ELLA: An Efficient Lifelong Machine Learning Algorithm

Paul Ruvolo       Eric Eaton

Bryn Mawr College Computer Science Department

This work was supported by ONR Grant #N00014-11-1-0139
- ELLA is a method for **online multi-task learning** in a lifelong learning setting

<table>
<thead>
<tr>
<th></th>
<th>Transfer Learning</th>
<th>Batch Multi-Task Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizes performance over</td>
<td>Target task</td>
<td>All tasks</td>
</tr>
<tr>
<td>Learns tasks consecutively</td>
<td>Yes, efficiently</td>
<td>Very inefficiently</td>
</tr>
<tr>
<td>Computational cost</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Lifelong learning includes elements of both transfer and multi-task learning
ELL: A method for online multi-task learning in a lifelong learning setting

ELL's Capabilities:
1. Learns tasks consecutively
2. Transfers knowledge from previous tasks
3. Optimizes performance over all tasks
4. Theoretical guarantees on performance and convergence

<table>
<thead>
<tr>
<th></th>
<th>Transfer Learning</th>
<th>Batch Multi-Task Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizes performance</td>
<td>Target task</td>
<td>All tasks</td>
</tr>
<tr>
<td></td>
<td>Learn tasks</td>
<td>Yes, efficiently</td>
</tr>
<tr>
<td></td>
<td>Consecutively</td>
<td>Very inefficiently</td>
</tr>
<tr>
<td>Computation cost</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Lifelong learning includes elements of both transfer and multi-task learning.
Overview

- ELLA is a method for **online multi-task learning** in a lifelong learning setting.

**ELLA’s Capabilities:**

1. Learns tasks consecutively
2. Transfers knowledge from previous tasks
3. Optimizes performance over all tasks
4. Theoretical guarantees on performance and convergence

<table>
<thead>
<tr>
<th></th>
<th>Transfer Learning</th>
<th>Batch Multi-Task Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizes performance</td>
<td>Target task</td>
<td>All tasks</td>
</tr>
<tr>
<td></td>
<td>Learn tasks</td>
<td>Very inefficiently</td>
</tr>
<tr>
<td></td>
<td>Computational cost</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
</tr>
</tbody>
</table>

Lifelong learning includes elements of both transfer and multi-task learning.

ELLA has equivalent accuracy to batch multi-task learning, but is over 1,000x faster and can learn online.
Lifelong Machine Learning

1.) Tasks are received sequentially

- previously learned tasks
- current task
- future learning tasks

\[ t-3 \quad t-2 \quad t-1 \quad t \quad t+1 \quad t+2 \quad t+3 \quad \ldots \]

previously learned knowledge \( L \)

Lifelong Learning System
Lifelong Machine Learning

1.) Tasks are received sequentially

- previously learned tasks
- current task
- future learning tasks

Labeled data: \( X^{(t)}, y^{(t)} \)

Previously learned knowledge: \( L \)

Lifelong Learning System
1.) Tasks are received sequentially.

2.) Knowledge is transferred from previously learned tasks.

Labeled data: \(X^{(t)}, y^{(t)}\)
1.) Tasks are received sequentially

2.) Knowledge is transferred from previously learned tasks

---

Lifelong Machine Learning

---

Lifelong Learning System
Lifelong Machine Learning

1.) Tasks are received sequentially

2.) Knowledge is transferred from previously learned tasks

3.) New knowledge is stored for future use

Labeled data $X^{(t)}, y^{(t)}$

Learned model $f_t$

Previously learned knowledge $L$

Lifelong Learning System
Lifelong Machine Learning

1.) Tasks are received sequentially
2.) Knowledge is transferred from previously learned tasks
3.) New knowledge is stored for future use
4.) Existing knowledge is refined

previously learned tasks


t-3  t-2  t-1  t  t+1  t+2  t+3

future learning tasks

labeled data \( X^{(t)}, y^{(t)} \)

learned model \( f_t \)

previously learned knowledge \( L \)

Lifelong Learning System
ELLLA fits a parametric model for each task $t$

$$f^{(t)}(x) = f(x; \theta^{(t)}) \quad \theta^{(t)} \in \mathbb{R}^d$$

The parameters $\theta^{(t)}$ are linear combinations of a shared basis $L$

$$\theta^{(t)} = Ls^{(t)} \quad L \in \mathbb{R}^{d \times k}, \ s^{(t)} \in \mathbb{R}^k$$
ELL A fits a parametric model for each task $t$

$$f^{(t)}(x) = f(x; \theta^{(t)}), \quad \theta^{(t)} \in \mathbb{R}^d$$

The parameters $\theta^{(t)}$ are linear combinations of a shared basis $L$

$$\theta^{(t)} = Ls^{(t)}, \quad L \in \mathbb{R}^{d \times k}, \quad s^{(t)} \in \mathbb{R}^k$$

