Lifelong Spectral Clustering

Gan Sun,[1,3*] **Yang Cong,**[2] **Qianqian Wang,**[3] **Jun Li,**[4] **Yun Fu**[3]
[1]University of Chinese Academy of Sciences, China. [†]
[2]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, China.
[3]Northeastern University, USA. [4]MIT, USA.
{sungan1412, congyang81, junl.mldl}@gmail.com, qianqian174@foxmail.com, yunfu@ece.neu.edu

Abstract

In the past decades, spectral clustering (SC) has become one of the most effective clustering algorithms. However, most previous studies focus on spectral clustering tasks with a fixed task set, which cannot incorporate with a new spectral clustering task without accessing to previously learned tasks. In this paper, we aim to explore the problem of spectral clustering in a lifelong machine learning framework, *i.e.,* Lifelong Spectral Clustering (L^2SC). Its goal is to efficiently learn a model for a new spectral clustering task by selectively transferring previously accumulated experience from knowledge library. Specifically, the knowledge library of L^2SC contains two components: 1) orthogonal basis library: capturing latent cluster centers among the clusters in each pair of tasks; 2) feature embedding library: embedding the feature manifold information shared among multiple related tasks. As a new spectral clustering task arrives, L^2SC firstly transfers knowledge from both basis library and feature library to obtain encoding matrix, and further redefines the library base over time to maximize performance across all the clustering tasks. Meanwhile, a general online update formulation is derived to alternatively update the basis library and feature library. Finally, the empirical experiments on several real-world benchmark datasets demonstrate that our L^2SC model can effectively improve the clustering performance when comparing with other state-of-the-art spectral clustering algorithms.

Introduction

Spectral clustering algorithms (Ng, Jordan, and Weiss 2002; Shi and Malik 2000) discover the corresponding embedding of data via utilizing manifold information embedded in the sample distribution, which has shown the state-of-the-art performance in many applications (Li and Chen 2015; Zhao, Ding, and Fu 2017; Wang, Ding, and Fu 2019). In addition to single spectral clustering task scenario, (Yang et al. 2015) proposes a multi-task spectral clustering model, and aims to perform multiple clustering tasks and make them reinforce each other. However, most recently-proposed models (Zhang et al. 2018; Pang et al. 2018; Kang et al. 2018) focus on clustering tasks with a fixed task set. When applied into a new task environment or incorporated into a new spectral clustering task, these models have to repeatedly access to previous clustering tasks, which can result in high energy consumption in real applications, *e.g.,* in mobile applications. In this paper, our work explores how to adopt the spectral clustering scenario into the setting of lifelong machine learning.

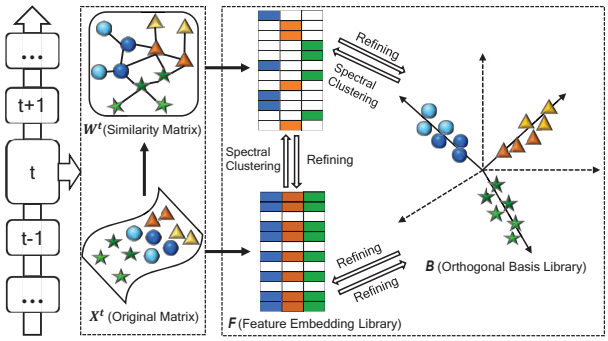


Figure 1: The demonstration of our lifelong spectral clustering model, where different shapes are from different clusters. When a new clustering task X^t is coming, the knowledge is iteratively transferred from orthogonal basis library B and feature embedding library F to encode the new task.

For the lifelong machine learning, recent works (Ruvolo and Eaton 2013; Isele, Rostami, and Eaton 2016; Xu et al. 2018; Sun et al. 2018a; 2019) have explored the methods of accumulating the single task over time. Generally, lifelong learning utilizes knowledge from previously learned tasks to improve the performance on new tasks, and accumulates a knowledge library over time. Although these models have been successfully adopted into supervised learning (Chen, Ma, and Liu 2018; Sun, Cong, and Xu 2018) and reinforcement learning (Ammar et al. 2014; Isele and Cosgun 2018), its application in spectral clustering, one of the most classical research problems in machine learning community, is still sparse. *Take the news clustering tasks as an example, the semantic meaning of* Artificial Intelligence *and* NBA *are very dissimilar in the newspaper of*

*The corresponding author is Gan Sun.
†This work has been done during Gan Sun visiting Northeastern University, and he was supported by NSFC (61722311).

year 2010, and should be divided into different clusters. The clustering task of year 2010 can thus contribute to the clustering task of year 2020 in a never-ending perspective, since the correlation information between Artificial Intelligence *and* NBA *of year 2020 is similar with that in year 2010.*

Inspired by the above scenario, this paper aims to establish a lifelong learning system with spectral clustering tasks, *i.e.,* lifelong spectral clustering. Generally, the main challenges among multiple consecutive clustering tasks are as follows: 1) **Cluster Space Correlation:** the latent cluster space should be consistent among multiple clustering tasks. For example, for the news clustering task, the cluster centers in year 2010 can be {Business, Technology, Science, etc}, while the cluster centers in year 2020 are similar to that in year 2010; 2) **Feature Embedding Correlation:** another correlation among different clustering tasks is feature correlation. For example, in consecutive news cluster tasks, the semantic meaning of *Artificial Intelligence* are very similar in year 2010 and year 2020. Thus, the feature embedding of *Artificial Intelligence* should be same for these two tasks.

To tackle the challenges above, as shown in Figure 1, we propose a Lifelong Spectral Clustering (*i.e.,* $\mathrm{L^2SC}$) model by integrating cluster space and feature embedding correlations, which can achieve never-ending knowledge transfer between previous clustering tasks and later ones. To achieve this, we present two knowledge libraries to preserve the common information among multiple clustering tasks, *i.e.,* orthogonal basis and feature embedding libraries. Specifically, 1) orthogonal basis library contains a set of latent cluster centers, *i.e.,* each sample of cluster tasks can be effectively assigned to multiple clusters with different weights; 2) feature embedding library can be modeled by introducing bipartite graph co-clustering, which can not only discover the shared manifold information among cluster tasks, but also maintain the data manifold information of each individual task. When a new spectral clustering task is coming, $\mathrm{L^2SC}$ can firstly encode the new task via transferring the knowledge of both orthogonal basis library and feature embedding library to encode the new task. Accordingly, these two libraries can be refined over time to keep on improving across all clustering tasks. For model optimisation, we derive a general lifelong learning formulation, and further optimize this optimization problem via applying an alternating direction strategy. Finally, we evaluate our proposed model against several spectral clustering algorithms and even multi-task clustering models on several datasets. The experimental results strongly support our proposed $\mathrm{L^2SC}$ model.

