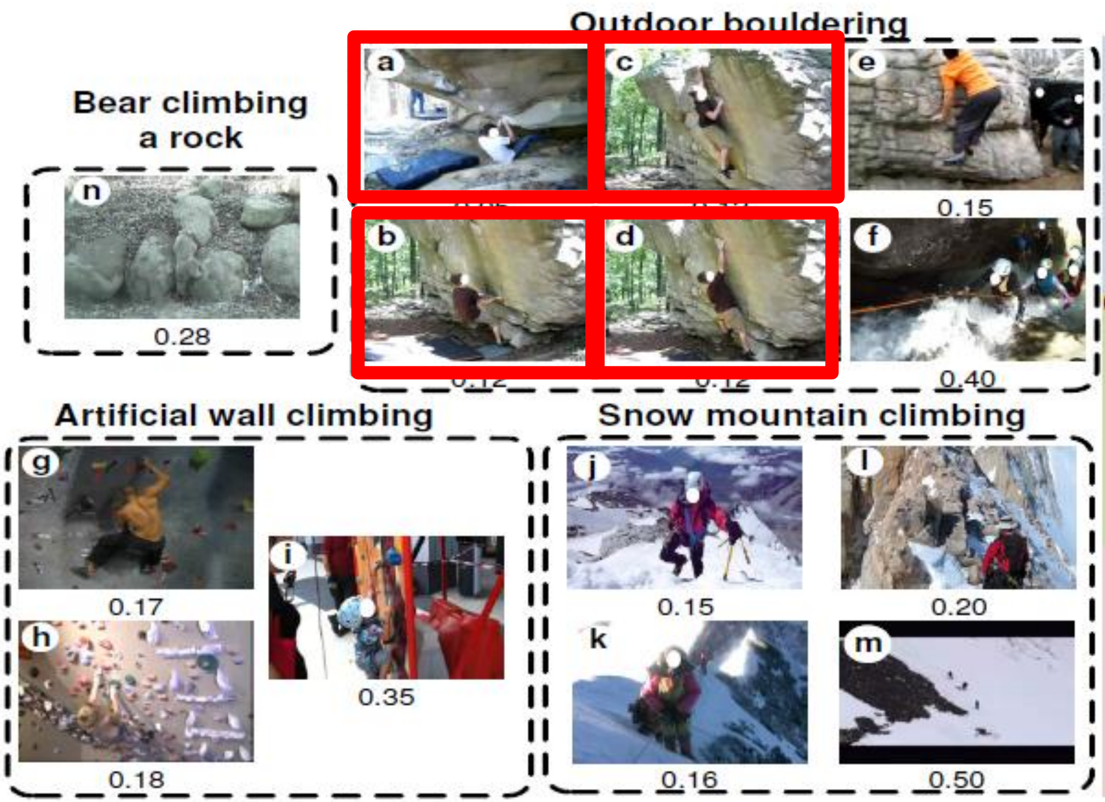
Learning with Diversity



Learning with Diversity

- Learning easy samples:

