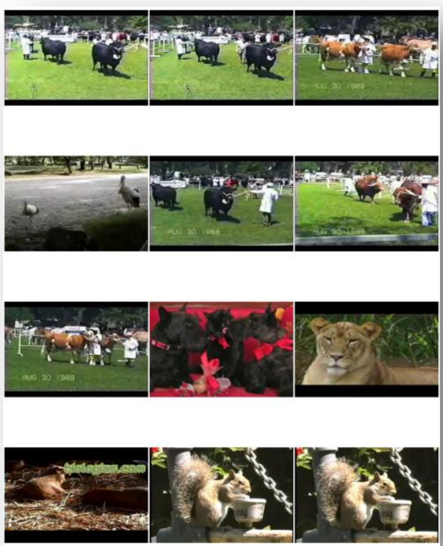
ONLINE MULTI-TASK LEARNING FOR SEMANTIC CONCEPT DETECTION IN VIDEO

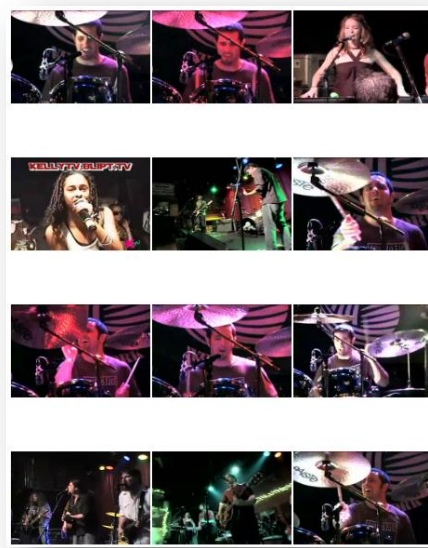
Foteini Markatopoulou[1,2], Vasileios Mezaris[1], and Ioannis Patras[2]
[1]Information Technologies Institute / Centre for Research and Technology Hellas
[2]Queen Mary University of London

Problem

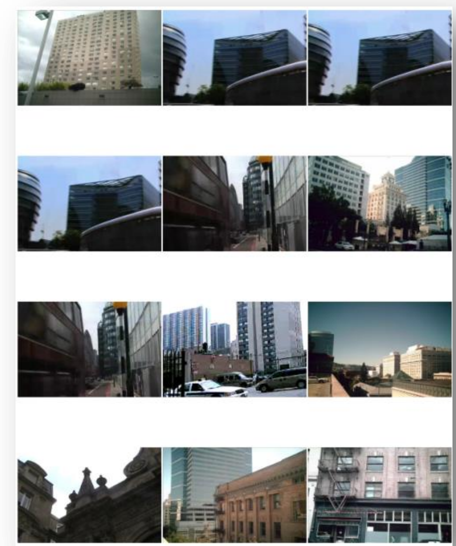
- Concept-based video retrieval (38 evaluated concepts)
- TRECVID SIN Task video dataset



