Effect of Curriculum on the Consolidation of Neural Network Task Knowledge

Ryan Poirier and Daniel L. Silver
Intelligent Information Technology Research Laboratory,
Jodrey School of Computer Science, Acadia University,
Wolfville, Nova Scotia, Canada B4P 2R6
E-mail: danny.silver@acadiau.ca

Abstract— Fundamental to the problem of lifelong machine learning is how to consolidate the knowledge of a learned task within a long-term memory structure (domain knowledge) without the loss of prior knowledge. We investigate the effect of curriculum, ie. the order in which tasks are learned, on the consolidation of task knowledge. Relevant background material on knowledge transfer and consolidation using multiple task learning (MTL) neural networks is reviewed. A large MTL network is used as the long-term memory structure and task rehearsal overcomes the stability-plasticity problem and the loss of prior knowledge. Experimental results demonstrate that curriculum has an important effect on the accuracy of consolidated knowledge particularly for the first few tasks that are learned. The results also suggest that, for given set of tasks and training examples, the mean accuracy of consolidated domain knowledge converges to the same level regardless of the curriculum.

I. INTRODUCTION

The majority of machine learning research has focused on the single task learning approach where a hypothesis for a single task is induced from a set of training examples with no regard to previous learning or to the retention of task knowledge for future learning. In contrast, humans take advantage of previous learning by retaining task knowledge and transferring this knowledge when learning a new and related task. Life-long learning is a relatively new area of machine learning research concerned with the persistent and cumulative nature of learning [13]. Life-long learning considers situations in which a learner faces a series of different tasks and develops methods of retaining and using task knowledge to improve the effectiveness (more accurate hypotheses) and efficiency (shorter training times) of learning. Our research investigates methods of knowledge retention and transfer within the context of artificial neural networks and applies these methods to life-long learning problems, such as learning accurate medical diagnostic models from small samples of a patient population [11].

One of the fundamental problems in developing a life-long learning system is devising a method of retaining task knowledge in an efficient and effective manner such that it can be later used when learning a new task. We argue that this requires the consolidation of new task knowledge with previously learned task knowledge within a domain knowledge structure. In [12] we present a theory of task knowledge consolidation within the context of multiple task learning (MTL) neural networks and test the theory on a synthetic domain of tasks. The results indicate that it is possible to sequentially consolidate task knowledge through the rehearsal of previously learned tasks so as to overcome the stability-plasticity and the loss of prior task knowledge. This paper expands on this work by investigating the effect of curriculum (the order in which tasks are learned) on the consolidation process. Through a series of experiments we demonstrate that curriculum has an important effect on the accuracy of consolidated knowledge particularly for the first few tasks that are learned. The experiments also suggest that, for a given set of tasks and training examples, the mean accuracy of consolidated domain knowledge converges to the same level regardless of the curriculum.

II. BACKGROUND

The constraint on a learning system's hypothesis space, beyond the criterion of consistency with the training examples, is called inductive bias [6]. For example, Occam's Razor suggests a bias for simple over more complex hypotheses. Inductive bias is essential for the development of a hypothesis with good generalization from a practical number of examples. Ideally, a life-long learning system can select its inductive bias to tailor the preference for hypotheses according to the task being learned. One type of inductive bias is prior knowledge of the task domain [1]. The retention and use of task domain knowledge as a source of inductive bias remains an open problem in machine learning [13], [2].

In [9] we define *knowledge-based inductive learning* as a life-long learning method that uses knowledge of the task domain as a source of inductive bias. As with a standard inductive learner, training examples are used to develop a hypothesis of a classification task. However, unlike a standard learning system, knowledge from each hypothesis is saved in a long-term memory structure called domain knowledge. When learning a new task, aspects of domain knowledge are selected to provide a positive inductive bias to the learning system. The result is a more accurate hypothesis developed in a shorter period of time. The method relies on the transfer of knowledge from one or more prior secondary tasks, stored in domain knowledge, to the hypothesis for a new primary task. Consequently, the problem of selecting an appropriate bias becomes one of selecting the most related task knowledge for
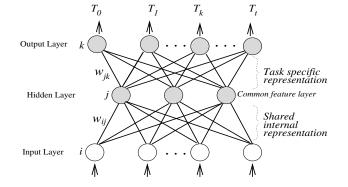
Fig. 1. A multiple task learning (MTL) network with an output node for each task being learned in parallel. The representation formed in the lower portion of the network is common to all tasks.

transfer. Much of our prior work has focused on knowledge transfer and the measurement of task relatedness for the purposes of learning a new task [10]. The following provides a review of those aspects of knowledge transfer that that are relevant to the discussion of knowledge retention and consolidation.

