Multitask Learning Via Interleaving: A Neural Network Investigation

David Mayo[*1], **Tyler R. Scott**[+], **Mengye Ren**[+†], **Gamaleldin Elsayed**[+],
Katherine Hermann[+], **Matt Jones**[+‡], **Michael C. Mozer**[+]
[+] Google Research, Brain Team
[‡] University of Colorado, Boulder
[†] New York University
[*] Massachusetts Institute of Technology

Abstract

The most common settings in machine learning to study multitask learning assume either that a random task is selected on each training trial, or that one task is trained to mastery and then training advances to the next. We study an intermediate setting in which tasks are interleaved, i.e., training proceeds on task $\mathcal{A}$ for some period of time, switches to another task $\mathcal{B}$ before $\mathcal{A}$ is mastered, and continues to alternate. We examine properties of modern neural net learning algorithms and architectures in this setting. The networks exhibit effects of task sequence that are qualitatively similar to established phenomena in human learning and memory, including: forgetting with relearning savings, task switching costs, and better memory consolidation with interleaved training. By improving our understanding of such properties, one can design learning schedules that are suitable given the temporal structure of the environment. We illustrate with a momentum optimizer that resets momentum following a task switch and leads to reliably better online cumulative learning accuracy.

Keywords: multitask learning; forgetting; relearning savings; task switching; interleaved training; weight consolidation

Natural environments demand that we learn to perform a diverse, unbounded set of tasks. We are challenged by the fact that we rarely have the opportunity to master one task before we are called on to perform another. Even when we achieve mastery, the demands of intervening tasks may lead to forgetting over time. The general setting of human learning is quite different than the setting in which machine learning is typically explored. Most machine-learning research in supervised and reinforcement learning focuses on learning a single task *de novo*. When multiple tasks are to be learned, the standard assumption is that training data arrives in episodes, each consisting of a distinct, novel task and/or input distribution. Figure 1 depicts this episodic training scenario, contrasted with the setting more typical of human experience in the natural world. Machine-learning paradigms that typically adopt episodic training include few-shot, transfer, continual, incremental, and meta-learning (see Murphy, 2022, Chapter 19).

In this paper, we explore the properties of neural networks trained in a structured multitask environment where tasks are interleaved (Marr, 1971; McClelland, McNaughton, & O'Reilly, 1995). We study the simplest possible scenario: one involving two tasks that alternate. We evaluate machine performance *continually* and *online*, i.e., each trial is both an opportunity for evaluating knowledge and for subsequent learning, just as it is for people. Our goal is to better understand the behavior of machine learning by examining the similarities and differences with human learning. We are not modeling human data per se; rather, we start by identifying phenomena in the human literature on multitask learning and then determine the extent to which a correspondence exists in machine learning. Discovering properties and phenomena that had not previously been identified in machine learning should prove useful for designing online, continual-learning machines, and may also offer new insights into the mechanisms underlying these phenomena in humans.

Consider catastrophic forgetting, frequently identified as a fundamental challenge to machine learning (e.g., French, 1999). But is it? People do forget catastrophically. What enables people to accumulate knowledge is the consolidation of memory (Nadel, Hupbach, Gomez, & Newman-Smith, 2012) and the slowing of forgetting with repeated experience (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Pavlik & Anderson, 2005). If machine learning exhibits the same property, then perhaps the focus on mitigating catastrophic forgetting in contrived training settings (e.g., Kirkpatrick et al., 2017; Robins, 1995) is unwarranted; instead, machine-learning research should examine emergent properties of memory in naturalistic settings (Figure 1) and then leverage these properties to propose mechanisms that supplement the natural resistance to forgetting in these settings (e.g., Flesch et al., 2018; Sprechmann et al., 2018; Russin et al., 2022).

Human Learning With Multiple Tasks

In this section, we discuss three phenomena relating to the influence of task sequence on human learning.

Forgetting With Relearning Savings

Ebbinghaus (1885/1913) performed the first experimental studies of learning and memory using himself as a subject. He

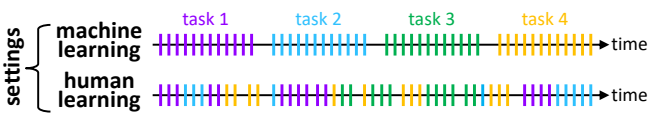

Figure 1: Contrasting the typical setting for machine learning and the setting for naturalistic human learning.

