Online Multi-Task Learning based on K-SVD

Paul Ruvolo, Eric Eaton

Bryn Mawr College, Computer Science Department

June 13, 2014
Presented by Zhengming Xing

Outline

- Introduction
- Multi-Task learning using K-SVD
- Lifelong learning
- Experiment Results

Introduction

Objective and novelty of the paper

Learning and sharing knowledge between multiple tasks to improve the performance.

- K-SVD based algorithm to reduce computation.
- Online Multi-Task learning

Review of K-SVD

Given the input signal $\{\mathbf{x}_n\}_{n=1}^{N} \in \mathbb{R}^d$, obtain an overcomplete dictionary and sparse representation of the signal.

$$\underset{\mathbf{L}}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{\mathbf{s}^{(i)}} \{||\mathbf{L}\mathbf{s}^{(i)} - \mathbf{x}_i||_2^2 + \mu||\mathbf{s}^{(i)}||_0\} \tag{1}$$

Introduction

Solving equation 1 includes two optimization steps. Let us define $\mathbf{S} = [\mathbf{s}^{(1)} \ldots \mathbf{s}^{(n)}]$ and $\mathbf{X} = [\mathbf{x}_{(1)} \ldots \mathbf{x}_{(n)}]$

- Sparse coding stage
  Given a fixed value of $\mathbf{L}$, solve the following problem
  $$\mathbf{s}^{(i)} \leftarrow \text{argmin}_{\mathbf{s}} \{ ||\mathbf{L}\mathbf{s} - \mathbf{x}_i||_2^2 + \mu ||\mathbf{s}||_0 \}$$

- Codebook update stage.
  $$||\mathbf{X} - \mathbf{L}\mathbf{S}||_2^2 = ||\mathbf{X} - \sum_{k \neq j} \mathbf{l}_k \mathbf{s}_k^T - \mathbf{l}_j \mathbf{s}_j^T||_2^2 = ||\mathbf{E}_j - \mathbf{l}_j \mathbf{s}_j^T||_2^2$$
  Using SVD to solve the above problem
  $$(\mathbf{U}, \Sigma, \mathbf{V}) = \text{SVD}(\mathbf{E}_j^{\mathcal{A}})$$
  $$\mathbf{l}_j \leftarrow \mathbf{u}_1, \mathbf{s}_j^{(\mathcal{A})} \leftarrow \sigma_{1,1} \mathbf{v}_1$$
  where $\mathbf{E}_j = \mathbf{X} - \sum_{k \neq j} \mathbf{l}_k \mathbf{s}_k^T$, $\mathcal{A}$ indicates the columns corresponds to nonzero s_{ji}.

Multi-task learning using K-SVD

Let $\{\mathcal{Z}^{(1)}, \mathcal{Z}^{(2)}, ..., \mathcal{Z}^{(T)}\}$ denotes a set of supervised learning task.

- Each task $\mathcal{Z}^{(t)} = (\hat{f}^{(t)}, \mathbf{X}^{(t)}, \mathbf{y}^{(t)})$ is defined by a mapping $\hat{f}^{(t)}$ from instance space $\mathbf{X}^{(t)}$ to a set of label $\mathbf{y}^{(t)}$.
- Each task model $f^{(t)}(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}^{(t)})$ is specified by a task-specific parameter vector $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^d$.
- Eack task parameter vector $\boldsymbol{\theta}^{(t)} = \mathbf{L}\mathbf{s}^{(t)}$, $\mathbf{L} \in \mathbb{R}^{d \times k}$ is shared for all task and $\mathbf{s}^{(t)}$ is sparse.

The object function is as following

$$e_T(\mathbf{L}) = \sum_{t=1}^{T} \min_{s^{(t)}} \{ \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(f(\mathbf{x}_i^{(t)}; \mathbf{L}\mathbf{s}^{(t)}), y_i^t) + \lambda ||\mathbf{L}\mathbf{s}^{(t)}||_2^2 + \mu ||\mathbf{s}^{(t)}||_0 \} \tag{2}$$

where $\mathcal{L}$ is some loss function for fitting the task model.

