ELLA: An Efficient Lifelong Learning Algorithm

Paul Ruvolo  
Bryn Mawr College

Eric Eaton  
Bryn Mawr College

Abstract

The problem of learning multiple tasks that arrive sequentially, known as lifelong learning, is of great importance to the creation of intelligent, general-purpose, and flexible machines. This paper develops a method for online multitask learning in the lifelong learning setting. The proposed Efficient Lifelong Learning Algorithm (ELLA) maintains a sparsely shared basis for all task models, transfers knowledge from the basis to learn each new task, and refines the basis over time to maximize performance across all learned tasks. The proposed method has strong connections to both online dictionary learning for sparse coding and current batch multi-task learning methods, and provides robust theoretical performance guarantees. Empirically, ELLA yields nearly identical performance to batch multitask learning while learning tasks sequentially in over 1,000x less time.

Introduction

Goal: Develop intelligent agents that
1. Quickly learn new tasks
2. Learn continually with experience
3. Exhibit versatility over multiple tasks

ELLA’s Capabilities:
1. Optimized performance over all tasks
2. Efficient learning of each new consecutive task via transfer
3. Equivalent performance to batch MTL with over 1,000x speedup

Lifelong Learning Framework

The problem of learning multiple tasks that arrive sequentially, known as lifelong learning, is of great importance to the creation of intelligent, general-purpose, and flexible machines. This paper develops a method for online multitask learning in the lifelong learning setting. The proposed Efficient Lifelong Learning Algorithm (ELLA) maintains a sparsely shared basis for all task models, transfers knowledge from the basis to learn each new task, and refines the basis over time to maximize performance across all learned tasks. The proposed method has strong connections to both online dictionary learning for sparse coding and current batch multi-task learning methods, and provides robust theoretical performance guarantees. Empirically, ELLA yields nearly identical performance to batch multitask learning while learning tasks sequentially in over 1,000x less time.

Task Structure Model

ELLA’s goal is to fit a parametric model for each task \( f(t^i) = f(x; \theta^{(t)}) \) \( \theta^{(t)} \in \mathbb{R}^d \)

The parameter vectors for each function are assumed to be linear combinations of a shared latent basis \( \mathbf{L} \)

\[ \theta^{(t)} = \mathbf{L} \mathbf{s}^{(t)} \quad \mathbf{L} \in \mathbb{R}^{d \times k}, \mathbf{s}^{(t)} \in \mathbb{R}^k \]

We minimize the following objective function to encourage models to utilize few latent basis vectors:

\[ e_T (\mathbf{L}) = \frac{1}{n_T} \sum_{t=1}^{n_T} \frac{1}{n_i} \sum_{i=1}^{n_i} \mathcal{L} (f(x_i^t; \mathbf{L} \mathbf{s}^{(t)}), y_i^t) + \mu \| \mathbf{s}^{(t)} \|_2 + \lambda \| \mathbf{L} \|_F^2 \]

Efficient Lifelong Learning

Minimizing \( e_T \) is computationally expensive for two reasons:
1. Evaluating the objective function scales with the number of training instances \( n_i \)
2. The number of optimization problems grows linearly with the number of tasks \( T \)

To address (1) we replace the inner optimization problem with the 2nd-order Taylor expansion around the optimal task-specific model: \( \theta^{(t)} = \arg \min_{\theta^{(t)}} \sum_{i=1}^{n_i} \mathcal{L} (f(x_i^t; \theta^{(t)}), y_i^t) \)

To address (2) we optimize \( s^{(t)} \) only when training on task \( t \) and not on other tasks

These simplifications yield the following update equations to learn given \( (X^t, y^t) \):

\[ s^{(t)} \leftarrow \arg \min_{\mathbf{s}^{(t)}} e_L (\mathbf{L}_0, \mathbf{s}^{(t)}, \mathbf{D}^{(t)}) \]

\[ \mathbf{L}_{m+1} \leftarrow \arg \min_{\mathbf{L}} \frac{1}{2} \| \mathbf{L} \|_F^2 + \sum_{t=1}^{T} e_L (\mathbf{L}, \mathbf{s}^{(t)}, \theta^{(t)}, \mathbf{D}^{(t)}) \]

where

\[ e_L (\mathbf{L}, \mathbf{s}, \theta, \mathbf{D}) = \mu \| \mathbf{s} \|_1 + \| \theta - \mathbf{L} \mathbf{s} \|_2^2 \]

\[ \mathbf{D}^{(t)} = \frac{1}{2} \mathbf{H} \text{ is the Hessian of the single-task loss evaluated at } \theta^{(t)} \]

Base Learning Algorithms

ELLA can support any base learner with a twice-differentiable loss function

Linear Regression: \( (y^t) \in \mathbb{R}^{n_t}, f(x; \theta) = \theta^T x, \text{ and } \mathcal{L}(\cdot) \text{ is squared loss} \)

\[ \theta^{(t)} = \left( X^T X^{(t)} \right)^{-1} X^T y^{(t)} \]

Logistic Regression: \( (y^t) \in [-1,1]^n, f(x; \theta) = \left( 1 + e^{-\theta^T x} \right)^{-1}, \text{ and } \mathcal{L}(\cdot) \text{ is log-loss} \)

\[ \theta^{(t)} = \frac{1}{2m} \sum_{i=1}^{m} f(x_i^t, \theta^{(t)}) \left( 1 - f(x_i^t, \theta^{(t)}) \right) x_i^T \]

Theory

Assumptions:
1. Tuples \((D^1, \theta^{(1)})\) are drawn iid from a distribution with compact support
2. The sparse coding solution is unique and is sensitive to changes in \( s^{(t)} \) (non-zero entries of \( s^{(t)} \) \( \forall T, D^t, \theta^{(t)} \) and \( \theta^{(t)} \) the smallest eigenvalue of \( L^T L \) \( L \geq 0 \) over all tasks

Theorems:
1. The basis \( L \) becomes more stable over time: \( L_{t+1} - L_t = O (\frac{1}{t}) \)
2. The penalty for not re-optimizing the \( s^{(t)} \) vanishes as \( T \) gets large:

\[ g_T (\mathbf{L}) = \lambda \| \mathbf{L} \|_F^2 + \sum_{t=1}^{T} \min_s (\mathbf{L} s, \mathbf{D}^{(t)}) \]

\[ g_T (\mathbf{L}) = \lambda \| \mathbf{L} \|_F^2 + \sum_{t=1}^{T} \min_s (\mathbf{L} s, \mathbf{D}^{(t)}) \]

as \( T \to \infty, g_T (\mathbf{L}) \to g_T (\mathbf{L}) \) converges a.s. to 0

3. The basis \( L \) converges to a fixed point of the expected loss \( g_T \)

Connections to Dictionary Learning for Sparse Coding:
Online dictionary learning for sparse coding (Mairal et al., 2009) is a special case of ELLA where the \( \theta^{(t)} \)’s are given instead of learned and the \( D^t \)’s are identity matrices

Results

Facial Expression Recognition:
identify presence of facial action units (85 upper lid raiser, #10 upper lid raiser, #12 lip corner pull)

Land Mine Detection from radar images

Student Exam Score Prediction

ELLA achieves nearly identical accuracy to batch MTL,

ELLA also exhibits reverse transfer

Acknowledgement:
This research was supported by ONR grant #N00014-11-1-0139