animal

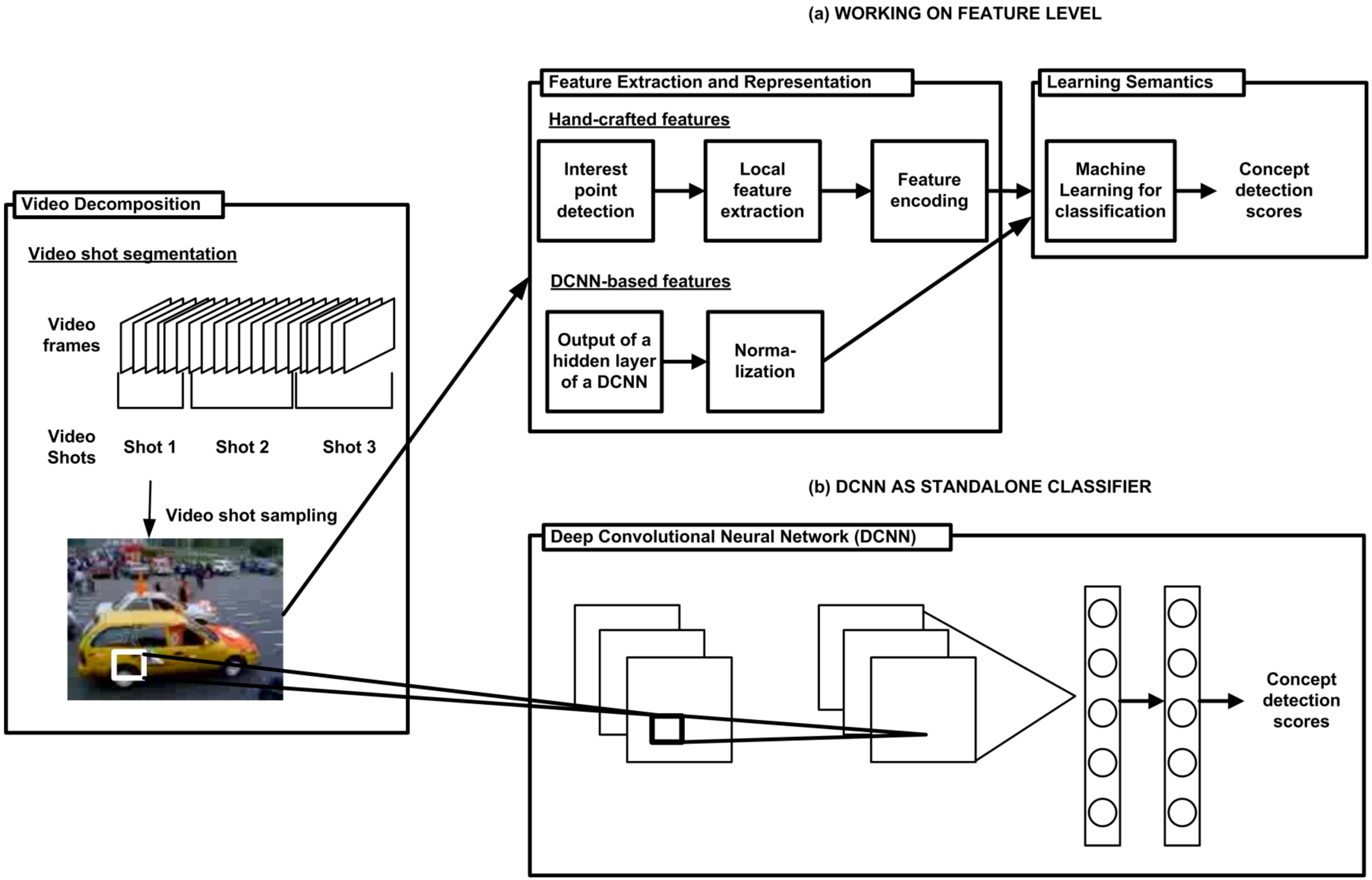


singing

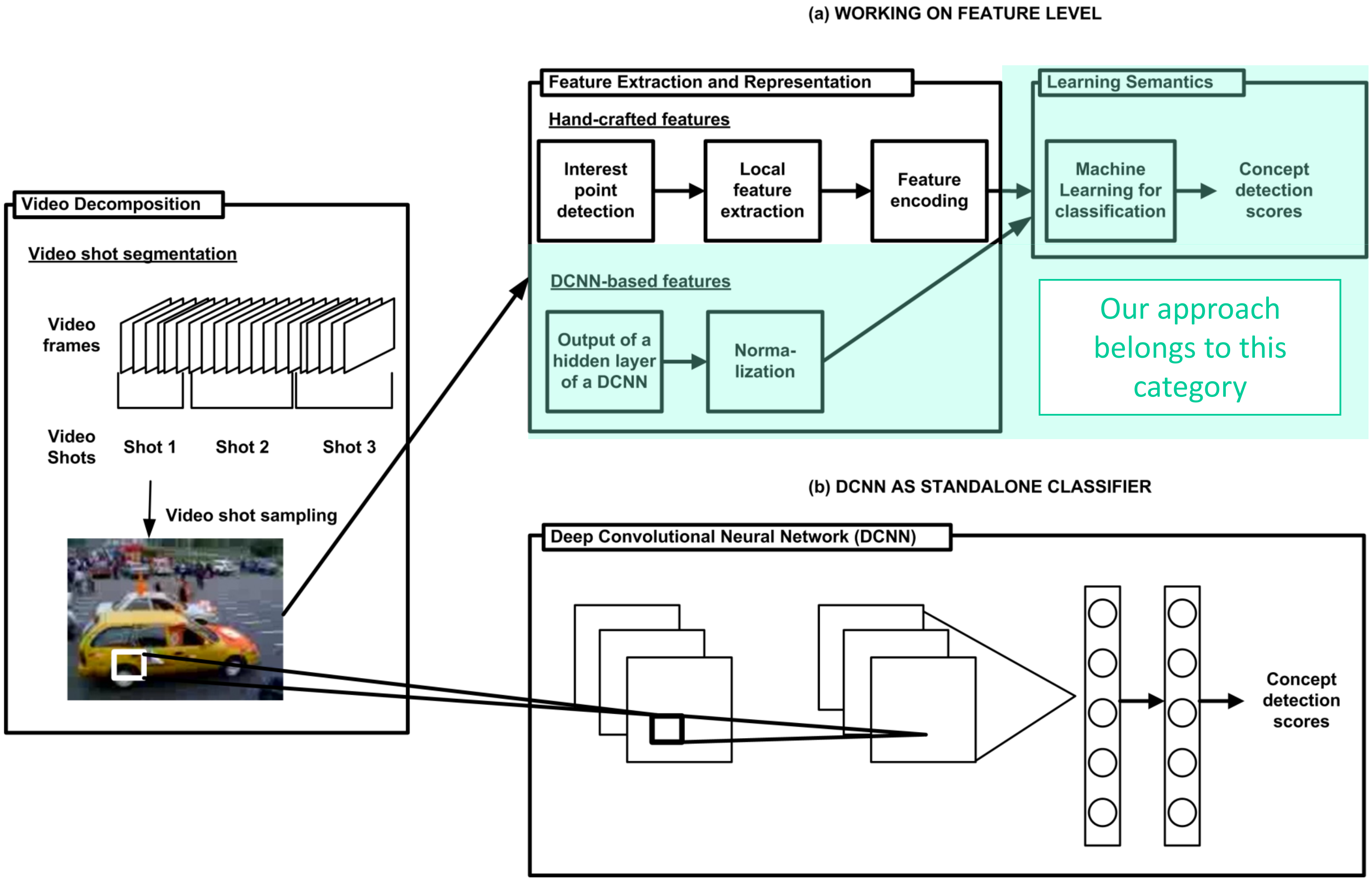


building

Typical solution



Typical solution



Information Technologies Institute

Queen Mary University of London

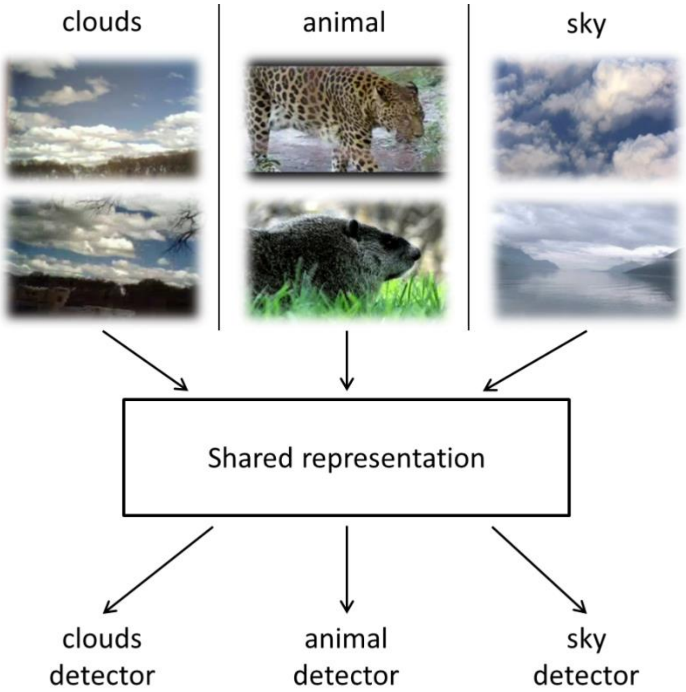
InVID

Motivation for going beyond the typical solution

- Typical concept detection: Train one supervised classifier separately for each concept; a single-task learning process (STL)
- However, concepts do not appear in isolation from each other

Label relations

Task relations



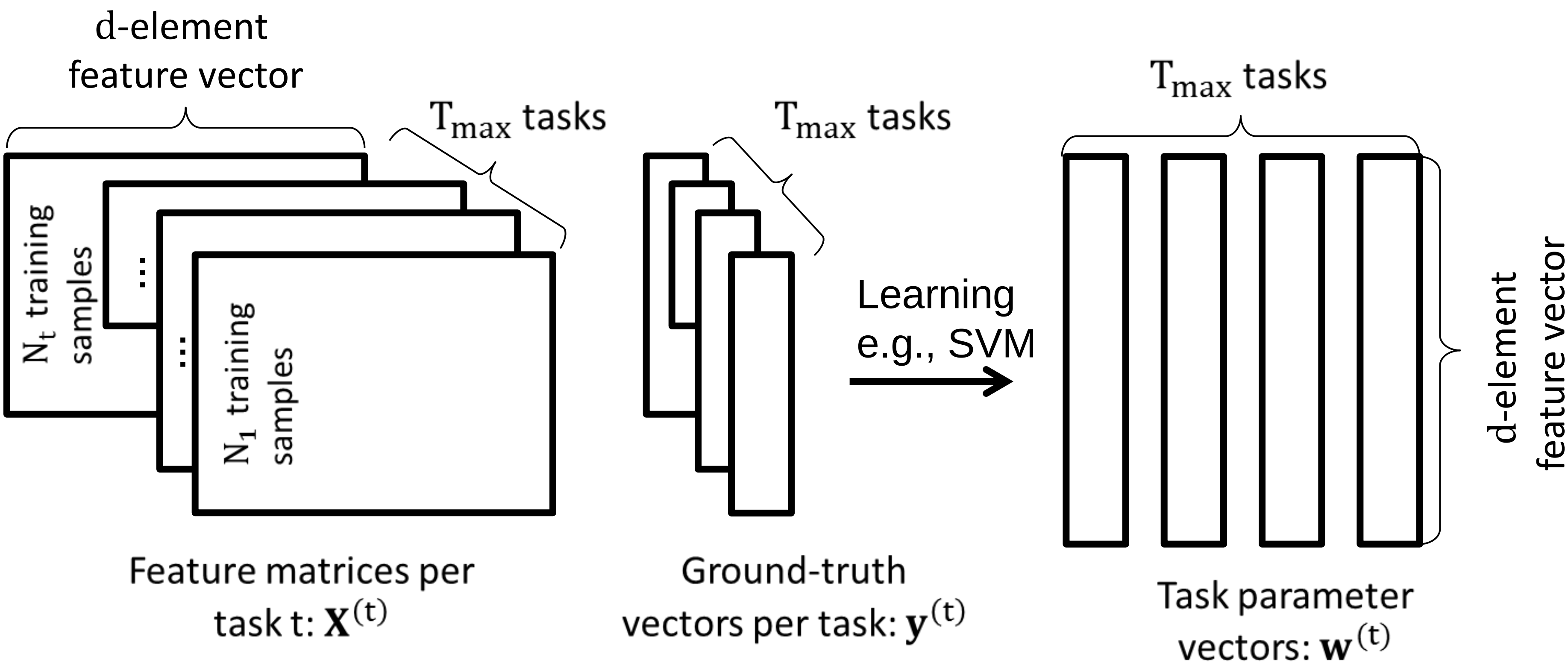
Literature review

- Multi-concept learning (MCL): Exploit concept relations
  - Stacking-based approaches (Smith et al. 2003), (Markatopoulou et al. 2014)
  - Inner learning approaches (Qi et al. 2007)
- Multi-task learning (MTL): Exploit task relations (learn many tasks together)
  - Assuming all tasks are related e.g., use regularization (Argyriou et al. 2007)
  - Some tasks may be unrelated e.g., CMTL (Zhou et al. 2011), AMTL (Sun et al. 2015), GO-MTL (Kumar et al. 2012)
  - Online MTL for lifelong learning e.g., ELLA (Eaton & Ruvolo 2013)