Multi-task learning using K-SVD

Second order Taylor expansion of the objective function around $\theta^{(t)}$. where $\theta^{(t)}$ is an optimal predictor for task t only.

$$g_T(\mathbf{L}) = \sum_{t=1}^{T} \min_{\mathbf{s}^{(t)}} \{ \frac{1}{n_t} ||\theta^{(t)} - \mathbf{L}\mathbf{s}^{(t)}||^2_{\mathbf{D}^{(t)}} + \mu ||\mathbf{s}^{(t)}||_0 \}$$

$||\mathbf{v}||_{\mathbf{A}} = \mathbf{v}^T \mathbf{A} \mathbf{v}$ and

$$\mathbf{D}^{(t)} = \lambda \mathbf{I} + \bigtriangledown^2_{\theta,\theta} \frac{1}{2n_t} \sum_{i=1}^{n_t} \mathcal{L}(f(\mathbf{x}_i^{(t)}; \theta), y_i^{(t)}) \mid_{\theta=\theta^{(t)}}$$
$$\theta^{(t)} = \text{argmin}_\theta \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(f(\mathbf{x}_i^{(t)}, \theta), y_i^{(t)}) + \lambda ||\theta||_2^2$$

Multi-task learning using K-SVD

- Sparse coding step

$$\mathbf{s}^{(i)} \leftarrow \text{argmin}_{\mathbf{s}}\{||\mathbf{L}\mathbf{s} - \mathbf{x}_i||^2_{\mathbf{D}^{(t)}} + \mu||\mathbf{s}||_0\}$$

- Codebook update step
  Replace the original SVD with generalized SVD

$$(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}) = \text{GSVD}(\mathbf{E}^{\mathcal{A}_j}, \mathbf{M}, \mathbf{W})$$

  where $\mathbf{M} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} \mathbf{D}^{(t)}$, W is diagonal with $w_{t,t} = \frac{\mathbf{1}^T\mathbf{D}^{(\mathcal{A}_t)}\mathbf{1}}{\sum_{t'\in\mathcal{A}}\mathbf{1}^T\mathbf{D}^{(\mathcal{A}_t)}\mathbf{1}}$. In this way, the update values of $\mathbf{l}_j$ and $s_j^{\mathcal{A}}$ are obtained via minimizing

$$\sum_{t1=1}^{|\mathcal{A}|}\sum_{t2=1}^{|\mathcal{A}|} w_{t1,t2}(\mathbf{e}_j^{\mathcal{A}_{t1}} - \mathbf{l}_j\mathbf{s}_j^{\mathcal{A}_{t1}})^T\mathbf{M}(\mathbf{e}_j^{\mathcal{A}_{t2}} - \mathbf{l}_j\mathbf{s}_j^{\mathcal{A}_{t2}})$$