[1] Research conducted as a student researcher at Google.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

practiced lists of 12-36 random syllables and recorded the number of times needed to go through a list in order to recite it back from memory. He did this over multiple days, and not surprisingly, from one day to the next he would forget the list and need to practice it again to achieve mastery. Despite the forgetting, the time to relearn decreased over successive days. Thus, forgetting, which is due to both the passage of time and interference from intervening tasks (Sadeh, Ozubko, Winocur, & Moscovitch, 2016), does not imply that the learner is back in the state they were in prior to initial learning. To the best of our knowledge, this *relearning savings with practice* has not been demonstrated for machine learners in the setting of alternating tasks. Indeed, one might expect neural nets not to exhibit savings, because once the weights are unlearned there is no trace of their previous values. However, a hint that networks might exhibit savings comes from early work of Hinton and Sejnowski (1986), who observed that, following damage to a neural network in the form of weight corruption, retraining the network was more efficient than the initial training.

Task Switching Cost

Whenever a person switches among tasks, a performance cost is incurred (Monsell, 2003). The cost can be either in latency or accuracy, both reflecting some degradation in skill.[2] For example, suppose an individual is presented a series of digits and is asked to perform one of two tasks: ($\mathcal{A}$) classify the digit as low (< 5) or high (≥ 5), or ($\mathcal{B}$) classify the digit as odd or even. In a series of trials ordered as $\mathcal{AAAABBBBAAAA}$..., the first trial after a switch is performed more slowly than the second trial. In deterministic sequences, performance reaches asymptote on the second trial. For non-deterministic sequences, performance improves as the run length increases, up to about 5 trials. Thus, even for simple, highly practiced tasks, any interruption by another task incurs a performance cost. The rapid recovery appears to reflect trial-to-trial adaptation of the cognitive architecture (Mozer, Kinoshita, & Shettel, 2007; Wilder et al., 2013), which is rational when operating in an environment with temporal autocorrelation (Jones & Sieck, 2003; Yu & Cohen, 2008), even if it results in a cost when switches occur. The closest analog to this situation in the machine-learning literature arises in few-shot learning, a setting in which a new task needs to be learned from a few examples, e.g., using the in-context window to guide a language model to specialize in a new task (Brown et al., 2020). However, this situation has been studied for novel tasks, not revisiting previously learned tasks.

Blocked Versus Interleaved Training

When learning new skills, students benefit from *interleaved* over *blocked* practice (e.g., Taylor & Rohrer, 2010; Rohrer, 2012). Interleaved practice refers to a series of problems which demand many different skills, blocked practice to a series in which most successive problems require application of the same skill. While learners may find blocked practice easier than interleaved practice—e.g., a set of problems pertaining to the most recent lesson versus problems drawn randomly from any preceding lesson—the latter boosts learning gains and resistance to forgetting on an educationally relevant time scale (Rohrer, Dedrick, & Stershic, 2015).

As noted earlier, machine-learning research typically studies blocked training. In the present work, we study varieties of interleaved training. The closest work to ours in machine learning involves the use of nonstationary data streams in continual learning, though the focus has been on covariate shift—a change of input distribution over time (Cai, Sener, & Koltun, 2021; Lin, Shi, Pathak, & Ramanan, 2022; Ren, Iuzzolino, Mozer, & Zemel, 2021), rather than a change in how inputs should be mapped to outputs.

Methodology

We study the simplest possible setting in which multiple tasks are interleaved during training: a setting involving two distinct, alternating tasks. As in the natural world, the task sequence has temporal contiguity such that task repetition are more common than task switches. The learner is provided with an explicit signal that specifies the task to be performed. The learner must process its input in a task-appropriate manner (similar to Davidson and Mozer, 2020; cf., Flesch, Balaguer, Dekker, Nili, and Summerfield, 2018, where the task is not provided and must be inferred from feedback).

In this setting, we investigate the learning behavior of a generic neural net trained with standard stochastic gradient descent. Our analyses focus on the properties described in the previous section: relearning savings, task switching costs, and blocked versus interleaved training.

In an initial experiment, we use the CIFAR-10 data set, comprising images labeled into 10 classes. We randomly partition the classes into two tasks, $\mathcal{A}$ and $\mathcal{B}$, each requiring five-class discrimination. We assume that all inputs are processed by a ResNet-50 network (He, Zhang, Ren, & Sun, 2016) with one output head per task. The task signal is used to determine which output head to read from. (This architecture maintains task-conditional class priors, which are distorted by an architecture with a single output head.) Training on one task will alter representations in lower layers, which may negatively impact performance on the other task.