The novelties of our proposed $\mathrm{L^2SC}$ model include:

- To our best knowledge, this work is the first attempt to study the problem of spectral clustering in the lifelong learning setting, *i.e.,* Lifelong Spectral Clustering ($\mathrm{L^2SC}$), which can adopt previously accumulated experience to incorporate new cluster tasks, and improve the clustering performance accordingly.
- We present two common knowledge libraries: orthogonal basis library and feature embedding libray, which can simultaneously preserve the latent clustering centers and capture the feature correlations among different clustering tasks, respectively.
- We propose an alternating direction optimization algorithm to optimize the proposed $\mathrm{L^2SC}$ model efficiently, which can incorporate fresh knowledge gradually from online dictionary learning perspective. Various experiments show the superiorities of our proposed model in terms of effectiveness and efficiency.

Related Work

In this section, we briefly provide a review on two topics: **Multi-task Clustering** and **Lifelong Learning**.

For the **Multi-task Clustering** (Zhang et al. 2018), the learning paradigm is to combine multi-task learning (Sun et al. 2017) with unsupervised learning, and the key issue is how to transfer useful knowledge among different clustering tasks to improve the performance. Based on this assumption, recently-proposed methods (Zhang et al. 2017; Huy et al. 2013) achieve knowledge transfer for clustering via using some sample from other tasks to form better distance metrics or K-nn graphs. However, these methods ignore employing the task relationships in the knowledge transfer process. To preserve task relationships, multi-task Bregman clustering (MBC) (Zhang and Zhang 2011) captures the task relationships by alternatively update clusters among different tasks. For the spectral clustering based multi-task clustering, multi-task spectral clustering (MTSC) (Yang et al. 2015) take the first attempt to extend spectral clustering into multi-task learning. By using the inter-task and intra-task correlations, a $\ell_{2,p}$-norm regularizer is adopted in MTSC to constrain the coherence of all the tasks based on the assumption that a low-dimensional representation is shared by related tasks. Then a mapping function is learned to predict cluster labels for each individual task.

For the **Lifelong Learning**, the early works on this topic focus on transferring the selective information from task cluster to the new tasks (Thrun and O'Sullivan 1996; Sun et al. 2018b), or transferring invariance knowledge in neural networks (Thrun 2012). In contrast, an efficient lifelong learning algorithm (ELLA) (Ruvolo and Eaton 2013) is developed for online learning multiple tasks in the setting of lifelong learning. By assuming that models of all related tasks share a common basis, each new task can be obtained by transferring knowledge from the basis. Furthermore, (Ammar et al. 2014) extends this idea into learn decision making tasks consecutively, and achieves dramatically accelerate learning on a variety of dynamical systems; (Isele, Rostami, and Eaton 2016) proposes a coupled dictionary to incorporate task descriptors into lifelong learning, which can enable performing zero-shot transfer learning. Since observed tasks in lifelong learning system may not compose an *i.i.d* samples, learning an inductive bias in form of a transfer procedure is proposed in (Pentina and Lampert 2015). Different from traditional learning models (Rannen Ep Triki et al. 2017), (Li and Hoiem 2016) proposes a learning without forgetting method for convolutional neural network, which can train the network only using the data of the new task, and retain performance on original tasks via knowledge distillation (Hinton, Vinyals, and Dean 2015), and train the network

using only the data of the new task. Among the discussion above, there is no works concerning lifelong learning in the spectral clustering setting, and our current work represents the first work to achieve lifelong spectral clustering.

Lifelong Spectral Clustering ($\mathrm{L^2SC}$)

This section introduces our proposed lifelong spectral clustering ($\mathrm{L^2SC}$) learning model. Firstly, we briefly review a general spectral clustering formulation for single spectral clustering task. Our $\mathrm{L^2SC}$ model for lifelong spectral clustering task problem is then given.

Revisit Spectral Clustering Algorithm

This subsection reviews a general spectral clustering algorithm with normalized cut. Given an undirected similarity graph $G^t = \{X^t, W^t\}$ with a vertex set $X^t \in \mathbb{R}^{d\times n_t}$ and an corresponding affinity matrix $W^t \in \mathbb{R}^{n_t\times n_t}$ for the clustering task t, where d is the number of the features, n_t is the total number of data samples for the task t, each element w_{ij}^t in symmetric matrix W^t denotes the similarity between a pair of vertices (x_i^t, x_j^t). The common choice for matrix W^t can be defined as follows:

$$w_{ij}^t = \begin{cases} \exp\Big(-\frac{\|x_i^t - x_j^t\|^2}{2\sigma^2}\Big), & \text{if } x_i^t \in \mathcal{N}(x_j^t) \text{ or } x_j^t \in \mathcal{N}(x_i^t) \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{N}(\cdot)$ is the function for searching K-nearest neighbors, and σ controls the spread of the neighbors. After applying the normalized Laplacian:

$$K^t = (D^t)^{-\frac{1}{2}} L^t (D^t)^{-\frac{1}{2}} = I - (D^t)^{-\frac{1}{2}} W^t (D^t)^{-\frac{1}{2}}, \quad (1)$$

where D^t is a diagonal matrix with the diagonal elements as $D_{ii}^t = \sum_j W_{ij}^t \, (\forall i)$. The final formulation of spectral clustering turns out to be the well-known normalized cut (Shi and Malik 2000), and can be expressed as:

$$\max_{F^t} \ \mathrm{tr}((F^t)^\top K^t F^t), \quad s.t., (F^t)^\top F^t = I_k, \quad (2)$$

where the optimal cluster assignment matrix F^t can be achieved via the eigenvalue decomposition of matrix K^t. Based on the relaxed continuous solution, then the final discrete solution of F^t can be obtained by spectral rotation or K-means, *e.g.,* the j-th element of f_i^t is 1, if the sample x_i^t is assigned to the j-th cluster; 0, otherwise.