$$f(\mathbf{v};\lambda) = -\lambda\|\mathbf{v}\|_1$$

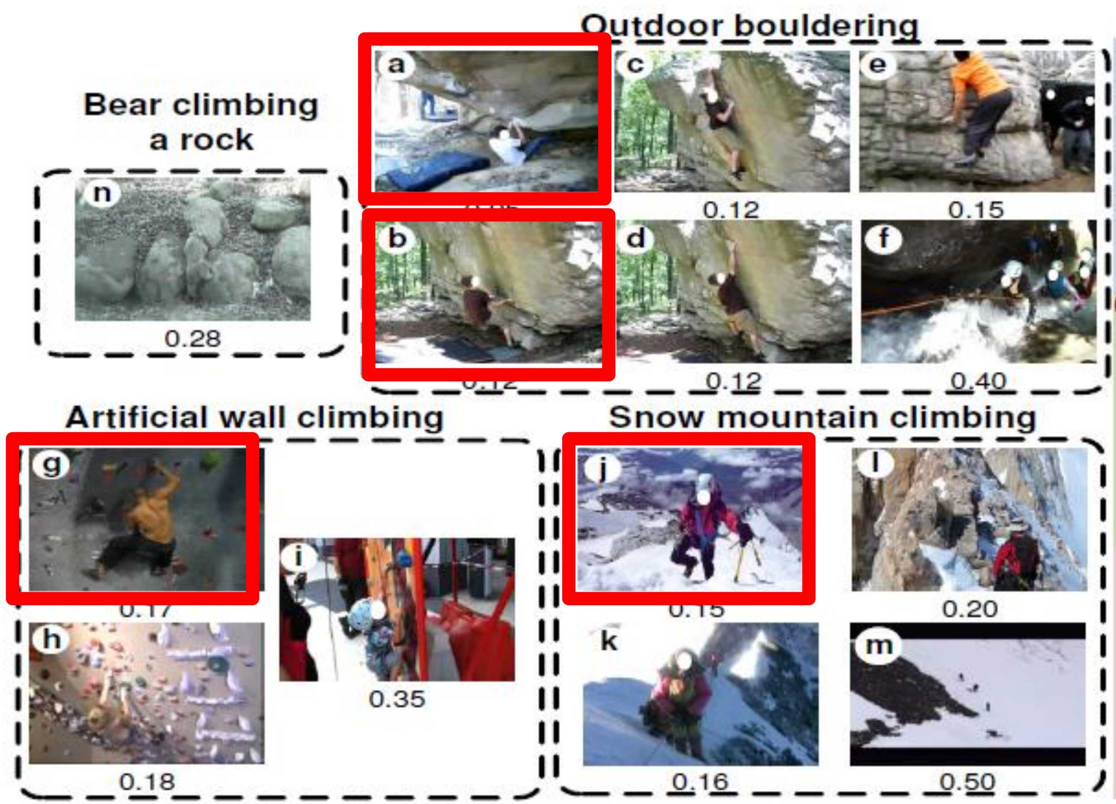


[1] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, A. Hauptmann, "Self-paced Learning with Diversity," in *NIPS*, 2014.

Learning with Diversity

- Learning easy and diverse samples[1]:

$$f(\mathbf{v};\lambda,\gamma) = -\lambda\|\mathbf{v}\|_1 - \gamma\|\mathbf{v}\|_{2,1}$$