We characterize the training environment in terms of a *run length*, ρ. Models are trained on ρ passes through the complete set of training instances in task $\mathcal{A}$ and then a switch is made to $\mathcal{B}$, alternating every ρ passes. As $\rho \to 0$, training becomes much like iid (independent and identically distributed) training, where a random task is chosen on each training trial. As $\rho \to \infty$, training becomes like the standard blocked procedure in which one task is trained to asymptote and then the next task is trained to asymptote (see right column of Figure 6). We investigate an intermediate range of ρ. With any $\rho \geq 1$, the training environment exhibits strong recency effects, i.e., $\Pr(\text{task}_{t+1} = \text{task}_t) \gg \Pr(\text{task}_{t+1} \neq \text{task}_t)$.

[2] In the cognitive modeling literature, output strength from a model is often related to latency, e.g., the drift-diffusion model of Ratcliff and McKoon (2008).

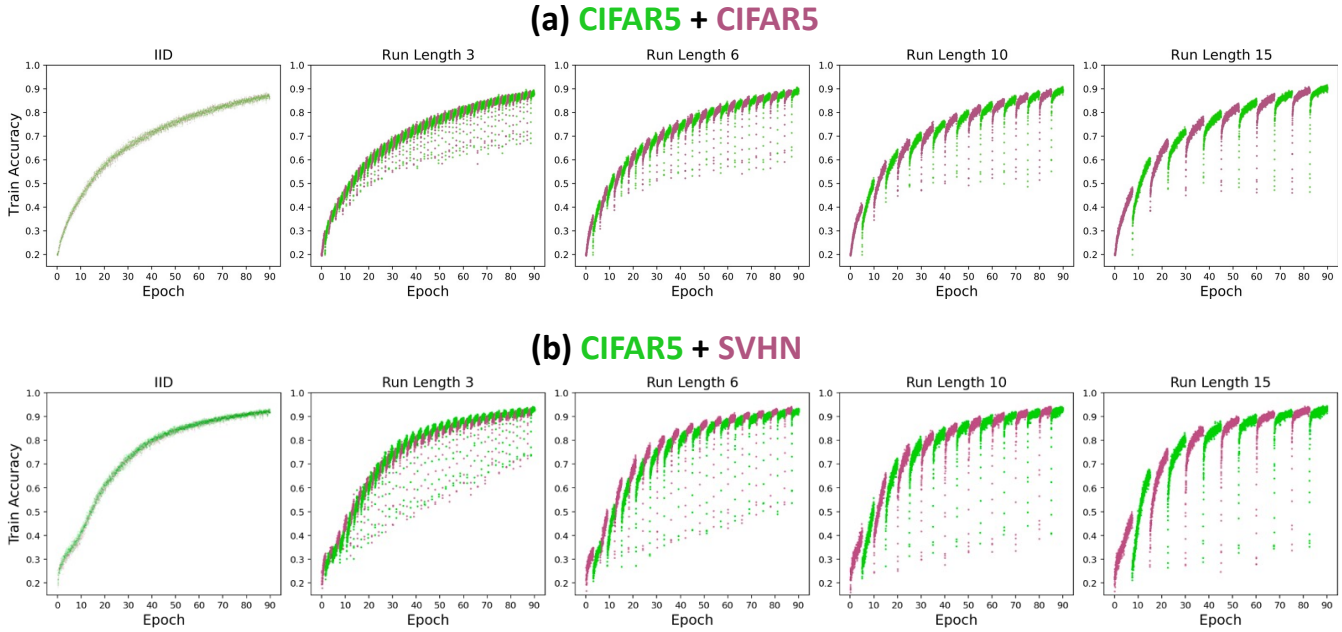

Figure 2: Training in a task-alternation environment with (a) two 5-way CIFAR tasks, (b) a 5-way CIFAR task and a 5-way SVHN task. Each graph shows accuracy per training batch for the current batch's task—indicated by color—over the course of training. Column 1 is the iid environment ($\rho = 1/48$) and columns 2-5 are for environments with run lengths $\rho \in \{3, 6, 9, 15\}$.

CIFAR-10 has 50k training instances total, half associated with each task. We train the model in *batches* of 512 same-task instances for implementation efficiency. Batches are generated at random each pass, with the constraint that examples are not shown again until all examples of a task have been presented. There are 48 full training batches per pass, and the fractional batch remaining is discarded.