Our approach

- Proposed method: ELLA_LC
  - ELLA_LC stands for Efficient Lifelong Learning Algorithm with Label Constraint
  - It jointly considers task and label relations
  - ELLA_LC is based on ELLA (Eaton & Ruvolo 2013)
  - ELLA is the online version of GO-MTL: Learning Task Grouping and Overlap in Multi-Task Learning (Kumar et al. 2012)

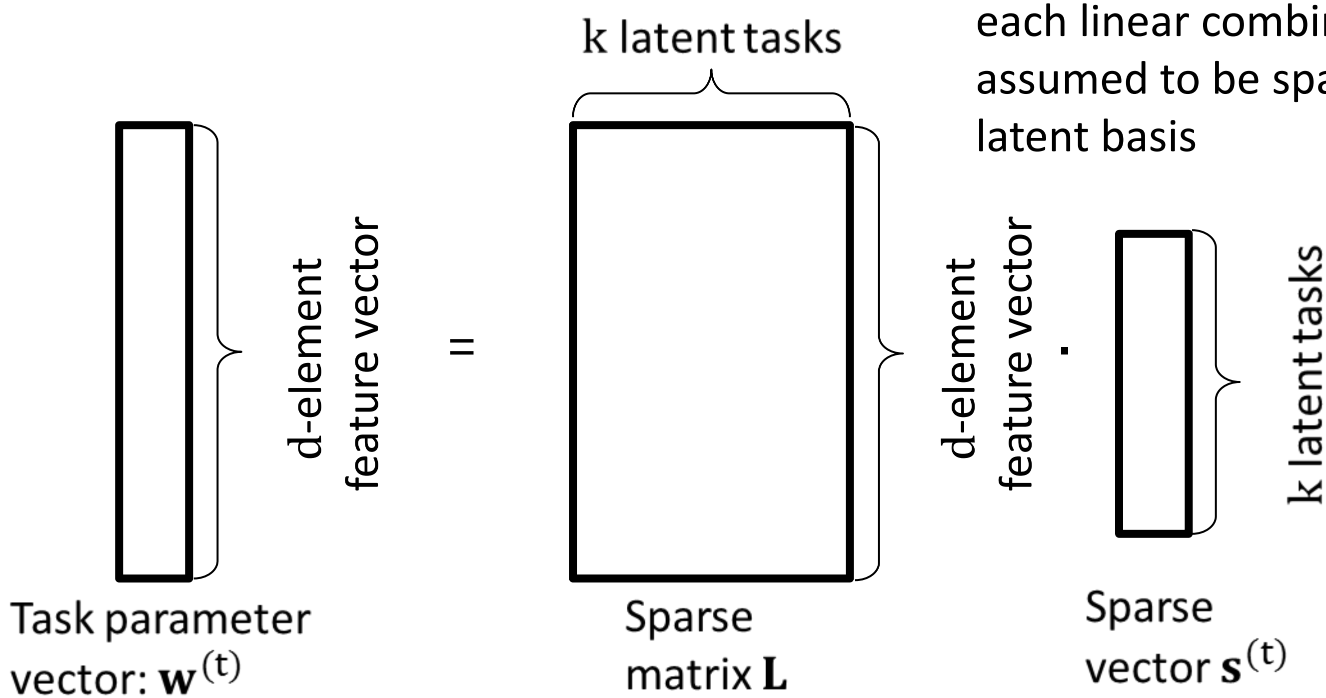
Background: Single-task learning



- We focus on linear models: $\mathbf{y}^{(t)} = \mathbf{X}^{(t)} \mathbf{w}^{(t)}$
- $\mathbf{X}^{(t)} \in \mathbb{R}^{N_t \times d}$, $\mathbf{y}^{(t)} \in \mathbb{R}^{N_t}$
- $\mathbf{w}^{(t)} \in \mathbb{R}^{d}$

Background: The GO-MTL algorithm

each linear combination is assumed to be sparse in the latent basis



ensuring sparsity of $\mathbf{L}$

ensuring sparsity of $\mathbf{S}$ (matrix $\mathbf{S}$ concatenates the weight vectors $\mathbf{s}^{(t)}$ from all the tasks)

- Knowledge shared basis: $\mathbf{L} \in \mathbb{R}^{d \times k}$
- Task-specific weight vector: $\mathbf{s}^{(t)} \in \mathbb{R}^{k}$
- $\mathbf{w}^{(t)} = \mathbf{L}\mathbf{s}^{(t)}$