Problem Statement

Given a set of m unsupervised clustering tasks $\mathcal{T}^1, \ldots, \mathcal{T}^m$, where each individual clustering task $\mathcal{T}^t$ has a set of n_t training data samples $X^t \in \mathbb{R}^{d\times n_t}$, and the dimensionality of feature space is d. The original intention of multi-task spectral clustering method (Yang et al. 2015) is to uncover the correlations among all the clustering tasks, and predict the cluster assignment matrices $\{F^t\}_{t=1}^m$ for each clustering task. However, learning incremental spectral clustering tasks without accessing to the previously-adopted clustering data is not considered in traditional single or multi-task spectral clustering models. In the setting of spectral clustering, a lifelong spectral clustering system encounters a series of spectral clustering tasks $\mathcal{T}^1, \ldots, \mathcal{T}^m$, where each task $\mathcal{T}^t$ is defined in Eq. (2), and intends to obtain new cluster assignment matrix $F^t \in \mathbb{R}^{n_t\times k}$ for the task t. For convenience, this paper assume that the learner in this lifelong machine learning system do not know any information about clustering tasks, *e.g.,* the task distributions, the total number of spectral clustering tasks m, etc. When lifelong spectral clustering system receives a batch of data for some spectral clustering task t (either a new spectral clustering task or previously learning task t) in each period, this system should obtain cluster assignment matrix of samples of encountered tasks. The goal is to obtain corresponding task assignment matrices $F^1, \ldots, F^m$ such that: **1)** Clustering Performance: each obtained assignment matrix F^t should preserve the data configuration of the t-th task, and partition the new clustering task more accurate; **2)** Computational Speed: in each clustering period, obtaining each F^t should be faster than that among traditional multi-task spectral clustering methods; **3)** Lifelong Learning: new F^t's can be arbitrarily and efficiently added when the lifelong clustering system faces with new unsupervised spectral clustering tasks.

The Proposed $\mathrm{L^2SC}$ Model

In this section, we introduce how to model the lifelong learning property and cross-task correlations simultaneously. Basically, there are two challenges in the $\mathrm{L^2SC}$ model:

1) Orthogonal Basis Library: in order to achieve lifelong learning, one of the major component is how to store the previously accumulated experiences, *i.e.,* knowledge library. To tackle this issue, inspired by (Han and Kim 2015) which employs the orthogonal basis clustering to uncover the latent cluster centers, each assignment matrix F^t can be decomposed into two submatrices, *i.e.,* a basis matrix $B \in \mathbb{R}^{k\times k}$ called orthogonal basis library, and a cluster encoding matrix $E^t \in \mathbb{R}^{n_t\times k}$, as $F^t = E^t B$. Then the multi-task spectral clustering formulation can be expressed as:

$$\begin{aligned} \max_{\{E^t\}_{t=1}^m} \frac{1}{m}\sum_{t=1}^m \mathrm{tr}((E^t B)^\top K^t E^t B), \\ s.t., B^\top B = I_k, (E^t)^\top E^t = I_k, \forall t = 1, \ldots, m, \end{aligned} \quad (3)$$

where the orthogonal constraint of matrix B encourages each column of B to be independent, and K^t is defined in the Eq. (1). Therefore, the orthogonal basis library B can be used to refine the latent cluster centers and further obtain an excellent cluster separation.

2) Feature Embedding Library: even though the latent cluster centers can be captured gradually in Eq. (3), it does not consider the common feature embedding transfer across multiple spectral clustering tasks. Motivated by (Jiang and Chung 2012) which adopts graph based co-clustering to control and achieve the knowledge transfer between two tasks, we propose to link each pair of clustering tasks together such that one embedding obtained in one task can facilitate the discover of the embedding in another task. We thus define an invariant feature embedding library $L \in \mathbb{R}^{d\times k}$ with group sparse constraint, and give the graph co-clustering term as:

$$\max_{L} \frac{1}{m}\sum_{t=1}^m \mathrm{tr}(L^\top \hat{X}^t E^t B) + \mu\|L\|_{2,1}, s.t., L^\top L = I_k, \quad (4)$$

Algorithm 1 Lifelong Spectral Clustering (L^2SC) Model

Input: Spectral clustering tasks: $X^1, \ldots, X^m$, Library: $B \leftarrow \mathbf{0}_{k\times k}$, $L \leftarrow \mathbf{0}_{d\times k}$, $\mu \geq 0, \lambda_t \geq 0, \forall t = 1, \ldots, m$, Statistical records: $M_0 \leftarrow \mathbf{0}_{k\times k}$, $C_0 \leftarrow \mathbf{0}_{d\times k}$;
while Receive clustering task data **do**
  New t-th task: (X^t, t);
  Construct matrices $\{K^t, \hat{X}^t\}$;
  while Not Converge **do**
    Update E^t via Eq. (7);
    Update B via Eq. (10);
    Update L via Eq. (14);
    Update Θ via $\Theta_{ii} = \frac{1}{2\|l_i\|_2}, (\forall i = 1, \ldots, d)$;
  end while
  Compute cluster assignment matrices via $\{E^t B\}_{t=1}^m$;
  Compute final indicator matrices via K-means;
end while

and $\hat{X}^t$ for the t-th task is defined as:

$$\hat{X}^t = (D_1^t)^{-\frac{1}{2}} X^t (D_2^t)^{-\frac{1}{2}}, \tag{5}$$

where $D_1^t = \mathrm{diag}(X^t\mathbf{1})$, and $D_2^t = \mathrm{diag}((X^t)^\top\mathbf{1})$. Intuitively, with this sharing embedding library L, multiple spectral clustering tasks can transfer embedding knowledge with each other in a perspective of common feature learning (Argyriou, Evgeniou, and Pontil 2008).

Given the same graph construction method and training data for each spectral clustering task, we solve the optimal cluster assignment matrix $\{F^t\}_{t=1}^m$ while encouraging each clustering task to share common knowledge in libraries B and L. By combining these two goals in Eq. (3) and Eq. (4), then lifelong spectral clustering model can be expressed as the following objective function:

$$\begin{aligned}\max_{B,L,\{E^t\}_{t=1}^m} \frac{1}{m}\sum_{t=1}^m \Big\{ &\mathrm{tr}((E^tB)^\top K^t E^t B) \\ &+ \lambda_t \mathrm{tr}(L^T \hat{X}^t E^t B)\Big\} + \mu \|L\|_{2,1}, \\ s.t.,\ B^\top B = I_k, &L^\top L = I_k, (E^t)^\top E^t = I_k,\end{aligned} \tag{6}$$

where λ_t's are the trade-off between the each spectral clustering task with the co-clustering objective. If λ_t's are set as 0, this model can reduce to the multi-task spectral clustering model with common cluster centers.