The training procedure for each weight update in the model is meant to be comparable to standard practice in machine learning, not a correspondence to individual learning trials for humans. The mismatch would be problematic if we were claiming to be modeling human behavior, but our goal is rather to discover new phenomena in deep nets using cognitive phenomena as a guide.

The task-alternation environment is of course periodic and completely predictable. Nonetheless, the nets we study have no means of exploiting the environment's determinism. For example, they cannot count the number of batches of a given task. Our simulations would not benefit from introducing stochasticity via a semi-Markov process, and doing so would only increase noise variability and weaken analyses.

Simulation Details

We run 30 replications of each simulation with different weight initializations, different splits of the 10 classes into tasks $\mathcal{A}$ and $\mathcal{B}$, and different batch compositions. For each class split, we counterbalance differences in intrinsic task difficulty by performing run pairs with either $\mathcal{A}$ or $\mathcal{B}$ leading the sequence.

For the first set of results reported here (Figures 2, 5), we train with a fixed learning rate of 0.02 and a stochastic gradient descent (SGD) optimizer without momentum. SGD-without-momentum is used to avoid the influence of previous batches on the current batch, despite the fact that resulting performance is not quite state-of-the-art. However, the results we present are qualitatively unaffected by the use of fancier optimizers and learning-rate decay.

Models are trained for exactly 90 passes through all the data (*epochs*). Consequently, the total number of training runs of both tasks combined is $180/\rho$, with the number of task alternations being $180/\rho - 1$. We explored $\rho \in \{1, 2, 3, 5, 6, 10, 15\}$. We consider the two extreme cases as well: an *iid* condition in which the task alternates each batch (1/48 of an epoch; with this granularity, there is essentially no forgetting of one task while the other is trained), and a *blocked* condition in which $\rho = 90$ and there is exactly one alternation.[3]

The simulation in Figure 2b uses a different task pair: a 5-way CIFAR-10 discrimination and a 5-way SVHN discrimination. SVHN (Netzer et al., 2011) consists of photos of house address numerals, which look quite unlike the CIFAR-10 classes, which are all natural categories, such as animals and human-made objects.

Results

During training, for each batch we first evaluate model accuracy and then take a gradient step based on that batch. Figure 2a plots training accuracy as a function of epoch for an iid environment (left graph) and environments with $\rho \in \{3, 6, 10, 15\}$. The two distinct tasks are indicated by

[3]For reference on terminology: 1 *run* = ρ *epochs*; 1 *epoch* = 48 *batches*; 1 *batch* = 512 *trials*.

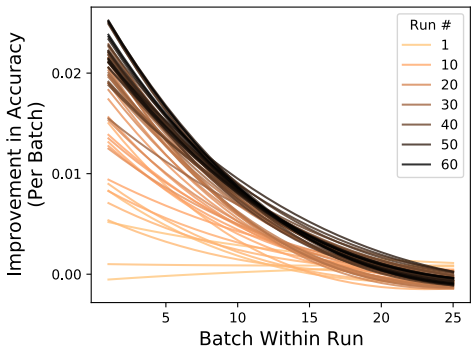

Figure 3: For $\rho = 3$, curves depicting rate of relearning—accuracy improvement within a run from batch k to batch $k+1$—as a function of k. The color coding indicates whether the training run is early (orange) or late (black) in training. Runs early in training don't show rapid relearning following a task switch, whereas later runs do. The identical pattern is observed whether we measure absolute increase in accuracy, as in the Figure, or relative increase in accuracy or relative decrease in error rate (neither of which is shown).

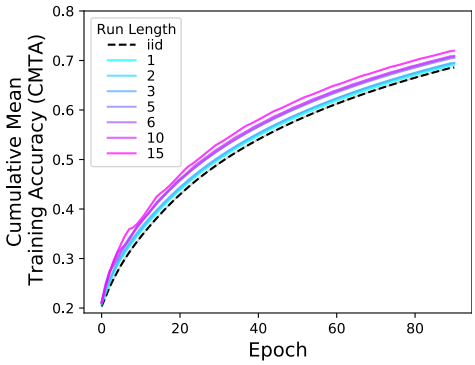

Figure 4: Cumulative mean training accuracy for environments with various run lengths (ρ) and an iid environment.

color. (Though not depicted here, we observe the same qualitative pattern for smaller ρ, including $\rho < 1$.) As run length increases, the repeated task benefits at the cost of forgetting when the task switches. These simulations differ only in the ordering of learning trials; they are controlled for overall frequency of each example and the number of gradient steps per task. As we will shortly show, evaluation (test) accuracy mirrors training accuracy.