A. Knowledge Transfer

Multiple task learning (MTL) neural networks are one of the better documented methods of knowledge functional transfer [2]. An MTL network is a feed-forward multi-layer network with an output for each task that is to be learned. The standard back-propagation of error learning algorithm is used to train all tasks in parallel. Consequently, MTL training examples are composed of a set of input attributes and a target output for each task. Figure 1 shows a simple MTL network containing a hidden layer of nodes that are common to all tasks. The sharing of internal representation is the method by which inductive bias occurs within an MTL network [1]. MTL is a powerful method of knowledge transfer because it allows two or more tasks to share all or part of internal representation to the extent to which it is mutually beneficial. The more that tasks are related the more they will share representation and create positive inductive bias.

B. Selective Transfer and Task Rehearsal

In [9] we developed a method of knowledge transfer that distinguishes knowledge from related and unrelated tasks; such a method was previously lacking [2], [13]. ηMTL, a modified version of MTL, was created to provide a solution to the problem of selective transfer of secondary task knowledge. Using a measure of secondary task to primary task relatedness an ηMTL network can favourably bias the induction of a hypothesis for a primary task. Various functional and structural measures of relatedness are proposed and compared in [10]. The conclusion is that a consolidated representation of all previously learned tasks would provide the best source for knowledge transfer because it provides the basis for measuring deep structural similarity between tasks.

In [11] the *task rehearsal method* was introduced as a knowledge-based inductive learning system that is able to

retain and recall learned task knowledge. Building on the theory of pseudo-rehearsal [8], previously learned but unconsolidated task representations are used to generate *virtual examples* as a source of functional knowledge. After a task T_k has been successfully learned (to a specified level of generalization accuracy), its hypothesis representation is saved in domain knowledge. This representation acts as a surrogate for the space of input-output examples that defines task T_k. Virtual examples of the input-output space for T_k can be produced (with the same level of generalization error) by passing inputs to the domain knowledge representation for T_k and recording the outputs. When learning a new task, T_0, the domain knowledge representations for tasks $T_1...T_k...T_t$ are used to generate corresponding *virtual* output values from the set of T_0 training examples. The resulting set of virtual examples is used to relearn or *rehearse* the domain knowledge tasks in parallel with the learning of T_0 in an MTL or ηMTL network. It is through the rehearsal of previously learned tasks that knowledge is transferred to the new task.

C. The Need for Consolidated Domain Knowledge

Long-term knowledge retention is necessary for a knowledge-based inductive learning system, however, it is not sufficient. Domain knowledge must be integrated in a systematic fashion for the purposes of efficient (minimize storage) and effective (maintain hypothesis accuracy) retention and for more efficient (rapid indexing) and effective (appropriate choice of related knowledge) transfer during future learning. The process of integration we define as *consolidation of task knowledge* and the resulting representation we define to be *consolidated domain knowledge* [12].

The question of how new task knowledge can be consolidated into an existing neural network without loss of prior knowledge is interesting and challenging. In fact, it is the stability-plasticity problem originally posed by [3] taken to the level of learning sets of tasks as opposed to learning sets of examples. The stability-plasticity problem points out the difficulty in trying to learn a new example within a neural network while at the same time trying to maintain knowledge of previously learned examples. The loss of the previously learned knowledge has been referred to as catastrophic forgetting [5].

D. Consolidation Through MTL and Task Rehearsal

Consolidation using a connectionist network was first proposed in [4]. The report suggests a method by which the neocortex of the mammalian brain consolidates new knowledge without loss of previous knowledge. Consolidation occurs through a slow process of interleaved learning of new and old knowledge within a long-term memory structure of sparse representation. Earlier work with MTL networks showed an ability to simultaneously learn a variety of tasks of a domain from small random weights without loss of accuracy provided there are sufficient training examples and sufficient internal representation [2]. However, there are two problems that must be overcome if MTL networks are to be used to sequentially

consolidate domain knowledge: (1) preventing the catastrophic forgetting of previously learned tasks already existing within the MTL network, particularly unrelated tasks, and (2) avoiding high-magnitude weight representations that frustrate the learning of new internal features. In [12] we propose that MTL and the task rehearsal method provides a mechanism for sequentially consolidating task knowledge that addresses these issues as well as those of efficient and effective retention and transfer.