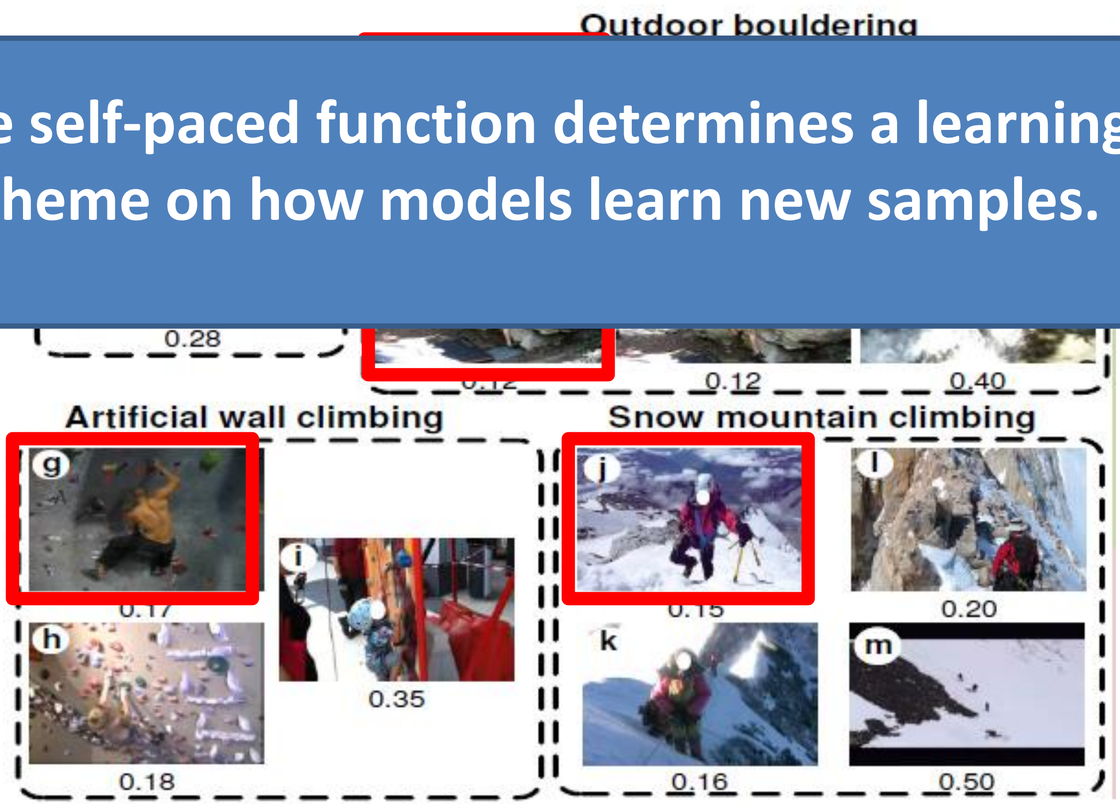


[1] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, A. Hauptmann, "Self-paced Learning with Diversity," in *NIPS*, 2014.

Learning with Diversity

- Learning easy and diverse samples[1]:

$$f(\mathbf{v}; \lambda, \gamma) = -\lambda \|\mathbf{v}\|_1 - \gamma \|\mathbf{v}\|_{2,1}$$



The self-paced function determines a learning scheme on how models learn new samples.

[1] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, A. Hauptmann, "Self-paced Learning with Diversity," in *NIPS*, 2014.

Outline

- Submission Review
- Self-paced Concept Learning
  - Motivation and Challenges
  - Theory
  - **Implementations**
- Summary

Practical lessons

- Training a quadratic programming problem with linear kernel
  - Primal (to obtain the parameters in the original space)
  - Dual (to obtain the support vectors)
- For nonlinear kernels, apply explicit feature mapping[1].

[1] Vedaldi, Andrea, and Andrew Zisserman. "Efficient additive kernels via explicit feature maps." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34.3 (2012): 480-492.

Practical lessons

Primal	Dual
Efficient in testing	**Efficient in training** with pre-computed kernel (preferred in shared memory)
Low memory usage	Minimum duplicate computation
	Good for high-dimensional dense vector

- It used to take 60 days on 1000 cores to extract SIN features for 100k videos using dual form. Now it takes **1 day on 32 cores using primal form**.
- **Pre-compute kernel → Training (dual form)→ Testing (primal)**

Outline

- Submission Review
- Self-paced Concept Learning
  - Challenges
  - Theory
  - Implementations
- **Summary**

Conclusions

- We have built tools for shot-based concepts training on big data.
  - Suppose we have **500 concepts** each of which has 1,000 positive videos (**500,000 in total**).
  - Using the improved dense trajectory feature (best single feature with 100k dimension).
  - We can finish the training within **48 hours on 512 CPU cores**.
  - After getting the models, the prediction for a shot/video **only takes 0.125s** on a 16-core machine with 16GB memory.

- The feature extracted by this pipeline can be used for some other tasks e.g. multimedia event detection (more tomorrow).

THANK YOU.

Q&A?

APPENDIX

Practical Discussions

- Practical lessons for applying self-paced learning in your problems:
  - Choose reasonable starting values using prior knowledge[1].
  - Pace positive/negative separately for unbalanced data
  - Pace the age parameter so that it includes a certain number of samples for the next iteration.
  - Use reasonable validation sets to determine the optimal age of the final model (when to stop), which follows a similar distribution as the test set . Physically it corresponds to mock exams used to evaluate the learning progress.

[1] L. Jiang, D. Meng, Q. Zhao, S. Shan, A. Hauptmann, "Self-paced Curriculum Learning" in *AAAI*, 2015.