  Generalized SVD, the matrix $\mathbf{A}$ decompose to

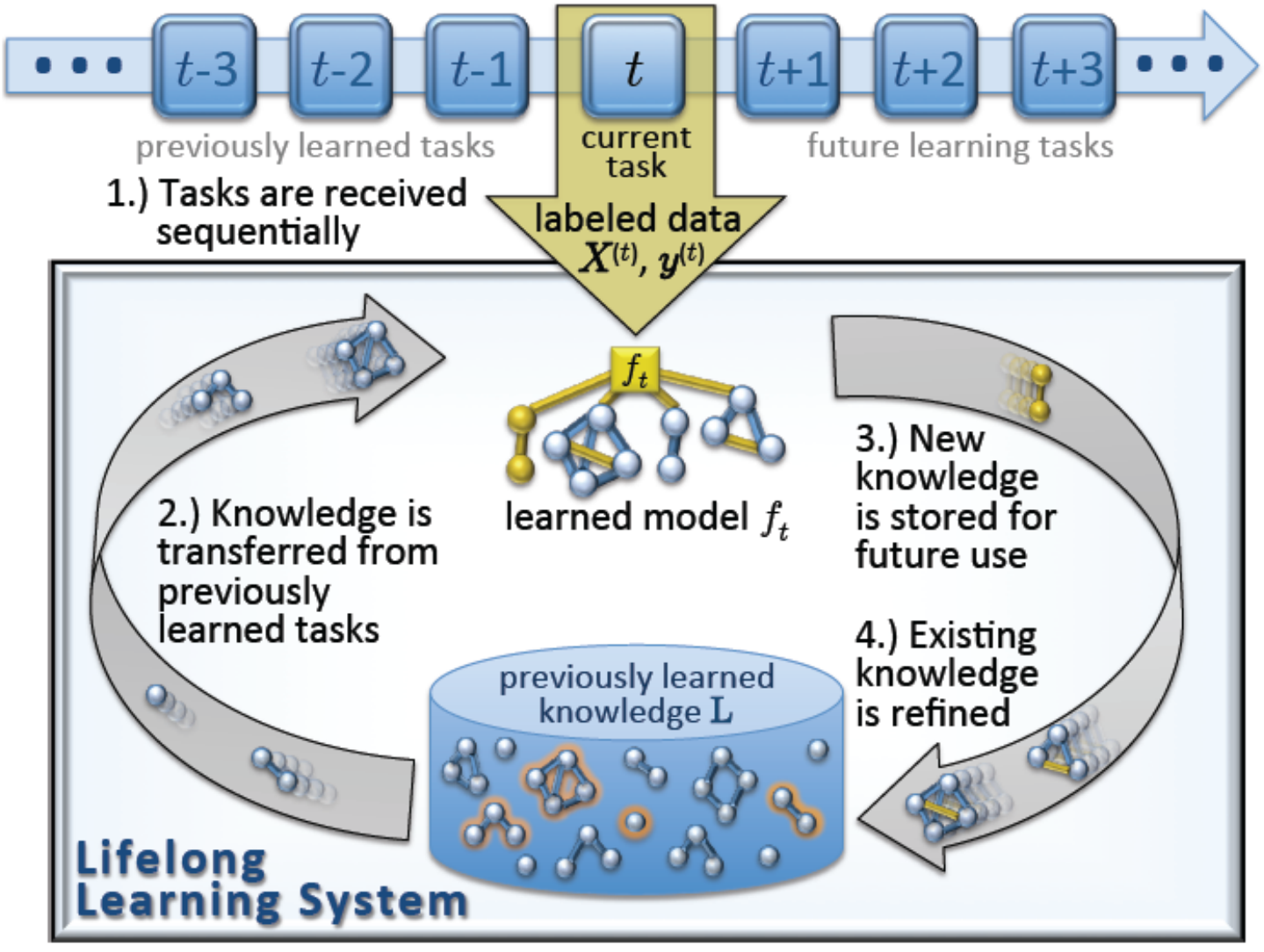
$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \text{ with: } \mathbf{U}^T\mathbf{M}\mathbf{U} = \mathbf{I} \text{ and } \mathbf{V}^T\mathbf{W}\mathbf{V} = \mathbf{I}$$

Algorithm

Algorithm 1 MTL-SVD

Input training data $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}), ..., (\mathbf{X}^{(T)}, \mathbf{y}^{(T)})$; dictionary size k
for $i = 1$ **to** T **do**
  $(\theta^{(i)}, \mathbf{D}^{(i)}) \leftarrow \text{singleTaskLeaner}(\mathbf{X}^{(i)}, \mathbf{y}^{(i)})$
end for
while convergence criteria not meet **do**
  for $t = 1$ **to** T **do**
    update $\mathbf{s}^{(t)}$
  end for
  for $j = 1$ **to** k **do**
    update $\mathbf{l}_j$ and $\mathbf{s}_j^{\mathcal{A}}$
  end for
end while