Figure 2b shows training curves for a pair of less similar tasks: CIFAR (natural images) and SVHN (photos of house numbers from street view). The qualitative pattern of forgetting-and-relearning matches Figure 2a, except that forgetting and relearning are amplified.

Forgetting and Recovery

Two striking features of Figure 2 are that *substantial forgetting occurs during an intervening task* and *forgetting increases with the duration of the intervening task*. Neither of these features is surprising in light of the literature on catastrophic forgetting. (Not visible from the Figure, but forgetting is exponential over batches.) A third feature is less obvious: *recovery from forgetting is far quicker than initial learning*. This feature is clear from the hockey-stick shape of the learning curves, where the recovery to the previous level of performance is quick and continued learning progresses more slowly. In fact, *the rate of recovery increases with practice*, as shown in Figure 3. This evidence for relearning savings matches the human memory phenomenon first noted by Ebbinghaus. Relearning savings indicates latent residual knowledge that can be unmasked by brief practice.

Task Switching

Does the structure of the environment impact learning effectiveness? Typically in machine learning, effectiveness is measured via accuracy on an evaluation set that is separate from the training set, but for an online learning agent interacting with a structured environment, effectiveness is often measured via *prequential evaluation* (Haug, Tramountani, & Kasneci, 2022), which in our case simply means that each input is both an evaluation trial and a learning trial. A prequential evaluation measure reflects the effectiveness of a learner in a given environment. We use the average prequential accuracy of all batches earlier in training—or cumulative mean training accuracy (*CMTA*)—as shown in Figure 4 for various run lengths over the course of training. Because it is a cumulative average, CMTA at a given time is indicative of historical (not instantaneous) performance and can be compared to total reward obtained by an agent. Environments in which task repetitions are more likely (i.e., longer run lengths) benefit because there are fewer task switches. Early trials in a run pay an accuracy cost, relative to the iid setting, whereas later trials in a run benefit (Figure 5). One striking feature of Figures 2 and 5 is that *even after significant task expertise has been acquired, a performance drop and rapid recovery is observed immediately after a task switch*. This drop and recovery arguably mirrors task-switching costs in human performance, albeit at a very different time scale.

Interleaved Versus Blocked Practice

The models in Figure 2 are trained for 90 epochs. It is not self-evident what will occur if training continues indefinitely. Can a model learn both tasks completely, or will inter-task interference yield sufficient forgetting that neither task is ever learned perfectly? Informally, polling our colleagues produces a roughly 50/50 split prediction. As it turns out, both tasks are learned, although the amount of training required is more than the twice the amount that would be required to learn either task alone. The left and right upper panels of Figure 6 show extended interleaved and blocked training, respectively. (To reach asymptotic performance faster, this simulation uses momentum of 0.9 with an optimized learning rate of 0.002. Qualitative behavior is unchanged without momentum.)

In retrospect, the fact that both tasks can be learned is not surprising, as long as each task run is long enough to slightly

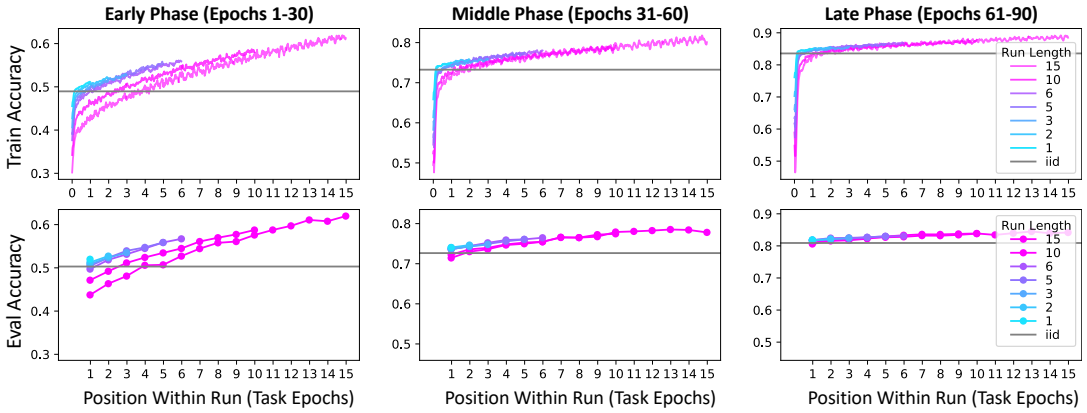

Figure 5: Accuracy measured on the current task over the course of each run. The three columns correspond to early, middle, and late phases of training. The top row shows training accuracy computed for each batch, and the bottom row shows accuracy on an evaluation set, computed only at the end of a full pass through all examples for the task (a *task epoch*). Run length is indicated by color and the iid environment is indicated by the grey horizontal line.