Model Optimization

This section shows how to optimize our proposed L^2SC model. Normally, standard alternating direction strategy using all the learned tasks is inefficient to this lifelong learning model in Eq. (6). Our goal in this paper is to build an lifelong clustering algorithm that both CPU time and memory space have lower computational cost than offline manner. When a new spectral clustering task m arrives, the basic ideas for optimizing Eq. (6) is: both L, B and E^m should be updated without accessing to the previously learned tasks, *e.g.,* the previous data in matrices $\{K^t, \hat{X}^t\}_{t=1}^{m-1}$. In the following, we briefly introduce the proposed update rules, and provide the convergence analysis in the experiment.

Updating E^m with fixed L and B: With the fixed L and B, the problem for solving encoding matrix E^m can be expressed as:

$$\max_{(E^m)^\top E^m = I_k} \mathrm{tr}((E^mB)^\top K^m E^m B) + \lambda_m \mathrm{tr}(L^\top \hat{X}^m E^m B). \tag{7}$$

With the orthonormality constraint, E^m can be updated in the setting of Stiefel manifold (Manton 2002), which is defined by the following Proposition.

Proposition 1. *Let $X \in \mathbb{R}^{n\times k}$ be a rank p matrix, where the singular value decomposition (*i.e.*, SVD) of X is $U\Sigma V^\top$. The projection of matrix X on Stiefel manifold is defined as:*

$$\pi(X) = \arg\min_{Q^\top Q = I} \|X - Q\|_F^2. \tag{8}$$

The projection could be calculated as: $\pi(X) = UI_{n,k}V^\top$.

Therefore, we can update E^m by moving it in the direction of increasing the value of the objective function, and the update operator can be given as:

$$E^m = \pi(E^m + \eta_m \nabla f(E^m)), \tag{9}$$

where η_m is the step size, $f(E^m)$ is the objective function of Eq. (7), and $\nabla f(E^m)$ can be defined as $2(K^m)^\top E^m BB^\top + \lambda_m(\hat{X}^m)^\top LB^\top$. To guarantee the convergence of the optimization problem in Eq. (7), we provide a convergence analysis at the experiment section.

Updating B with fixed L and $\{E^t\}_{t=1}^m$: With the obtained encoding matrix E^m for the new coming m-th task, the optimization problem for variable B can be:

$$\max_{B^\top B = I_k} \frac{1}{m}\sum_{t=1}^m \mathrm{tr}(B^\top (E^t)^\top K^t E^t B) + \lambda_t \mathrm{tr}(L^\top \hat{X}^t E^t B). \tag{10}$$

Based on the orthonormality constraint $B^\top B = I_k$, we can rewrite Eq. (10) as follows:

$$\begin{aligned}&\max_{B^\top B = I_k} \frac{1}{m}\sum_{t=1}^m \mathrm{tr}(B^\top((E^t)^\top K^t E^t + \lambda_t BL^\top \hat{X}^t E^t)B), \\ \Leftrightarrow &\max_{B^\top B = I_k} \mathrm{tr}(B^\top(\frac{1}{m}\sum_{t=1}^m (E^t)^\top K^t E^t + \frac{1}{m}\sum_{t=1}^m \lambda_t BL^\top \hat{X}^t E^t)B)\end{aligned} \tag{11}$$

To better store the previous knowledge of learned clustering tasks, we then introduce two statistical variables:

$$M_m = M_{m-1} + (E^m)^\top K^m E^m, C_m = C_{m-1} + \lambda_m \hat{X}^m E^m, \tag{12}$$

where $M_{m-1} = \sum_{t=1}^{m-1}(E^t)^\top K^t E^t$, and $C_{m-1} = \sum_{t=1}^{m-1}\lambda_t \hat{X}^t E^t$. Therefore, knowledge of new task is $(E^m)^\top K^m E^m$ and $\hat{X}^m E^m$. With B as a warm start, so:

$$B = \arg\max_{B^\top B = I_k} \mathrm{tr}(B^\top(M_m/m + BL^\top C_m/m)B). \tag{13}$$

It is well-known that the solution of B can be relaxedly obtained by the eigen-decomposition of $(M_m/m +$

$BL^\top C_m/m$). Notice that even though the input parameter of Eq. (13) contains B, the above solution is also effective since the proposed algorithm converges very quickly in the online manner.

Updating L with fixed B and $\{E^t\}_{t=1}^m$: With the obtained center library B and encoding matrix E^m for the new coming m-th task, the optimization problem for variable L can be denoted as:

$$\max_{L^\top L=I_k} \frac{1}{m}\sum_{t=1}^{m}\lambda_t \mathrm{tr}(L^\top \hat{X}^t E^t B) + \mu \left\|L\right\|_{2,1}, \tag{14}$$

and the equivalent optimization problem can be formulated as following equations:

$$\begin{aligned}
&\min_{L^\top L=I_k} -\mathrm{tr}(L^\top(\frac{1}{m}\sum_{t=1}^{m}\lambda_t \hat{X}^t E^t)B + \mu\Theta L)),\\
\Leftrightarrow &\min_{L^\top L=I_k} \left\| L - ((\frac{1}{m}\sum_{t=1}^{m}\lambda_t \hat{X}^t E^t)B + \mu\Theta^{-1}L)\right\|_F^2,\\
\Leftrightarrow &\min_{L^\top L=I_k} \left\| L - (C_m B + \mu\Theta^{-1}L)\right\|_F^2,
\end{aligned} \tag{15}$$

which is also definition of projection of $(C_m B + \mu\Theta L)$ on the Stiefel manifold. Further, Θ denotes a diagonal matrix with each diagonal element as: $\Theta_{ii} = \frac{1}{2\|l_i\|_2}$ (Nie et al. 2010), where l_i is the i-th row of L.

Finally, the cluster assignment matrices for all learned tasks can be computed via $\{E^t B\}_{t=1}^m$, and final indicator matrices are obtained using K-means. The whole optimization procedure is summarized in **Algorithm 1**.

Experiments

This section evaluates the clustering performance of our proposed $\mathrm{L^2SC}$ model via several empirical comparisons. We firstly introduce the used competing models. Several adopted datasets and experimental results are then provided, followed by some analyses of our model.

Comparison Models and Evaluation

The experiments in this subsection evaluate our proposed $\mathrm{L^2SC}$ model with three single spectral clustering models, and five multi-task clustering models.