1) Effective Storage: Experimental results described in [12] demonstrate that consolidation can be accomplished within an MTL network provided that: (1) task rehearsal is used to maintain prior task knowledge while knowledge of each new task is integrated, (2) there are sufficient training examples to ensure that features are maintained and created for all tasks, (3) there is sufficient internal representation within the network for learning each task independent of all others, (4) learning occurs slowly (small learning rate) in order to increase the probability of creating internal features that are useful to all tasks and (5) there is a method of early stopping to prevent the over-fitting of new task knowledge and the creation of high magnitude weights [7]. This approach can also be seen as a step toward a general solution to the stability-plasticity problem that results in the catastrophic loss of prior neural network knowledge.

2) Efficient Storage: MTL networks provide an efficient representational form for the consolidation of knowledge from multiple tasks. Network representation provides significant space savings over the original training examples and the shared use of internal representation by related hypotheses reduces redundant storage. The time required to consolidate each new task into an ever-growing MTL network would seem to continually increase. However, we show in [12] that the average time to consolidate a new task into an existing MTL network will decrease as the number of previously consolidated tasks increases. This is because the probability of the MTL network already having the required internal representation (common features) for a new task increases as the number of previously consolidated tasks increases.

3) Effective Retrieval: Our previous research into knowledge transfer has shown that the most effective methods of determining the relatedness between two tasks are structural measures that calculate a task's use of a common internal representation. MTL networks provide a common internal representation and the degree to which tasks share features of this representation have been shown to be powerful measures of task relatedness [11].

4) Efficient Retrieval: We are actively working on structural measures of relatedness based on statistics such as the cosine similarity of hidden to output weights and the mutual information of task outputs with respect to the features generated at the hidden node layer. These measures can be computed in time polynomial in the number of tasks stored in consolidated domain knowledge.

III. CURRICULUM AND THE CONSOLIDATION OF TASK KNOWLEDGE

A. Definition

The American Heritage Dictionary of the English Language, Fourth Edition, defines *curriculum* as: "a group of related courses, often in a special field of study". Intuitively, one can imagine a curriculum of courses to have a particular structure or hierarchy. Certain courses must be taken before others since they provide the fundamental groundwork and provide a basis for knowledge transfer. In other words, things that one has learned in introductory courses will enable more efficient and effective learning in upper-year courses. In the same way, individual courses are structured in such a way that they begin with the basics and work their way to more complex material.

By extension, we can apply the same idea to a domain of tasks to be learned by a learning system. In [12], we simply considered the sequential consolidation of tasks without paying attention to the *order in which they were being consolidated*. This paper begins the exploration into the task order, which we define as the curriculum.

B. Curriculum Types

We propose that there are various types of curricula, each of which use different techniques of capturing the various internal features used by the tasks of the domain. Recall that an internal feature is a particular combination of inputs that one or more hidden nodes in the network is capable of detecting. The following outlines the range of different types of curricula based on the speed of feature acquisition.

On one end of the spectrum is the *rapid feature capture curriculum*; this ordering of tasks forces the network to create all the internal features as soon as possible. Consider k tasks from a domain that each utilize g internal features from a total of f features. The rapid feature capture curriculum will capture g features for each of the first $f \mod g$ tasks. This provides all subsequent tasks with the necessary prior knowledge for developing accurate models. Unfortunately, this means that in the early stages of consolidation there is no knowledge transfer. During this time, the overall mean accuracy of the consolidated tasks may be adversely affected.

On the opposite end of the spectrum is the *gradual feature capture curriculum*; this ordering spreads out the acquisition of the various features over the greatest number of tasks. This class of curricula would capture g features when learning the first task, then 1 feature over the next $f + 1 - g$ tasks. This curriculum has the advantage that each new task can receive beneficial prior knowledge from the consolidated tasks. Therefore, the overall mean accuracy of the consolidated tasks will remain stable or increase. Its disadvantage is the number of tasks that must be learned in order to acquire all the internal features used by tasks in the domain.