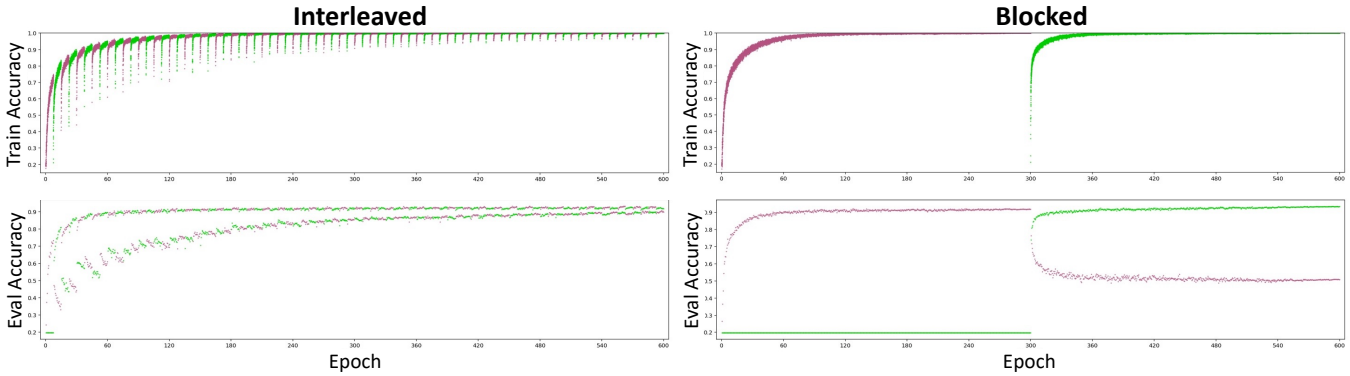

Figure 6: Extended interleaved and blocked training (left and right columns, respectively). Training accuracy in the first row, evaluation accuracy in the second row. Interleaved training is with $\rho = 15$, thus each run consists of roughly 370k training examples of one task or the other. With interleaved training, both tasks are eventually learned without resulting in overfitting.

improve performance over where the model was at the end of the previous run of the same task. As the training error drops, the amount of learning during a run drops, and the interference on the other task is reduced, allowing both tasks to be learned.

Although training accuracy will reach 100% in the interleaved condition, it is not self-evident how the model will generalize: is high training accuracy achieved at the expense of generalization? That is, in order to reduce the error with interleaved training, does the model need to memorize and thereby overfit the data? The lower-left graph in Figure 6 indicates that generalization performance continues to improve, although a small amount of forgetting on the untrained task is apparent, corresponding to the task-switching phenomenon described earlier. In contrast, with blocked training—the lower-right graph—we observe the expected catastrophic forgetting of task $\mathcal{A}$ when training shifts to task $\mathcal{B}$.

Behaviorally, the network exhibits a type of *memory consolidation with interleaved practice*. The network weights find a way to accommodate both tasks. Essentially, one can think of the individual synapses in the network as being in one of three states: (a) driven in the same direction by both tasks, (b) driven by one task and receiving no gradient signal from the other, (c) driven in opposite directions by the two tasks. As learning progresses, weight magnitudes will grow for synapses in groups (a) and (b), but the synapses in group (c) will be unable to contribute to the performance of either task. Consequently, interleaved training will naturally produce a sort of separation of knowledge about the two tasks, performing each task in a way that does not interfere with the other, even though it was not explicitly trained to do so.

Research in machine learning has focused on techniques to prevent catastrophic interference between tasks by separating the knowledge needed to perform each task in the network. Aljundi et al. (2018) propose to learn a parameter for each synapse that indicates how important it is for previously learned tasks and to modulate synaptic plasticity for new tasks based on this parameter. Zenke et al. (2017) propose intelligent synapses that seek a solution to a new task while staying near the solution for the previous task. Cheung et al. (2019) propose a method for orthogonalizing weights for different

tasks to prevent interference. Wallingford et al. (2022) adapt a pretrained model to multiple tasks using task-specific masks on synaptic plasticity. Our work argues that in the context of certain training environments, these specialized mechanisms are unnecessary, and thus it is valuable to place more emphasis on the nature of the training environment and how it affects learning. We return to this topic in the next section.