Lifelong learning problem



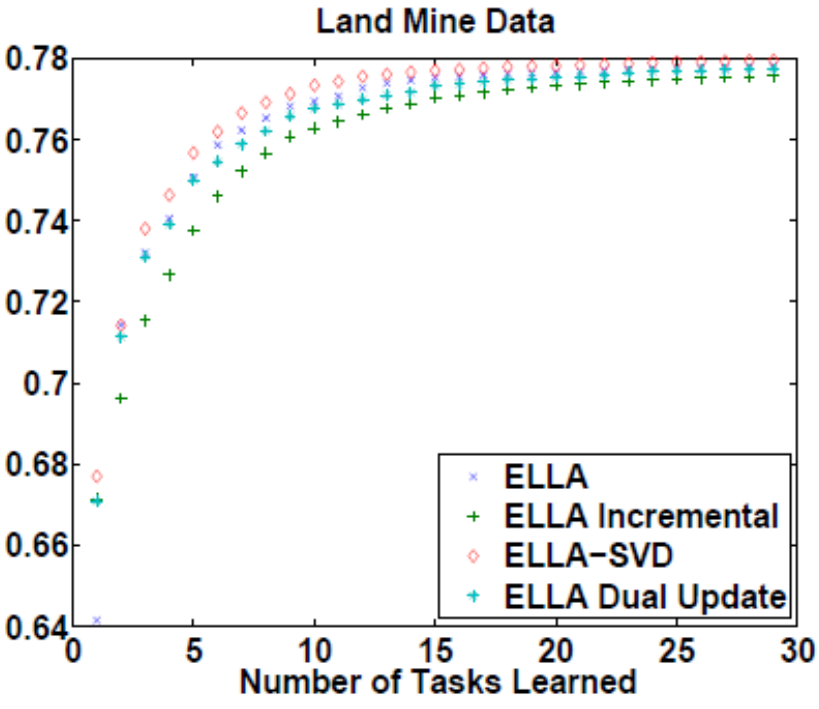
Lifelong learning problem

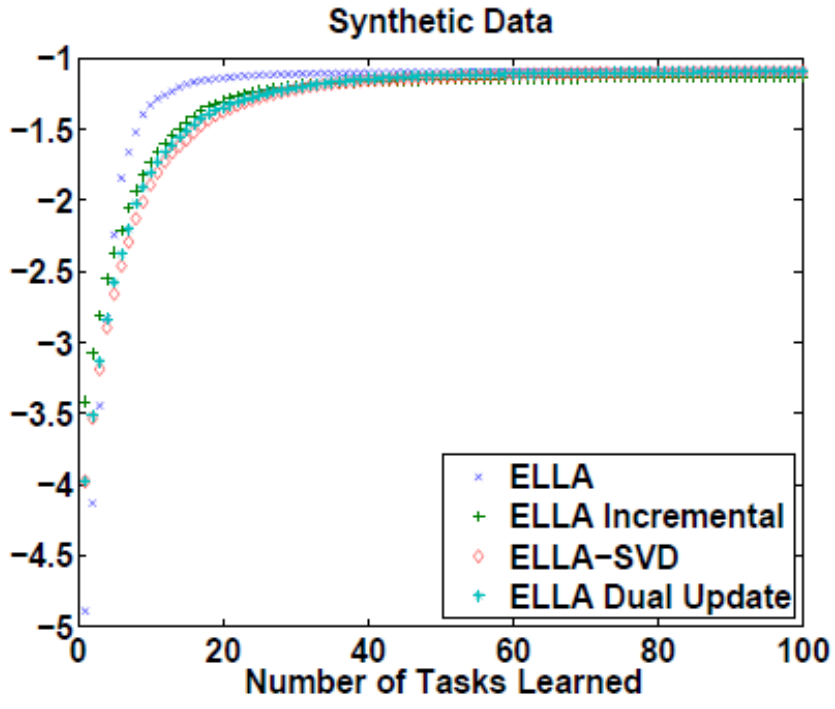
Adaptions to allow efficient lifelong learning