**Objective Function:**

$$e_T(L) = \frac{1}{T} \sum_{t=1}^{T} \min_{s^{(t)}} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L} \left( f \left( x_i^{(t)}; Ls^{(t)} \right), y_i^{(t)} \right) + \mu \| s^{(t)} \|_1 \right\} + \lambda \| L \|_F^2$$

- #tasks seen so far
- model fit to data
- sparsity
- complexity
Efficient Lifelong Learning

Objective Function:

\[ e_T(L) = \frac{1}{T} \sum_{t=1}^{T} \min_{s^{(t)}} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L} \left( f \left( x_i^{(t)}; Ls^{(t)} \right), y_i^{(t)} \right) + \mu \| s^{(t)} \|_1 \right\} + \lambda \| L \|_F^2 \]

**Problem 1:** The complexity of the inner summation scales linearly with the number of training instances

**Our solution:** Replace the model-fit-to-data term with the second-order Taylor expansion around the optimal single task model:

\[ g_T(L) = \frac{1}{T} \sum_{t=1}^{T} \min_{s^{(t)}} \left\{ \| \theta^{(t)} - Ls^{(t)} \|_{D^{(t)}}^2 + \mu \| s^{(t)} \|_1 \right\} + \lambda \| L \|_F^2 \]

where, \( \theta^{(t)} = \arg \min_{\theta} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L} \left( f \left( x_i^{(t)}; \theta \right), y_i^{(t)} \right) \)

\( D^{(t)} \) is \( \frac{1}{2} \) the Hessian of the single-task loss evaluated at \( \theta^{(t)} \)

\[ \| x \|_{D}^2 = x^T Dx \]
Efficient Lifelong Learning

**Objective Function:**

\[
g_T(L) = \frac{1}{T} \sum_{t=1}^{T} \min_{s^{(t)}} \left\{ \| \theta^{(t)} - Ls^{(t)}\|_{D^{(t)}}^2 + \mu \| s^{(t)} \|_1 \right\} + \lambda \| L \|_F^2
\]

**Problem 2:** The complexity of the outer summation grows linearly with the number of tasks \( T \)

**Our solution:** Optimize \( s^{(t)} \) only when training on task \( t \) and not on any other tasks

- We prove that the penalty for not re-optimizing the other \( s^{(t)} \)'s vanishes as \( T \) gets large
**Efficient Lifelong Learning Algorithm**

MTL Objective Function:

\[
e_T(L) = \frac{1}{T} \sum_{t=1}^{T} \min_{s(t)} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} L(f(x_i^{(t)}; Ls^{(t)}), y_i^{(t)}) + \mu \|s^{(t)}\|_1 \right\} + \lambda \|L\|_F^2
\]

**ELLA:** Given a new task \( t \),

1. Train a single-task model \( \theta^{(t)} \) for task \( t \)
2. Reconstruct \( \theta^{(t)} \) in the current basis (LASSO)
   \[
s^{(t)} \leftarrow \arg \min_{s^{(t)}} \ell(L_m, s^{(t)}, \theta^{(t)}, D^{(t)})
\]
3. Update the basis
   \[
   L_{m+1} \leftarrow \arg \min_L \lambda \|L\|_F^2 + \frac{1}{T} \sum_{t=1}^{T} \ell(L, s^{(t)}, \theta^{(t)}, D^{(t)})
   \]

in practice, \( L \) is constructed incrementally with each task

where \[ \ell(L, s, \theta, D) = \mu \|s\|_1 + \|\theta - Ls\|_D^2 \]

\( D^{(t)} \) is \( \frac{1}{2} \) the Hessian of the single-task loss evaluated at \( \theta^{(t)} \)

\[ \|x\|_D^2 = x^T Dx \]
ELLÀ’s per-task computational complexity is:

1. Independent of the number of tasks $T$
2. Independent of the numbers of training instances for previous tasks

We show a variety of theoretical guarantees on ELLÀ’s performance and convergence

Online dictionary learning for sparse coding

[Mairal et al ICML’09] is a special case of ELLÀ
Facial Expression Recognition: identify presence of facial action units (#5 upper lid raiser, #10 upper lip raiser, #12 lip corner pull)

21 Classification Tasks:
• 7 subjects
• 3 action units
• 450-999 images each

2,880 Gabor Features:
• 2 spatial scales
• 4 orientations
• 576 locations

PCA
100 features + bias

ELLA

Models
Applications

**Facial Expression Recognition:** identify presence of facial action units (#5 upper lid raiser, #10 upper lip raiser, #12 lip corner pull)

- 21 Classification Tasks
- 2,880 Gabor Features
- **PCA** 100 features + bias
- **ELLA**
- **29 ClassificaMon Tasks**
- 2,880 Gabor Features

**Land Mine Detection** from radar images [Xue et al. 2007]

- 29 Classification Tasks:
  - 29 regions
  - 2 terrain types
  - 14,820 instances total

**Exam Score Prediction** for London schools [Kumar et al. 2012]

- 139 Regression Tasks:
  - 139 schools
  - 15,362 students total
  - 4 school-specific features
  - 3 student-specific features
  - Exam year + bias term
## Empirical Results