Discussion

This paper explores properties of canonical deep nets when trained on interleaved task streams. Our work is motivated by phenomena in the psychological literature, with the aim of exploring whether qualitative analogs to these phenomena can be observed in networks. Identifying such properties should allow us to better understand—and eventually improve—neural network learning. We emphasize that cognitive modeling is not our goal. That is, we are not trying to develop models that serve as accounts of psychological processes. The time scale of training in our simulations is orders of magnitude different than the time scale of human experimental work on many dimensions, including training batch size, amount of training required, how much is learned from each trial, etc. We summarize our results as follows.

- Significant forgetting occurs between tasks, what is classically referred to as catastrophic forgetting (McCloskey & Cohen, 1989). Because output heads are separate for the two tasks, observed interference is due to adaptation of latent representations in the network.
- Although forgetting can be dramatic when a task switch occurs, recovery following the switch can be as dramatic. Relearning savings were first observed by Hinton and Sejnowski (1986) in the context of damaged networks. We show that another sort of corruption—adaptation to task $\mathcal{B}$ on performance of task $\mathcal{A}$—also yields relearning savings.
- Relearning savings increase with practice, consistent with human memory phenomena (Ebbinghaus, 1885/1913). However, even after significant expertise has been acquired, performance drops immediately after a change of task, consistent with the human task-switching literature (Monsell, 2003). Several weight updates are required before the resumed task is performed with full efficacy.
- With extended interleaved training, a type of consolidation occurs such that knowledge from both tasks becomes less susceptible to erasure. Related claims have been made in both the human literature (Rohrer, 2012; Rohrer et al., 2015) and machine-learning literature (Flesch et al., 2018; Davidson & Mozer, 2020) that task alternation increases resistance to forgetting.
- Not surprisingly, online learning performance is sensitive to the structure of the environment. Controlling for overall task frequency, performance improves as the probability of task repetitions increases (Figure 4).

Our investigation points to the importance of the temporal structure of the environment in determining how networks learn and retain knowledge. The typical environment in machine learning is either iid or some artificially blocked setting such as that shown in Figure 1. Only recently have datasets been proposed with naturalistic temporal structure. These datasets include sequences of images obtained while walking through a physical building (Ren et al., 2021), photos posted online over the course of a decade (Cai et al., 2021), and evolving real-world visual concepts such as 'computer' (Lin et al., 2022). Just as human learning and memory appears to be optimized to the structure of natural environments (Anderson & Schooler, 1991), we might aim to optimize machine learning methods to the environment in which they operate (Jones et al., 2023). For instance, if knowledge consolidation occurs for tasks that are repeatedly practiced as we have shown here, and if relearning is efficient, forgetting may not be the fundamental issue it is considered to be in the continual-learning literature.

We close with an illustration of an extremely simple optimizer specific to the task-interleaving environment studied in this paper. Figure 7 presents batch-wise training accuracy for $\rho = 10$. Traces from two alternative learning procedures are superimposed: a standard momentum optimizer with decay of 0.9, labeled MOMENTUM (green points), whose hyperparameters we tuned to the task, and a momentum optimizer that is sensitive to task switches, MOMENTUM+TBR (blue points). At a task boundary, MOMENTUM+TBR resets the momentum state, which prevents continued fine-tuning to the previous task. As Figure 7 shows, MOMENTUM+TBR yields less forgetting, as evidenced by lower accuracy for MOMENTUM immediately following a switch, and also a reliable improvement in within-run accuracy, as evidenced by the upper envelope of MOMENTUM+TBR's learning curve lying above that of MOMENTUM. Our long-term objective is to devise more sophisticated learning procedures that are appropriate given the environment structure.

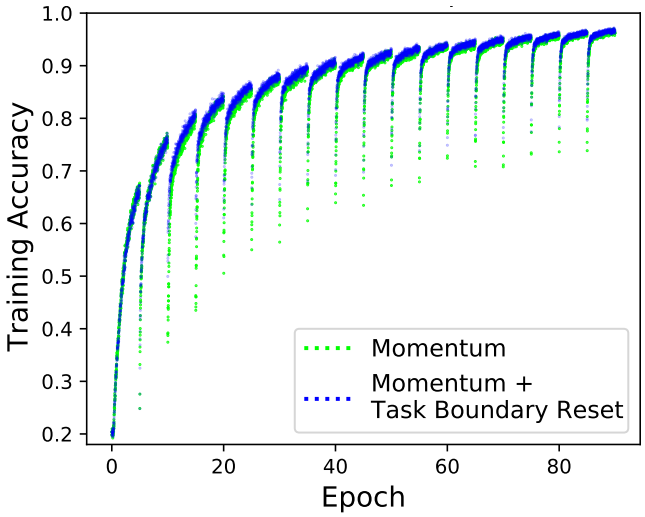

Figure 7: Accuracy per training batch as a function of epoch for $\rho = 10$, superimposing standard SGD with momentum (green) versus a variant that is sensitive to the task switch by resetting momentum (blue).