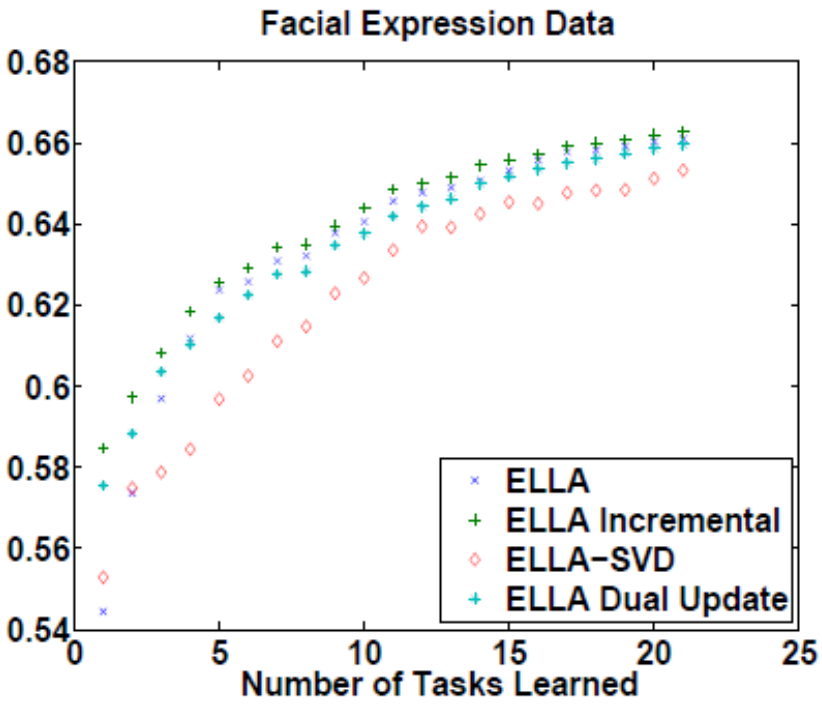
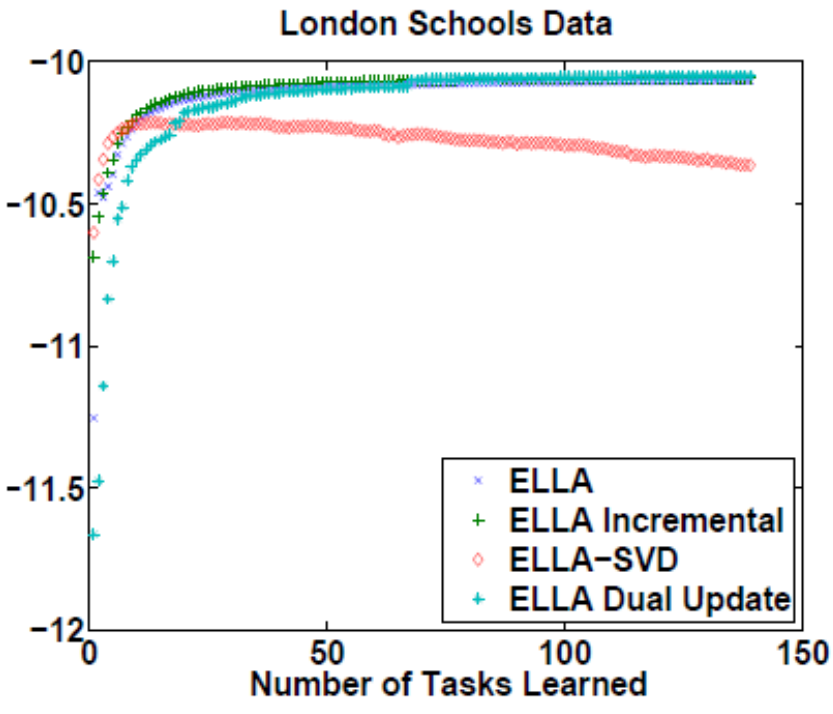
- update the value of $\mathbf{s}^{(t)}$ only based on the training data for task t.
- update subset of columns of $\mathbf{L}$ corresponding to non-zero entries of $\mathbf{s}^{(t)}$.
- the above two steps only performed once per patch.

Experiment Results

Synthetic Regression Tasks: 100 tasks with $d = 13$ and $n_t = 100$.
Land Mine Detection: 14820 data instances divided in to 29 different geographical regions.





Experiment Results

Facial Expression Recognition: 3 facial action units. 21 tasks and each with $450 - 999$ images.
London School Data: Exam score from 15362 students from 139 schools with $d = 27$.