Assuming there are sufficient training examples to learn each task to an appropriate accuracy level, we propose that the mean accuracy over all consolidated tasks will converge to the same level regardless of the curriculum. This will occur if all curricula learn the same tasks.

IV. Experiments

To test our theory of the effect of curriculum on the sequential consolidation of task knowledge in a MTL network we conduct a series of experiments on a synthetic domain of tasks. Our objectives are (1) to examine the effect on consolidated task accuracy under various short curricula where the order of tasks varies by their relatedness to each other and (2) to examine the effect on task accuracy when learning a larger set of tasks under various curricula.

The first experiment consolidates six different curricula of three tasks where the relatedness between each task varies from no relation to full reuse of previously learned internal representation. The study examines the variation in consolidated task accuracy (based on an independent test set) as the system moves through the curriculum. We anticipate that gradual feature capture curricula will maintain or increase their accuracies, whereas rapid feature capture curricula may show some loss of accuracy. The second experiment examines the variation in consolidated task accuracy under eight different curricula of six tasks. Our expectation is that the mean accuracy over all tasks will tend to converge to a narrow band of values regardless of the curricula because all internal features are eventually learned and shared.

Both experiments were conducted using the Research and Application System for Lifelong Learning, RASL3, developed at Acadia[1]. RASL3 is capable of single task learning, variants of multiple task learning and task rehearsal via the generation of virtual examples from retained task knowledge (consolidated or unconsolidated).

A. Test Domain

The Logic domain consists of eight synthetic tasks where each task's output is a boolean logic function of 4 variables. Table I presents the entire domain of tasks. The goal in designing the domain was to create a set of non-linearly separable classification tasks that shared two linearly separable features in various ways. The features used are boolean expressions of the form $(x > 0.5 \vee y > 0.5)$, where x and y are input variables. The domain has ten inputs, labelled a through j which generate 6 unique features. The tasks of the domain vary in their degree of relatedness to one another based on their shared use of these features. For example, T_5 shares the feature $(g > 0.5 \vee h > 0.5)$ with tasks T_2 and T_3 and feature $(i > 0.5 \vee j > 0.5)$ with T_4. Because all tasks are non-linearly separable, at least two hidden nodes are required to form an internal representation of the features for each task. For each task, 2 hidden nodes form a disjunction of two inputs and then an output node forms a conjunction of 2 hidden nodes. This suggests that the minimum configuration of an MTL network for consolidating all tasks would be 10 input, 6 hidden and 7 output nodes.

The examples of the Logic domain tasks were randomly generated for each repeated study. 500 examples were gen-

[1]See (http://birdcage.acadiau.ca/iitrl)

Task Name	Logical Expression for Task
T_0	$(a > 0.5 \vee b > 0.5) \wedge (c > 0.5 \vee d > 0.5)$
T_1	$(c > 0.5 \vee d > 0.5) \wedge (e > 0.5 \vee f > 0.5)$
T_2	$(c > 0.5 \vee d > 0.5) \wedge (g > 0.5 \vee h > 0.5)$
T_3	$(e > 0.5 \vee f > 0.5) \wedge (g > 0.5 \vee h > 0.5)$
T_4	$(e > 0.5 \vee f > 0.5) \wedge (i > 0.5 \vee j > 0.5)$
T_5	$(g > 0.5 \vee h > 0.5) \wedge (i > 0.5 \vee j > 0.5)$
T_6	$(i > 0.5 \vee j > 0.5) \wedge (a > 0.5 \vee c > 0.5)$
T_7	$(a > 0.5 \vee b > 0.5) \wedge (e > 0.5 \vee f > 0.5)$

erated for each task as a training set; 200 examples for a validation set and 1000 examples for an independent test set.

B. General Method

The MTL neural networks used in the experiments have an input layer of 10 nodes, one hidden layer (common feature layer) of 20 nodes, and an output layer of up to 7 nodes, one for each task. As each task is learned one output node is added to the network and the weights between the hidden nodes and this output node are initialized to small random values. To maintain prior domain knowledge, all secondary tasks are rehearsed within the network as the new task is learned. The existing consolidated representation of the secondary tasks is used to generate virtual examples for rehearsal by passing the training examples for the new task through the representation and producing the corresponding target outputs.