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 139–154).

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396-408.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cai, Z., Sener, O., & Koltun, V. (2021). *Online continual learning with natural distribution shifts: An empirical study with visual data.* arXiv. Retrieved from https://arxiv.org/abs/2108.09020

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Cheung, B., Terekhov, A., Chen, Y., Agrawal, P., & Olshausen, B. (2019). Superposition of many models into one. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc.

Davidson, G., & Mozer, M. C. (2020, June). Sequential mastery of multiple visual tasks: Networks naturally learn to learn and forget to forget. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Retrieved from https://arxiv.org/abs/1905.10837

Ebbinghaus, H. (1885/1913). (T. by H. A. Ruger & C. E. Bussenius, Eds.).

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313-E10322. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1800755115 doi: 10.1073/pnas.1800755115

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135.

Haug, J., Tramountani, E., & Kasneci, G. (2022). *Standardized evaluation of machine learning methods for evolving data streams.* arXiv. Retrieved from https://arxiv.org/abs/2204.13625 doi: 10.48550/ARXIV.2204.13625

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. volume 1: Foundations* (pp. 282–317). Cambridge, MA: MIT Press.

Jones, M., Scott, T. R., Ren, M., Elsayed, G. F., Hermann, K., Mayo, D., & Mozer, M. C. (2023). Learning in temporally structured environments. In *International Conference on Learning Representations (ICLR).*

Jones, M., & Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 626–640.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... others (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Lin, Z., Shi, J., Pathak, D., & Ramanan, D. (2022). *The CLEAR benchmark: Continual LEArning on Real-world imagery.* Retrieved from https://arxiv.org/abs/2201.06289

Marr, D. (1971). Simple memory: a theory for archicortex. *Philos Trans R Soc Lond B Biol Sci.*, *262*, 23–81.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev.*, *102*, 419–457.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, *24*, 109–165.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140.

Mozer, M. C., Kinoshita, S., & Shettel, M. (2007). Sequential dependencies offer insight into cognitive control. In W. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 180–193). Oxford University Press.

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press. Retrieved from probml.ai

Nadel, L., Hupbach, A., Gomez, R., & Newman-Smith, K. (2012). Memory formation, consolidation and transformation. *Neuroscience and Biobehavioral Reviews*, *36*, 1640–1645.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *Nips workshop on deep learning and unsupervised feature learning.*

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559–586.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ren, M., Iuzzolino, M. L., Mozer, M. C., & Zemel, R. (2021). Wandering within a world: Online contextualized few-shot learning. In *International Conference on Learning Representations.* Retrieved from https://openreview.net/

forum?id=oZIvHV04XgC

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, *7*(2), 123-146.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*(3), 355–367.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900.

Russin, J. L., Zolfaghar, M., Park, S. A., Boorman, E., & O'Reilly, R. C. (2022). *A neural network model of continual learning with cognitive control.*

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2016). Forgetting patterns differentiate between two forms of memory representation. *Psychological Science*, *27*(6), 810-820.

Sprechmann, P., Jayakumar, S. M., Rae, J. W., Pritzel, A., Badia, A. P., Uria, B., ... Blundell, C. (2018). *Memory-based parameter adaptation.*

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848.

Wallingford, M., Li, H., Achille, A., Ravichandran, A., Fowlkes, C., Bhotika, R., & Soatto, S. (2022, June). Task adaptive parameter sharing for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 7561-7570).

Wilder, M. H., Jones, M., Ahmed, A. A., Curran, T., & Mozer, M. C. (2013). The persistent impact of incidental experience. *Psychonomic Bulletin & Review*, *20*(6), 1221-–1231.

Yu, A. J., & Cohen, J. D. (2008). Sequential effects: Superstition or rational behavior? In *Advances in Neural Information Processing Systems 21* (p. 1873—1880).

Zenke, F., Poole, B., & Ganguli, S. (2017). Continual learning through synaptic intelligence. In *International Conference on Machine Learning* (pp. 3987–3995).