- Objective function:

$$\min_{(\mathbf{L},\mathbf{s}^{(t)})} \sum_{t=1}^{T_{max}} \left\{ \sum_{i=1}^{N_t} \mathcal{L}\left(D\left(\mathbf{x}_i^{(t)}; \mathbf{L}\mathbf{s}^{(t)}\right), y_i^{(t)}\right) + \mu\|\mathbf{S}\|_1 + \lambda\|\mathbf{L}\|_F^2 \right\}$$

loss function

base learner e.g., LSVM, LR

Information Technologies Institute

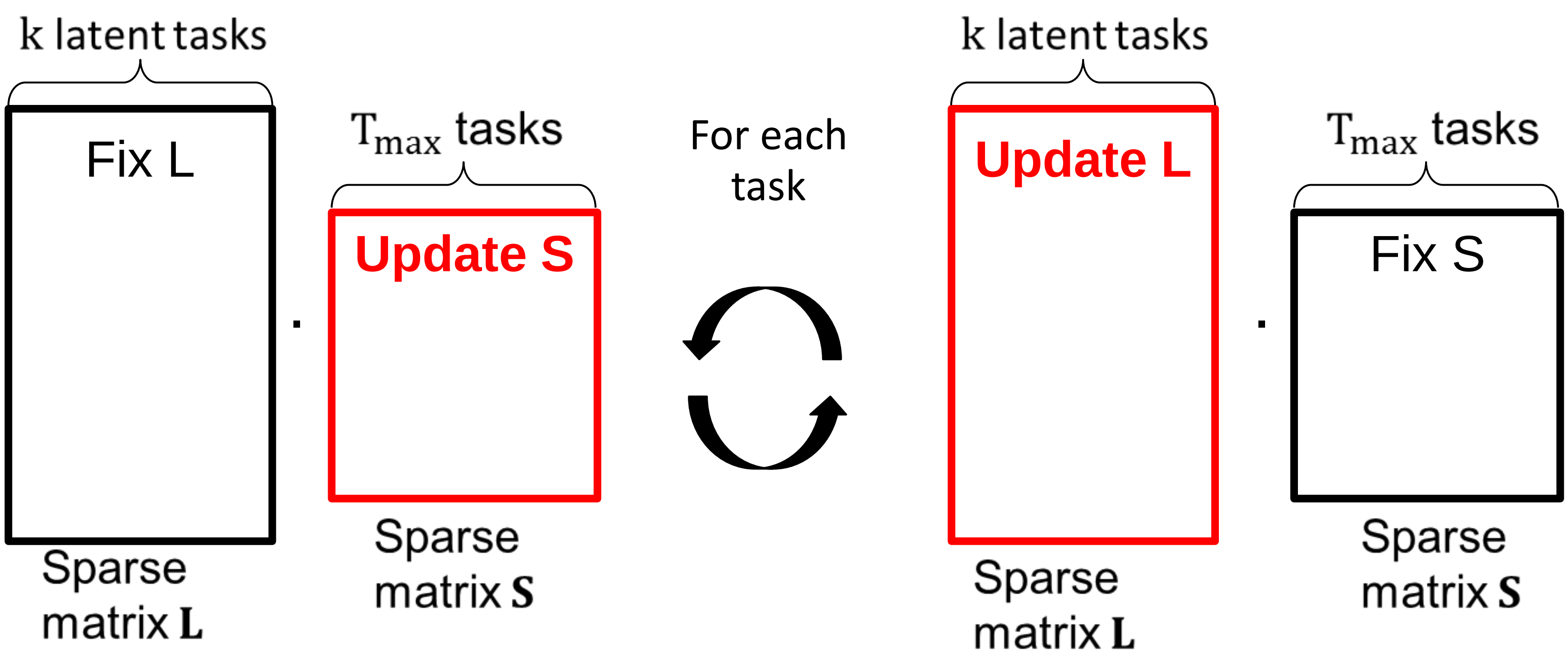
Queen Mary University of London

InVID

Background: The GO-MTL algorithm

Iterative optimization with respect to $\mathbf{L}$ and $\boldsymbol{S}$:

$$\min_{(\mathbf{L},\mathbf{s}^{(t)})} \sum_{t=1}^{T_{max}} \left\{ \sum_{i=1}^{N_t} \mathcal{L}\left(D\left(\mathbf{x}_i^{(t)}; \mathbf{L}\mathbf{s}^{(t)}\right), y_i^{(t)}\right) + \mu\|\mathbf{S}\|_1 + \lambda\|\mathbf{L}\|_F^2 \right\}$$



Background: The ELLA algorithm

- ELLA is the online version of GO-MTL (useful in lifelong learning scenarios)

Average the model losses across tasks

$$\min_{(\mathbf{L},\mathbf{s}^{(t)})} \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{N_t}\sum_{i=1}^{N_t}\mathcal{L}\left(D\left(\mathbf{x}_i^{(t)};\mathbf{L}\mathbf{s}^{(t)}\right),y_i^{(t)}\right)+\mu\left\|\mathbf{s}^{(t)}\right\|_1\right. + \lambda\|\mathbf{L}\|_F^2$$

First inefficiency: due to the explicit dependence of the above equation on all of the previous training data (through the inner summation)

- Solution: Approximate the equation using the second-order Taylor expansion of

$\frac{1}{N_t}\sum_{i=1}^{N_t}\mathcal{L}\left(D\left(\mathbf{x}_i^{(t)};\boldsymbol{w}^{(t)}\right),y_i^{(t)}\right)$ around $\mathbf{w}^{(t)}$

Second inefficiency: In order to evaluate a single candidate $\mathbf{L}$, an optimization problem must be solved to recompute the value of each of the $\mathbf{s}^{(t)}$'s

- Solution: Compute each $\mathbf{s}^{(t)}$ only when training data for task t are available and do not update it when new tasks arrive