Single spectral clustering models: 1) Spectral Clustering (stSC) (Ng, Jordan, and Weiss 2002): standard spectral clustering model; 2) Spectral clustering-union (uSC) (Ng, Jordan, and Weiss 2002): spectral clustering model, which can be achieved via collecting all the clustering task data (*i.e.,* "pooling" all the task data and ignoring the multi-task setting); 3) One-step spectral clustering (OnestepSC) (Zhu et al. 2017): single spectral clustering task model.

Multi-task clustering models: 1) Multi-task Bregman Clustering (MBC) (Zhang and Zhang 2011): this model consists of average Bregman divergence and a task regularization; 2) Smart Multi-task Bregman Clustering (SMBC) (Zhang, Zhang, and Liu 2015): unsupervised transfer learning model, which focuses on clustering a small collection of target unlabeled data with the help of auxiliary unlabeled data; 3) Smart Multi-task Kernel Clustering (SMKC) (Zhang, Zhang, and Liu 2015): this model can deal with nonlinear data by introducing Mercer kernel; 4) Multi-Task Spectral Clustering (MTSC) (Yang et al. 2015): this model performs spectral clustering over multiple related tasks by using their inter-task correlations; 5) Multi-Task Clustering with Model Relation Learning (MTCMRL) (Zhang et al. 2018): this model can automatically learn the model parameter relatedness between each pair of tasks.

For the evaluation, we adopt three performance measures: normalized mutual information (NMI), clustering purity (Purity) and rand index (RI) (Schütze, Manning, and Raghavan 2008) to evaluate the clustering performance. The bigger the value of NMI, Purity and RI is, the better the clustering performance of the corresponding model will be. We implement all the models in MATLAB, and all the used parameters of the models are tuned in $\{10^{-3}\times i\}_{i=1}^{10}\cup\{10^{-2}\times i\}_{i=2}^{10}\cup\{10^{-1}\times i\}_{i=2}^{10}\cup\{2\times i\}_{i=1}^{10}\cup\{40\times i\}_{i=1}^{20}$. Although different λ_t's are allowed for different tasks in our model, this paper we only differentiate between μ and $\lambda = \lambda_t > 0$.

Real Datasets & Experiment Results

According to whether the number of cluster center is consistent or not, there are two different scenarios for multi-task clustering tasks: **Cluster-consistent** and **Cluster-inconsistent**. For the **Cluster-consistent** dataset, it can be roughly divided into: same clustering task and different clustering tasks with same number of cluster centers. We thus use two datasets in this paper: WebKB4[1] with 2500 dimensions and Reuters[2] with 6370 dimensions, respectively. For the WebKB4 dataset, which includes web pages collected from computer science department websites at 4 universities: Cornell, Texas, Washington and Wisconsin, and 7 categories. Following the setting in (Zhang et al. 2018), 4 most populous categories (*i.e.,* course, faculty, project and student) are chosen for clustering. Accordingly, for the Reuters dataset, 4 most populous root categories (*i.e.,* economic index, energy, food and metal) are chosen for clustering, and the total number of task is 3. For the **Cluster-inconsistent** dataset, we also adopt 20NewsGroups[3] dataset with 3000 dimensions by following (Zhang et al. 2018), which consists of the news documents under 20 categories. Since "negative transfer" (Zhou and Zhao 2015) will happen when the cluster centers of multiple consecutive spectral tasks have significant changes, 4 most populous root categories (*i.e.,* comp, rec, sci and talk) are selected for clustering, while the 1-th and 3-th tasks are set as 3 categories, and the 2-th and 4-th tasks are set as 4 categories.

The experimental results (competing models with parameter setting are averaged over 10 random repetitions) are provided in Table 1, Table 2 and Table 3, where the task sequence for our $\mathrm{L^2SC}$ is in a random way. From the presented

[1] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo20/www/data/

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

[3] http://qwone.com/ jason/20Newsgroups/

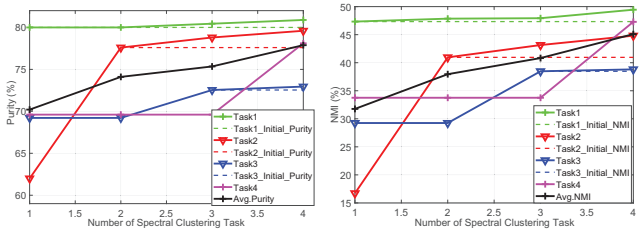
Table 1: Comparison results in terms of 3 different metrics (mean ± standard deviation) on WebKB4 dataset.

	Metrics	stSC	uSC	OnestepSC	MBC	SMBC	SMKC	MTSC	MTCMRL	Ours
Task1	Purity(%)	62.66±0.00	59.78±0.31	66.89±0.63	63.95±4.07	64.62±4.05	60.59±3.70	65.92±0.68	74.40±1.16	**80.00±1.25**
	NMI(%)	13.95±0.00	13.15±1.68	14.56±3.44	26.44±3.73	25.53±2.74	14.14±4.38	25.73±0.98	38.71±1.47	**49.07±1.41**
	RI(%)	59.89±0.00	58.83±0.04	64.76±1.06	61.64±3.58	62.58±2.65	59.45±1.62	62.85±0.76	73.47±0.64	**79.05±3.67**
Task2	Purity(%)	62.00±0.00	67.00 ±0.28	68.40±0.02	68.12±1.81	68.06±0.92	60.73±2.56	69.00±0.84	72.08±2.19	**74.40±1.13**
	NMI(%)	16.72±0.00	20.28 ±1.81	20.56±2.39	27.22±3.92	27.02±3.61	13.58±3.52	26.57±1.63	33.42±3.25	**41.89±1.49**
	RI(%)	57.12±0.00	60.38±2.06	64.81±1.52	68.04±2.46	68.32±3.29	58.31± 1.19	66.57±0.85	69.94±1.72	**74.79±0.13**
Task3	Purity(%)	69.21±0.27	59.80±0.27	69.80±0.55	64.86±5.36	68.04±2.28	66.01±4.13	68.23±0.55	**76.47±3.15**	74.12±1.10
	NMI(%)	29.24±0.30	15.60±2.42	22.55±2.36	26.50±3.97	28.32±3.86	22.09±5.95	29.33±0.99	40.97±5.26	**44.69±3.68**
	RI(%)	66.57±0.19	61.84±0.60	66.16±0.22	65.86±4.09	67.34±3.23	65.02±2.41	65.56±0.87	76.34±4.85	**78.53±1.97**
Task4	Purity(%)	69.61±0.00	70.42±0.23	71.31±0.92	72.18±4.17	71.21±4.08	69.82±2.58	69.93±0.46	78.23±2.68	**80.06±0.18**
	NMI(%)	33.75±0.00	33.15±0.49	36.84±0.59	39.97±5.24	39.53±2.74	30.31±4.17	45.64±0.66	49.23±2.17	**49.26±0.79**
	RI(%)	66.93±0.00	67.50±0.54	68.69±0.94	70.27±3.59	70.29±2.65	67.62±1.85	60.72±1.15	**79.01±1.54**	77.94±0.97
	Avg.Purity(%)	65.87±0.07	64.25±0.27	69.10±0.53	67.28±3.85	67.98±2.83	64.29±3.24	68.27±0.64	75.19±2.25	**77.14±0.92**
	Avg.NMI(%)	23.42±0.07	20.55±1.60	23.63±2.19	30.03±4.22	30.10±4.05	20.03±4.50	31.82±1.07	40.58±3.04	**46.26±1.84**
	Avg.RI(%)	62.63±0.05	62.14±0.81	66.11±0.94	66.45±3.43	70.29±2.65	62.60±1.76	63.93±0.91	74.69±2.19	**77.58±1.68**