ELLA achieves nearly identical accuracy to batch MTL:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Problem Type</th>
<th>Batch MTL Accuracy</th>
<th>ELLA Relative Accuracy</th>
<th>OMTL Relative Accuracy</th>
<th>STL Relative Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land Mine</td>
<td>Classification</td>
<td>0.7802 ± 0.013 (AUC)</td>
<td>99.73 ± 0.7%</td>
<td>82.2 ± 3.0%</td>
<td>97.97 ± 1.5%</td>
</tr>
<tr>
<td>Facial Expr.</td>
<td>Classification</td>
<td>0.6577 ± 0.021 (AUC)</td>
<td>99.37 ± 3.1%</td>
<td>97.58 ± 3.8%</td>
<td>97.34 ± 3.9%</td>
</tr>
<tr>
<td>Syn. Data</td>
<td>Regression</td>
<td>−1.084 ± 0.006 (-rMSE)</td>
<td>97.74 ± 2.7%</td>
<td>N/A</td>
<td>92.91 ± 1.5%</td>
</tr>
<tr>
<td>London Sch.</td>
<td>Regression</td>
<td>−10.10 ± 0.066 (-rMSE)</td>
<td>98.90 ± 1.5%</td>
<td>N/A</td>
<td>97.20 ± 0.4%</td>
</tr>
</tbody>
</table>

Batch MTL = [Kumar & Daumé III, ICML’12]  
OMTL = [Saha et al, AISTATS’11]
ELL A achieves nearly identical accuracy to batch MTL:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Problem Type</th>
<th>Batch MTL Accuracy</th>
<th>ELLA Relative Accuracy</th>
<th>OMTL Relative Accuracy</th>
<th>STL Relative Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land Mine</td>
<td>Classification</td>
<td>0.7802 ± 0.013 (AUC)</td>
<td>99.73 ± 0.7%</td>
<td>82.2 ± 3.0%</td>
<td>97.97 ± 1.5%</td>
</tr>
<tr>
<td>Facial Expr.</td>
<td>Classification</td>
<td>0.6577 ± 0.021 (AUC)</td>
<td>99.37 ± 3.1%</td>
<td>97.58 ± 3.8%</td>
<td>97.34 ± 3.9%</td>
</tr>
<tr>
<td>Syn. Data</td>
<td>Regression</td>
<td>−1.084 ± 0.006 (-rMSE)</td>
<td>97.74 ± 2.7%</td>
<td>N/A</td>
<td>92.91 ± 1.5%</td>
</tr>
<tr>
<td>London Sch.</td>
<td>Regression</td>
<td>−10.10 ± 0.066 (-rMSE)</td>
<td>98.90 ± 1.5%</td>
<td>N/A</td>
<td>97.20 ± 0.4%</td>
</tr>
</tbody>
</table>

While obtaining speedups of:
- over 1,000x for learning all tasks

Batch MTL = [Kumar & Daumé III, ICML’12]  
OMTL = [Saha et al, AISTATS’11]
ELL A achieves nearly identical accuracy to batch MTL:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Problem Type</th>
<th>Batch MTL Accuracy</th>
<th>ELLA Relative Accuracy</th>
<th>OMTL Relative Accuracy</th>
<th>STL Relative Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land Mine</td>
<td>Classification</td>
<td>0.7802 ± 0.013 (AUC)</td>
<td>99.73 ± 0.7%</td>
<td>82.2 ± 3.0%</td>
<td>97.97 ± 1.5%</td>
</tr>
<tr>
<td>Facial Expr.</td>
<td>Classification</td>
<td>0.6577 ± 0.021 (AUC)</td>
<td>99.37 ± 3.1%</td>
<td>97.58 ± 3.8%</td>
<td>97.34 ± 3.9%</td>
</tr>
<tr>
<td>Syn. Data</td>
<td>Regression</td>
<td>−1.084 ± 0.006 (-rMSE)</td>
<td>97.74 ± 2.7%</td>
<td>N/A</td>
<td>92.91 ± 1.5%</td>
</tr>
<tr>
<td>London Sch.</td>
<td>Regression</td>
<td>−10.10 ± 0.066 (-rMSE)</td>
<td>98.90 ± 1.5%</td>
<td>N/A</td>
<td>97.20 ± 0.4%</td>
</tr>
</tbody>
</table>

While obtaining speedups of:

- over 1,000x for learning all tasks
- over 38,000x for learning each new task

**Batch MTL** = [Kumar & Daumé III, ICML’12]  
**OMTL** = [Saha et al, AISTATS’11]
Earlier task models improve from later learning without retraining on the earlier tasks.
ELLA: An Efficient Lifelong Learning Algorithm
Paul Ruvolo & Eric Eaton

Thank you!

Code for ELLA is available at cs.brynmawr.edu/~eeaton

ELLA has equivalent accuracy to batch multi-task learning, but is over 1,000x faster and can learn online.