In all experiments, the mean square error (MSE) cost function is minimized by the back-propagation algorithm that uses a momentum term. Preliminary training determined that a learning rate of 0.01 and momentum term of 0.9 were good choices for the experiment. Prior to consolidation the weights of all MTL networks are randomly initialized to values in the range -0.1 to 0.1. For all experiments, training proceeds for up to 10,000 iterations or until the average MSE over the validation sets decrease to a specified maximum level. The network representation is saved at this point. Each experiment reports the results of five repetitions using different random initial weights for each repetition.

Performance of the methods is compared in terms of the effectiveness of maintaining the accuracy of the consolidated domain knowledge tasks within the MTL network. Effectiveness is measured as the mean proportion of correct classifications (accuracy), over all repetitions, made by the hypotheses against a 1000 example test set.

C. Experiment 1: Consolidating Six Curricula of Three Tasks.

1) Method: This experiment examines the consolidation of six different curriculum of three tasks as shown on Table II. The relatedness between each task varies from no relation to full reuse of previously learned internal representation. The study examines the variation in consolidated task accuracy (based on an independent test set) as the system moves through the curriculum. We expect that curricula C011 and C012 that gradually capture internal features will increase their mean

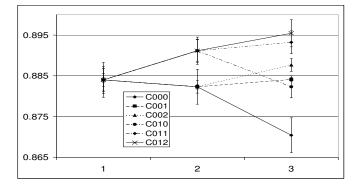| ID	Curriculum of Tasks	Comments
C000	T_0, T_3, T_6	rapid feature capture
C001	T_0, T_3, T_5	
C002	T_0, T_3, T_2	
C010	T_0, T_1, T_5	
C011	T_0, T_1, T_2	
C012	T_0, T_1, T_7	gradual feature capture

Fig. 2. Results of Experiment 1. Shown is the mean classification accuracy and standard error versus the number of tasks as a function of curriculum.

accuracies over the sequence of tasks. In contrast, we expect C000 to suffer from a lack of knowledge transfer over the curriculum because its three tasks are unrelated (do not share any features).

When consolidation of each task begins, the errors on the previously learned tasks are very small. Only the new task shows significant error. This guides the back-propagation algorithm to find and/or create the necessary internal representation for the new task. This process will interfere with the rehearsal of the previously consolidated tasks and drive their error rates upward temporarily. However, over several thousand iterations, sufficient internal representation will be found for all tasks and the mean error rate should drop below the tolerance level. In this way task rehearsal is used to maintain the accuracy of prior task knowledge while the new task is consolidated into the MTL network.

2) Results and Discussion: Figure 2 show the results from the six curricula studied in this experiment. Each graph shows the variation in test set accuracy as the first, second and third task is integrated into the consolidated MTL network. In general, the results indicate that it is possible to consolidate new task knowledge into the MTL network without significant loss of prior task knowledge. More specifically, the results verify that the C00 curriculum suffers from the least amount of knowledge transfer during consolidation. The mean accuracy over these tasks drops off because the hypothesis for T_3 and then T_6 do not receive inductive bias (previously learned features) from their predecessors. C001 does slightly better because the hypothesis for T_5 within the network makes use of the previously learned feature that detects inputs g and h (refer to Table I). The C002 curriculum results in even higher mean accuracy because the final hypothesis for T_2 can take advantage of internal features for inputs c, d, g and h. C010 is interesting because it shows an increase in the mean accuracy when learning the second task, T_1, but a drop in accuracy when learning the final task, T_5. This makes sense because the hypothesis for T_1 can take advantage of the feature that detects inputs c and d whereas the hypothesis for T_5 must learn its features strictly from the training examples. Finally, C011 and C012 do well as predicted. There tasks continually use previously consolidated task knowledge while learning new features.

D. Experiment 2: Consolidating Eight Curricula of Six Tasks.

1) Method: This experiment tests the consolidation method on eight different curricula of tasks that vary in their relatedness to one another. The objective is to show that under the proper conditions that the mean accuracy over all tasks will tend to converge to a narrow band of values regardless of the curricula because all internal features are eventually learned and shared under each curriculum. We will consider the proper conditions to be the values chosen for the four factors examined in the first experiment: a learning rate of 0.01, 20 hidden nodes, 500 training examples and with a validation set of 200 examples to prevent over-fitting. The eight curricula contain C1 = $T_0, T_1, T_2, T_3, T_4, T_5$ and seven others that were randomly selected. C1 is one of the best possible orderings as it is a gradual feature capturing curriculum; each new task can take advantage of one or more internal features that have been learned from a previous task.