Table 2: Comparison results in terms of 3 different metrics (mean ± standard deviation) on Reuters dataset.

	Metrics	stSC	uSC	OnestepSC	MBC	SMBC	SMKC	MTSC	MTCMRL	Ours
Task1	Purity(%)	95.63±0.00	85.44±0.00	94.66±0.00	73.30±9.27	89.90±1.40	95.75±0.72	97.57±0.00	97.57±0.00	**98.06±0.00**
	NMI(%)	82.72±0.00	60.54±0.00	75.89±1.52	61.39±2.32	77.92±3.31	84.17±2.05	89.49±0.00	89.49±0.00	**91.19±0.00**
	RI(%)	94.64±0.00	82.22±0.00	91.44±1.06	73.83±7.26	88.35±1.77	94.35±0.88	96.83±0.00	96.83±0.00	**97.43±0.00**
Task2	Purity(%)	84.62±0.00	70.00±0.00	86.92±0.00	70.19±0.73	92.88±0.38	90.96±1.15	96.15±0.54	97.31±0.54	**98.23±0.09**
	NMI(%)	62.91±0.00	53.17±0.00	64.45±0.00	53.43±7.81	79.54±1.27	75.76±2.65	84.89±1.62	88.93±2.46	**91.70±1.01**
	RI(%)	80.83±0.00	75.95±0.00	82.52±0.00	71.77±1.08	90.44±0.44	88.12±1.35	95.07±0.55	96.41±0.77	**98.11±0.05**
Task3	Purity(%)	75.26±0.00	82.63±0.00	76.05±1.86	72.36±9.78	75.24±2.98	76.50±2.07	90.79±0.37	94.21±0.00	**95.26±0.74**
	NMI(%)	54.00±0.00	59.85±0.00	61.74±1.44	46.35±6.70	54.11±5.41	52.72±2.79	73.37±0.66	**79.45±0.00**	78.62±0.47
	RI(%)	70.14±0.00	78.01±0.00	74.64±1.54	74.34±3.64	70.01±4.33	72.73±2.89	88.33±0.49	**93.13±0.00**	93.07±0.51
	Avg.Purity(%)	85.17±0.00	79.36±0.00	85.88±0.62	71.95±6.59	86.01±1.59	87.74±1.32	94.96±0.46	96.36±0.18	**97.18±0.74**
	Avg.NMI(%)	66.54±0.00	79.36±0.18	67.35±0.99	53.72±5.61	70.52±3.33	70.88±2.50	83.63±1.14	85.96±0.82	**87.71±0.47**
	Avg.RI(%)	81.87±0.00	78.73±0.90	82.87±0.87	73.31±7.33	82.93±2.18	85.07±1.71	93.54±0.52	95.45±0.26	**96.23±0.50**

results, we can notice that: **1)** Our proposed lifelong spectral clustering model outperforms the single-task spectral clustering methods, since L^{2}SC can exploit the information among multiple related tasks, whereas the single-task spectral clustering model only use the information within each task. MTCMRL performs worse than our proposed L^{2}SC in most cases, because even though it incorporates the cross-task relatedness with the linear regression model, it does not consider the feature embedding correlations among each pair of clustering tasks. The reason why MTCMRL performs better than our L^{2}SC in Task1 of 20NewsGroups is that we set $k = 4$ in this **Cluster-inconsistent** dataset, whereas the number of cluster center is 3 in Task1. **2)** In addition to MTCMRL and single-task spectral clustering models, our L^{2}SC performs much better than the comparable multi-task clustering model cases. It is because that L^{2}SC can not only learn the latent cluster center between each pair of tasks via the orthogonal basis library B, but also control the number of embedded features common across the clustering tasks. **3)** Additionally, Table 4 also shows that the runtime comparisons between our L^{2}SC model and other single/multi-task clustering models. L^{2}SC is faster and better than the most multi-task clustering models on WebKB, Reuters and 20NewsGroups datasets, *e.g.,* SMBC and MTSC, also OnestepSC. However, L^{2}SC is little slower than stSC and uSC. This is because both stSC and uSC can obtain the cluster assignment matrix via closed-form solution, *i.e.,* eigenvalue decomposition of the K^t in Eq. (2). We perform all the experiments on the computer with Intel i7 CPU, 8G RAM.



(a) WebKB4 Dataset (b) WebKB4 Dataset

Figure 2: The influence of the number of learned tasks on WebKB4 datasets in terms of Purity and NMI metrics, where the vertical and horizontal axes denote the clustering performance and number of learned tasks, respectively. The initial clustering performance of each task (except for the first task) of each dataset is achieved using stSC algorithm.