Information Technologies Institute

Queen Mary University of London

InVID

ELLA_LC objective function

- Contributions:
  1. We add a new **label-based constraint** that considers concept correlations
  2. We solve the objective function of ELLA using **quadratic programming** instead of solving the Lasso problem
  3. We use linear **SVMs** as base learners instead of logistic regression

$$\min_{(\mathbf{L},\mathbf{s}^{(t)})} \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{N_t}\sum_{i=1}^{N_t}\mathcal{L}\left(D\left(\mathbf{x}_i^{(t)};\mathbf{L}\mathbf{s}^{(t)}\right),y_i^{(t)}\right)+\mu\left\|\mathbf{s}^{(t)}\right\|_1 \right. \quad (1)$$

$$\boxed{\left.+\beta\left(\sum_{\substack{t'=1\\t'\neq t}}^{T}\frac{1}{T-1}\left|\phi_{t,t'}\right|\left\|\mathbf{L}\left(\mathbf{s}^{(t)}-\operatorname{sign}(\phi_{t,t'})\mathbf{s}^{(t')}\right)\right\|^2\right)\right\}}+\lambda\|\mathbf{L}\|_F^2$$

extra term added to ELLA's objective function that considers concept correlations

ϕ-correlation coefficient between t and t'

If $\phi_{t,t'} > 0$ (positive correlation)

$$\phi_{t,t'}\left\|\mathbf{L}\mathbf{s}^{(t)}-\mathbf{L}\mathbf{s}^{(t')}\right\|^2$$

Otherwise (negative correlation)

$$-\phi_{t,t'}\left\|\mathbf{L}\mathbf{s}^{(t)}+\mathbf{L}\mathbf{s}^{(t')}\right\|^2$$

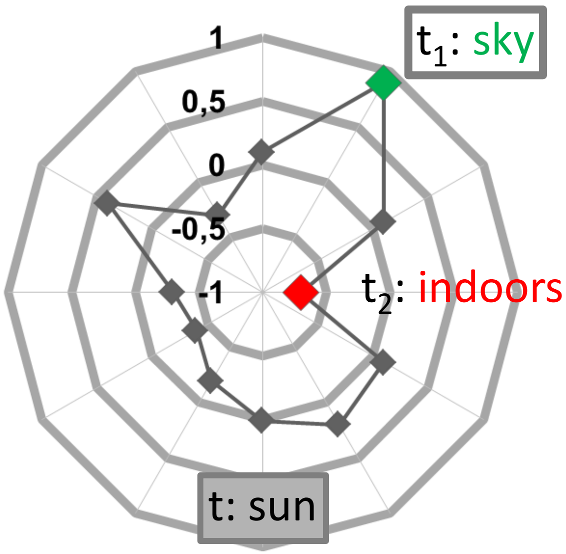
Information Technologies Institute

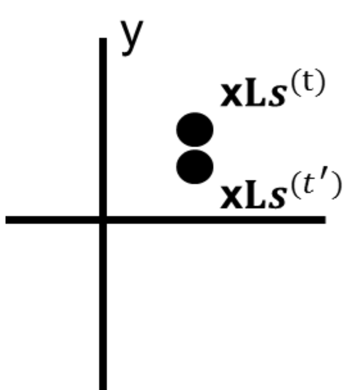
Queen Mary University of London

InVID

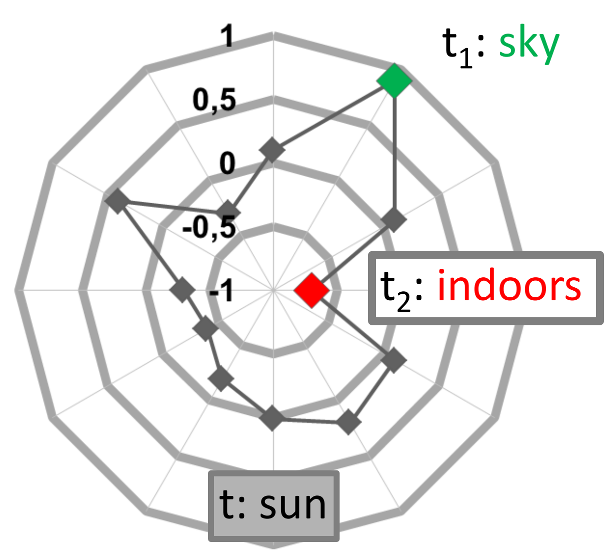
ELLA_LC label constraint

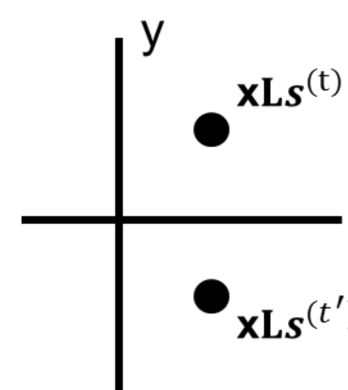
Positive correlation: force task parameters to be similar, linear classifiers return similar scores





Negative correlation: force task parameters to be opposite, linear classifiers return opposite scores