After each new task is learned the consolidated MTL network representation is saved. Before training begins on the next task, this consolidated MTL network is used as the initial representation (*e.g.* for C1, T_1 will be learned starting from the MTL network representation for T_0; T_2 will be learned starting with the consolidated MTL network representation for T_0 and T_1; etc). Only the weights between the hidden nodes and the new task output node are initialized to small random values. All previously learned tasks of the sequence are rehearsed within the network when consolidating a new task. The existing consolidated MTL network is used as the source of the virtual examples for this rehearsal.

2) Results and Discussion: Figure 3 shows a wide variation in mean classification accuracy over the various curricula. This is partially because of differences in the training examples for each of the tasks and partially because of the differences in the curricula. Curriculum C1 exhibits what we expected when learning tasks T_0, T_1 and T_2 but then a slow but study decline in accuracy is observed through to consolidation of the final task, T_5. This is because of an averaging effect over all of the tasks that becomes increasing similar across curricula as all internal features of domain of tasks is learned.

V. SUMMARY AND CONCLUSION

This paper extends the work reported in [12]. It addresses the question of how the order of tasks can affect the overall accuracy of the neural network model. We have defined this order to be the curriculum of tasks and outlined different types,
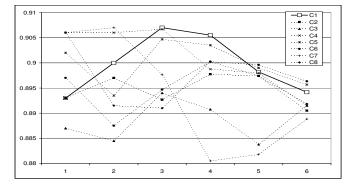
Fig. 3. Results of Experiment 2. Shown is the mean classification accuracy versus the number of tasks as a function of curriculum.

including the two extreme cases being the rapid feature capture curriculum and the gradual feature capture curriculum.

As shown in the experimental results, task curriculum has an important effect on the accuracy of consolidated knowledge. In particular, the choice in curriculum will greatly affect accuracy for the first few tasks that are learned. The results also suggest that if all the tasks of a domain are learned, the mean accuracy of domain knowledge converges to the same overall level regardless of curriculum. Future work in this area will continue to probe into the effects of curriculum using a variety of data domains.

REFERENCES

[1] Jonathan Baxter. Learning model bias. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 169–175. The MIT Press, 1996.
[2] Richard A. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.
[3] Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–64, 1987.
[4] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Technical Report PDP.CNS.94.1*, 1994.
[5] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: the sequential learning problem. *The Psychology of Learning and Motivation*, 24:109–165, 1989.
[6] Tom M. Mitchell. *Machine Learning*. McGraw Hill, New York, NY, 1997.
[7] Lorien Y. Pratt. Discriminability-based transfer between neural networks. *Advances in Neural Information Processing Systems 5*, 5:204–211, 1993. ed. C. L. Giles and S. J. Hanson and J.D. Cowan.
[8] Anthony V. Robins. Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science*, 7:123–146, 1995.
[9] Daniel L. Silver and Robert E. Mercer. The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness. *Connection Science Special Issue: Transfer in Inductive Systems*, 8(2):277–294, 1996.
[10] Daniel L. Silver and Robert E. Mercer. Selective functional transfer: Inductive bias from related tasks. *Proceedings of the International Conference on Artificial Intelligence and Soft Computing (ASC2001)*, pages 182–189, 2001.
[11] Daniel L. Silver and Robert E. Mercer. The task rehearsal method of life-long learning: Overcoming impoverished data. *Advances in Artificial Intelligence, 15th Conference of the Canadian Society for Computational Studies of Intelligence (AI'2002)*, pages 90–101, 2002.
[12] Daniel L. Silver and Ryan Poirier. Sequential consolidation of learned task knowledge. *Lecture Notes in Artificial Intelligence, 17th Conference of the Canadian Society for Computational Studies of Intelligence (AI'2004)*, pages 217–232, 2004.
[13] Sebastian Thrun. Lifelong learning algorithms. *Learning to Learn*, pages 181–209.1 Kluwer Academic Publisher, 1997.