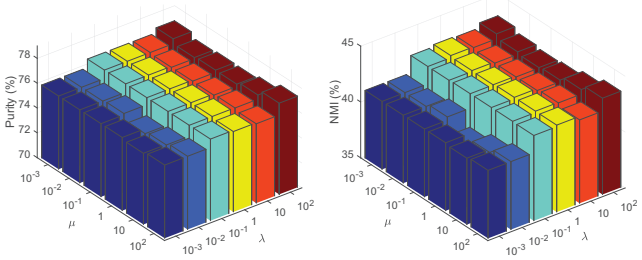
Evaluating Lifelong learning: This subsection studies the lifelong learning property of our L^{2}SC model by following (Ruvolo and Eaton 2013), *i.e.,* how well the clustering performance will be as the number of clustering tasks t increases. We adopt the WebKB4 dataset, set the sequence of learned t tasks as: Task1, Task2, Task3 and Task4, and present the clustering performance in Figure 2. Obviously, as new clustering task is imposed step-by-step, the performances (*i.e.,* Purity and NMI) for both learned and learning task are improved gradually when comparing with stSC (initial clustering result of each line in Figure 2), which justifies

Table 3: Comparison results in terms of 3 different metrics (mean ± standard deviation) on 20NewsGroups dataset.

	Metrics	stSC	uSC	OnestepSC	MBC	SMBC	SMKC	MTSC	MTCMRL	Ours
Task1	Purity(%)	63.89±0.15	44.52±0.49	66.53±1.98	47.69±2.13	50.45±5.41	73.89±1.36	77.27±0.78	**81.59±1.45**	81.05±1.05
	NMI(%)	30.77±0.33	4.35±0.33	38.74±1.10	19.29±2.76	24.80±3.18	37.75±2.68	45.35±0.83	**49.38±1.55**	46.38±0.62
	RI(%)	61.27±0.30	56.32±0.24	65.54±1.48	48.93±7.45	54.19±0.72	72.09±1.17	74.31±0.69	78.45±1.47	**78.60±0.15**
Task2	Purity(%)	53.54±0.48	40.89±0.00	55.97±0.13	48.56±2.96	50.46±1.31	66.81±1.44	63.55±0.78	65.06±0.77	**73.47±0.09**
	NMI(%)	34.68±0.20	9.92±0.00	32.86±0.08	21.27±3.45	23.23±7.97	40.76±2.88	42.52±0.33	44.21±0.39	**52.75±0.41**
	RI(%)	60.08±0.66	65.51±0.00	62.54±0.17	64.31±2.16	63.82±4.60	76.26±1.01	70.23±0.21	72.19±0.18	**81.17±0.05**
Task3	Purity(%)	59.07±0.00	54.74±0.00	59.87±1.68	49.85±3.05	52.34±1.43	60.40±2.15	68.86±1.26	77.86±0.69	**83.73±0.11**
	NMI(%)	34.58±0.09	17.63±0.00	39.25±1.93	20.53±5.41	23.37±4.01	30.24±1.12	38.81±1.56	46.05±1.31	**55.54±0.37**
	RI(%)	61.08±0.01	58.10±0.00	61.47±1.51	48.35±2.76	52.67±0.89	65.23±0.98	64.06±1.30	75.14±0.58	**82.06±0.14**
Task4	Purity(%)	51.51±0.14	52.35±0.45	54.37±0.29	46.33±2.86	**75.18±4.77**	68.69±0.35	67.35±0.35	74.85±0.89	72.08±3.19
	NMI(%)	32.53±0.32	26.13±0.87	34.12±0.73	21.37±3.48	44.09±4.78	41.15±0.95	44.03±0.31	54.02±0.65	**56.71±1.33**
	RI(%)	52.54±0.19	64.70±0.25	56.27±0.38	46.61±2.70	78.99±2.71	74.68±0.41	70.35±0.41	78.56±0.74	**82.29±1.25**
	Avg.Purity(%)	56.99±0.19	48.12±0.23	59.18±1.02	48.11±2.75	57.01±4.35	67.45±1.33	69.25±0.64	74.91±0.93	**77.73±1.11**
	Avg.NMI(%)	33.03±0.24	14.51±0.30	36.24±0.96	20.62±3.78	28.12±4.98	37.48±1.93	42.68±0.76	48.39±0.98	**52.84±0.68**
	Avg.RI(%)	58.73±0.29	61.16±0.12	61.46±0.89	52.05±3.77	62.42±2.23	72.07±0.89	69.74±0.41	76.15±0.85	**81.12±0.39**

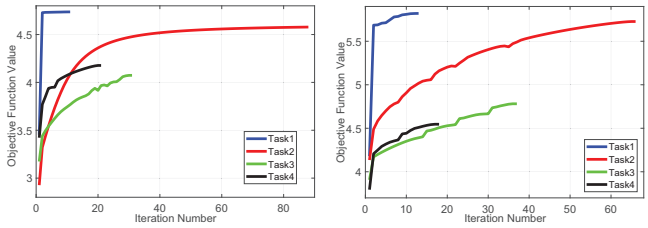
Table 4: Runtime (seconds) on a standard CPU of all competing models.

	stSC	uSC	OnestepSC	MBC	SMBC	SMKC	MTSC	MTCMRL	Ours
WebKB4(s)	1.22±0.01	1.21±0.03	600.91±26.60	6.97±1.08	5.77±0.14	34.79±0.47	69.72±1.26	14.51±1.30	2.69±0.02
Reuters(s)	0.87±0.20	1.31±0.22	1410.47±47.47	3.91±0.19	5.47±0.14	16.86±0.84	71.79±1.20	8.26±0.28	1.32±0.01
20NewsGroups(s)	2.92±0.07	5.27±0.02	3500.16±77.70	19.19±1.04	26.54±1.30	316.22±3.53	44.01±3.53	384.52±19.55	9.95±0.29



(a) WebKB4 Dataset (b) WebKB4 Dataset

Figure 3: Parameter analysis of our proposed L^2SC model on WebKB4 dataset.



(a) WebKB4 Dataset (b) 20NewsGroups Dataset

Figure 4: Convergence analysis of our proposed L^2SC model on (a) WebKB4 and (b) 20NewsGroups datasets, where lines with different colors denote different tasks in each dataset.

L^2SC can accumulate continually knowledge and accomplish lifelong learning just like "human learning". Furthermore, the performance of early clustering tasks can improve obviously than succeeding ones, *i.e.,* the early spectral clustering tasks can benefit more from the stored knowledge than later ones.

Parameter Investigation: In order to study how the parameters λ and μ affect the clustering performance of our L^2SC. For the WebKB4 dataset, we repeat the L^2SC ten times by fixing one parameter and tuning the other parameters in $[0.001, 0.01, 0.1, 1, 10, 100]$. As depicted in Figure 3, we can notice that clustering performance changes with different ratio of parameters, which give the evidence that the appropriate parameters can make the generalization performance better, *e.g.,* $\lambda = 100$ for WebKB4 dataset.