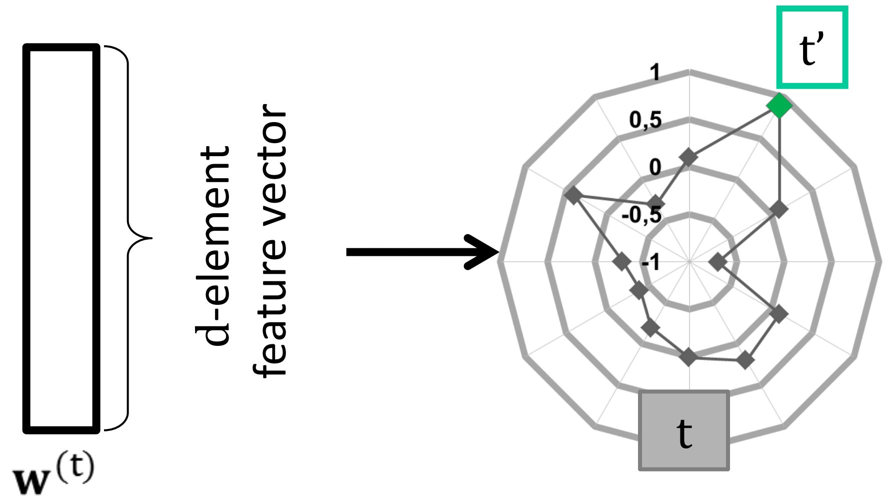


Correlation between **sun** and all the other concepts

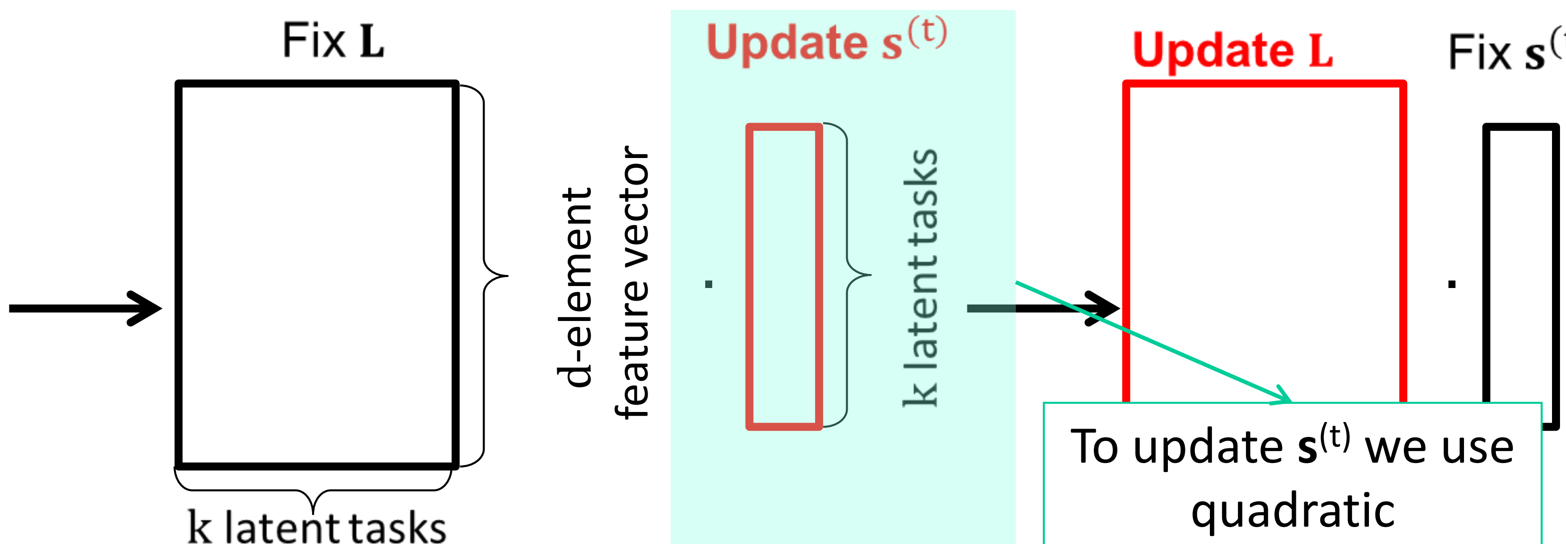
ELLA_LC solution

For each task arriving $(\mathbf{X}^{(t)}, \mathbf{y}^{(t)})$

Learn e.g., SVM



Compute the ϕ-correlation coefficient of the concept learned in task t with all the previously learned concepts



To update $\mathbf{s}^{(t)}$ we use quadratic programming

Information Technologies Institute

Queen Mary University of London

InVID

Experimental setup: Compared methods

Dataset: TRECVID SIN 2013

- 800 and 200 hours of internet archive videos for training and testing
- One keyframe per video shot
- Evaluated concepts: 38, Evaluation measure: MXinfAP

We experimented with 8 different feature sets

- The output from 4 different pre-trained ImageNet DCNNs (CaffeNet, ConvNet, GoogLeNet-1k, GoogLeNet-5k)
- The output from 4 fine-tuned networks on the TRECVID SIN dataset