Convergence Analysis: To investigate the convergence of our proposed optimisation algorithm for solving L^2SC model, we plot the value of total loss terms for each new task on WebKB4 and 20NewsGroups datasets. As shown in Figure 4, the objective function values increase with respect to iterations, and the values for each new task approach to be a fixed point after a few iterations (*e.g.,* less than 20 iteration for Task 4 on both datasets), *i.e.,* although the convergence analysis of L^2SC cannot be proved directly in our paper, we find it converge asymptotically on the real-world datasets.

Conclusion

This paper studies how to add spectral clustering capability into original spectral clustering system without damaging existing capabilities. Specifically, we propose a lifelong learning model by incorporating spectral clustering: lifelong spectral clustering (L^2SC), which learns a library of orthogonal basis as a set of latent cluster centers, and a library of embedded features for all the spectral clustering tasks. When a new spectral clustering task arrives, L^2SC can transfer knowledge embedded in the shared knowledge libraries to encode the coming spectral clustering task with encoding matrix, and redefine the libraries with the fresh knowledge. We have conducted experiments on several real-world datasets; the experimental results demonstrate the effectiveness and efficiency of our proposed L^2SC model.

References

Ammar, H. B.; Eaton, E.; Ruvolo, P.; and Taylor, M. 2014. Online multi-task learning for policy gradient methods. In *ICML-14*, 1206–1214.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* (73):243–272.

Chen, Z.; Ma, N.; and Liu, B. 2018. Lifelong learning for sentiment classification. *arXiv preprint arXiv:1801.02808*.

Han, D., and Kim, J. 2015. Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*, 5016–5023.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huy, T. N.; Shao, H.; Tong, B.; and Suzuki, E. 2013. A feature-free and parameter-light multi-task clustering framework. *Knowledge and information systems* 36(1):251–276.

Isele, D., and Cosgun, A. 2018. Selective experience replay for lifelong learning. *arXiv preprint arXiv:1802.10269*.

Isele, D.; Rostami, M.; and Eaton, E. 2016. Using task features for zero-shot knowledge transfer in lifelong learning. In *IJCAI*, 1620–1626.

Jiang, W., and Chung, F.-l. 2012. Transfer spectral clustering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 789–803. Springer.

Kang, Z.; Peng, C.; Cheng, Q.; and Xu, Z. 2018. Unified spectral clustering with optimal graph. In *AAAI*.

Li, Z., and Chen, J. 2015. Superpixel segmentation using linear spectral clustering. In *CVPR*, 1356–1363.

Li, Z., and Hoiem, D. 2016. Learning without forgetting. In *ECCV*, 614–629.

Manton, J. H. 2002. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing* 50(3):635–650.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, 849–856.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint l2, 1-norms minimization. In *NIPS*, 1813–1821.

Pang, Y.; Xie, J.; Nie, F.; and Li, X. 2018. Spectral clustering by joint spectral embedding and spectral rotation. *IEEE transactions on cybernetics*.

Pentina, A., and Lampert, C. H. 2015. Lifelong learning with non-iid tasks. In *Advances in Neural Information Processing Systems*, 1540–1548.

Rannen Ep Triki, A.; Aljundi, R.; Blaschko, M.; and Tuytelaars, T. 2017. Encoder based lifelong learning. In *Proceedings ICCV 2017*, 1320–1328.

Ruvolo, P., and Eaton, E. 2013. Ella: An efficient lifelong learning algorithm. In *ICML*, 507–515.

Schütze, H.; Manning, C. D.; and Raghavan, P. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22(8):888–905.

Sun, G.; Cong, Y.; Hou, D.; Fan, H.; Xu, X.; and Yu, H. 2017. Joint household characteristic prediction via smart meter data. *IEEE Transactions on Smart Grid*.

Sun, G.; Cong, Y.; Li, J.; and Fu, Y. 2018a. Robust lifelong multi-task multi-view representation learning. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, 91–98. IEEE.

Sun, G.; Cong, Y.; Liu, J.; Liu, L.; Xu, X.; and Yu, H. 2018b. Lifelong metric learning. *IEEE transactions on cybernetics* (99):1–12.

Sun, G.; Cong, Y.; Wang, Q.; Zhong, B.; and Fu, Y. 2019. Representative task self-selection for flexible clustered lifelong learning. *arXiv preprint arXiv:1903.02173*.

Sun, G.; Cong, Y.; and Xu, X. 2018. Active lifelong learning with" watchdog". In *AAAI*.

Thrun, S., and O'Sullivan, J. 1996. Discovering structure in multiple learning tasks: The tc algorithm. In *ICML*, 489–497.

Thrun, S. 2012. *Explanation-based neural network learning: A lifelong learning approach*, volume 357. Springer Science & Business Media.

Wang, L.; Ding, Z.; and Fu, Y. 2019. Low-rank transfer human motion segmentation. *IEEE Transactions on Image Processing* 28(2):1023–1034.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. S. 2018. Lifelong domain word embedding via meta-learning. *arXiv preprint arXiv:1805.09991*.

Yang, Y.; Ma, Z.; Yang, Y.; Nie, F.; and Shen, H. T. 2015. Multitask spectral clustering by exploring intertask correlation. *IEEE transactions on cybernetics* 45(5):1083–1094.

Zhang, J., and Zhang, C. 2011. Multitask bregman clustering. *Neurocomputing* 74(10):1720–1734.

Zhang, X.; Zhang, X.; Liu, H.; and Liu, X. 2017. Multitask clustering through instances transfer. *Neurocomputing* 251:145–155.

Zhang, X.; Zhang, X.; Liu, H.; and Luo, J. 2018. Multi-task clustering with model relation learning. In *IJCAI*, 3132–3140.

Zhang, X.; Zhang, X.; and Liu, H. 2015. Smart multitask bregman clustering and multitask kernel clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10(1):8.

Zhao, H.; Ding, Z.; and Fu, Y. 2017. Multi-view clustering via deep matrix factorization. In *AAAI*, 2921–2927. AAAI Press.

Zhou, Q., and Zhao, Q. 2015. Flexible clustered multi-task learning by learning representative tasks. *IEEE transactions on pattern analysis and machine intelligence* 38(2):266–278.

Zhu, X.; He, W.; Li, Y.; Yang, Y.; Zhang, S.; Hu, R.; and Zhu, Y. 2017. One-step spectral clustering via dynamically learning affinity matrix and subspace. In *AAAI*, 2963–2969.