Compared methods

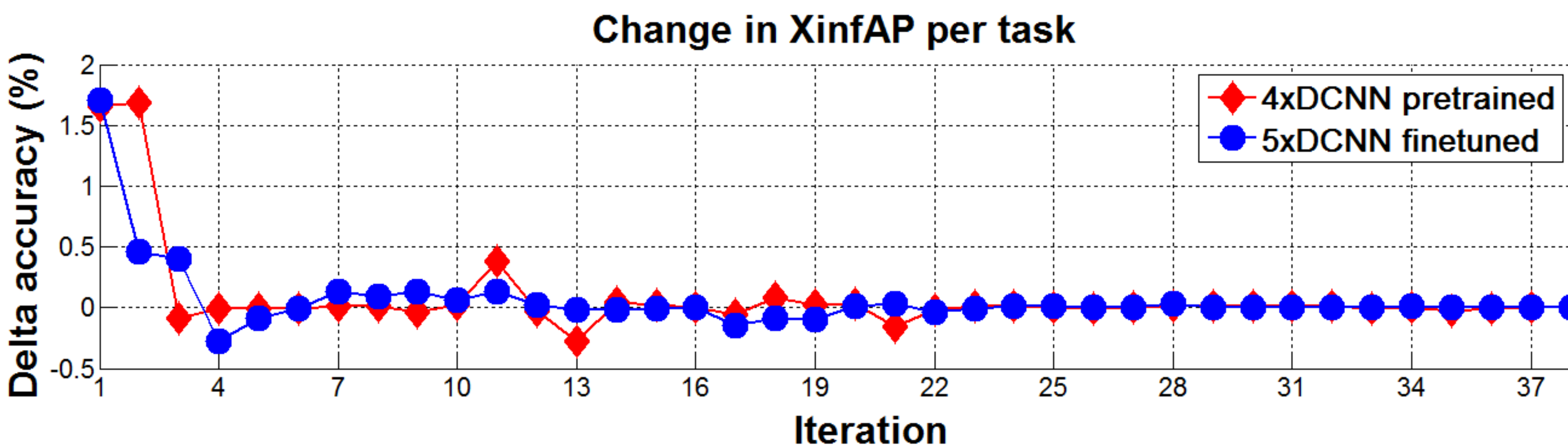
- STL using: a) LR, b) LSVM, c) kernel SVM with radial kernel (KSVM)
- The label powerset (LP) multi-label learning algorithm that models only label relations (Markatopoulou et al. 2014)
- AMTL (Sun et al. 2015) and CMTL (Zhou et al. 2011), two batch MTL methods
- ELLA (Eaton & Ruvolo 2013), an online MTL method (what we extend in this study)

Experimental results

			Single-task learning			Joint concept learning				Proposed multi-task learning			
R#	**Features**	**Direct output**	**LR**	**LSVM**	**KSVM**	**LP [9]**	**AMTL [17]**	**CMTL [16]**	**ELLA [2]**	**ELLA_QP LR**	**ELLA_QP LSVM**	**ELLA_LC LR**	**ELLA_LC LSVM**
	(i) Using the output of ImageNet-based networks as features												
1	CaffeNet1k	-	13.00	14.20	12.81	11.77	12.90	11.56	13.14	13.99	16.27 *	14.28	**16.36**
2	ConvNet1k	-	17.58	19.29	15.62	1608	17.58	16.09	17.88	18.45	21.02 *	18.94	**21.10**
3	GNET1k	-	16.10	17.73	14.17	15.00	16.34	14.43	15.79	17.07	19.86 *	17.48	**19.98**
4	GNET5k	-	20.89	22.68	20.73	20.54	21.01	19.99	15.65	21.88	24.05 *	22.16	**24.14**
5	**4xDCNN**	-	21.77	24.29	22.64	19.58	22.96	21.42	21.17	23.66	25.97 *	24.18	$\boxed{\textbf{26.10}}$
	(ii) Using the output of networks finetuned on different subsets of the TRECVID SIN 2013 training set as features												
6	CaffeNet1k-345	20.29	22.21	24.16	23.00	21.29	24.22	24.03	16.63	23.09	25.47 *	23.51	**25.88**
7	GNET1k-60	19.77	24.51	24.30	23.07	25.06 *	22.56	22.25	23.71	24.56	**26.05**	24.51	25.90 *
8	GNET1k-60	19.90	24.71	24.78	22.90	25.20 *	23.87	22.87	24.57	24.69	**26.24**	24.52	**26.24**
9	GNET1k-323	23.97	26.67	28.65	27.79	27.22	28.67	28.09	25.75	27.56	29.86	28.19	**30.23**
10	GNET5k-323	22.78	27.13	29.32	28.53	28.21	29.47	29.27	27.15	28.61	30.80 *	28.90	**31.01**
11	**5xDCNN FT**	25.35	28.56	30.60	29.93	30.27	30.94	30.15	28.19	29.89	31.82 *	30.32	$\boxed{\textbf{32.10}}$

- Results of our experiments in terms of MXinfAP
- ELLA_QP: an intermediate version of the proposed ELLA_LC that does not use the label constraint of ELLA_LC but uses quadratic programming
- Statistical significance from the best performing method using the paired t-test (at 5% significance level); the absence of * suggests statistical significance

Experimental results



- Change in XinfAP for each task between the iteration that the task was first learned and the last iteration (where all tasks had been learned), divided by the position of the task in the task sequence
- Reverse transfer occurred, i.e., a positive change in accuracy for a task indicates this, mainly for the tasks that were learned early
- As far as the pool of tasks increases early tasks get new knowledge from many more tasks, which explains why the benefit is bigger for them

Conclusions

- Proposed ELLA_LC: an online MTL method for video concept detection
- Learning the relations between many task models (one per concept) in combination with the concept correlations that can be captured from the ground-truth annotation outperforms other SoA single-task and multi-task learning approaches
- The proposed ELLA_QP and ELLA_LC perform better than the STL alternatives both when LR and when LSVM is used as the base learner
- The proposed ELLA_QP and ELLA_LC perform better than the MTL ELLA algorithm (the one that they extend) both when LR and when LSVM is used as the base learner
- Serving as input more complicated keyframe representations (e.g., combining many DCNNs instead of using a single DCNN) improves the accuracy of the proposed ELLA_QP and ELLA_LC
- Fine-tuning is a process that improves the retrieval accuracy of ELLA_QP and ELLA_LC

Thank you for your attention!
Questions?

More information and contact:
Dr. Vasileios Mezaris
bmezaris@iti.gr
http://www.iti.gr/~bmezaris