Sequential mastery of multiple tasks: Networks naturally learn to learn and forget to forget

Guy Davidson
NYU Center for Data Science
guy.davidson@nyu.edu

Michael C. Mozer
Google Research
University of Colorado at Boulder
mcmozer@gmail.com

Abstract

We explore the behavior of a standard convolutional neural net in a continual-learning setting that introduces classification tasks sequentially and requires the net to master new tasks while preserving mastery of previously learned tasks. This setting corresponds to that which human learners face as they acquire domain expertise serially, for example, as an individual studies a textbook. Through simulations involving sequences of ten related tasks, we find reason for optimism that nets will scale well as they advance from having a single skill to becoming multi-skill domain experts. We observe two key phenomena. First, forward facilitation—the accelerated learning of task \( n + 1 \) having learned \( n \) previous tasks—grows with \( n \). Second, backward interference—the forgetting of the \( n \) previous tasks when learning task \( n + 1 \) diminishes with \( n \). Amplifying forward facilitation is the goal of research on metalearning, and attenuating backward interference is the goal of research on catastrophic forgetting. We find that both of these goals are attained simply through broader exposure to a domain.

In a standard supervised setting, neural networks are trained to perform a single task, such as classification, defined in terms of a discriminative distribution \( p(y|x,D) \) for labels \( y \) conditioned on input \( x \) and data set \( D \). Although such models are useful in engineering applications, they do not reflect the breadth required for general intelligence, which includes the ability to select among many tasks. Multitask learning (Caruana 1997) is concerned with training models to perform any one of \( n \) tasks, typically via a multi-headed neural network, where head \( i \) represents the distribution \( p(y_i|x,D_1,\ldots,D_n) \). Related tasks serve as regularizers on one another (Caruana 1993, Ruder 2017).

Continual or lifelong learning (Thrun 1996, Parisi et al. 2019) addresses a naturalistic variant in which tasks are tackled sequentially and mastery of previously learned tasks must be maintained while each new task is mastered. Lifelong learning requires consideration of two issues: catastrophic forgetting (McCloskey and Cohen 1989) and metalearning (Schmidhuber 1987, Bengio 1993, Bengio and Cloutier 1991, Thrun 1996). Catastrophic forgetting is characterized by a dramatic drop in task 1 performance following training on task 2, i.e., the accuracy of the model \( p(y_1|x,D_1 \rightarrow D_2) \) is significantly lower than accuracy of the model \( p(y_1|x,D_1) \), where the arrow denotes training sequence. Metalearning aims to facilitate mastery on task \( n \) from having previously learned tasks \( 1,2,\ldots,n-1 \). Success in metalearning is measured by a reduction in training-trials-to-criterion or an increase in model accuracy given finite training for the \( n \)th task, \( p(y_n|x,D_1 \rightarrow \cdots \rightarrow D_n) \), relative to the first task, \( p(y_1|x,D_1) \).

Researchers have proposed a variety of creative approaches—specialized mechanisms, learning procedures, and architectures—either for mitigating forgetting or for enhancing transfer. We summarize these approaches in the next (related work) section. Although the literatures on catastrophic forgetting and metalearning have been considered separately for the most part, we note that they have a complementary relationship. Whereas catastrophic-forgetting reflects backward interference of a new task on previously learned tasks, metalearning reflects forward facilitation of previously learned tasks on a new task (Lopez-Paz and Ranzato 2017). Whereas catastrophic forgetting research has focused on the first task learned, metalearning research has focused on the last task learned. We thus view these two topics as endpoints of a continuum.

To unify the topics, we examine the continuum from the first task to the \( n \)th. We train models on a sequence of related tasks and investigate the consequences of introducing each new task \( i \). We measure how many training trials are required to learn the \( i \)th task while maintaining performance on tasks \( 1 \ldots i-1 \) through continued practice. Simultaneously, we measure how performance drops on tasks \( 1 \ldots i-1 \) after introducing task \( i \) and how many trials are required to retrain tasks \( 1 \ldots i-1 \). We believe that examining scaling behavior—performance as a function of \( i \)—is critical to assessing the efficacy of sequential multitask learning. Scaling behavior has been mostly overlooked in recent deep-learning research, which is odd considering its central role in computational complexity theory, and therefore,
in assessing whether existing algorithms offer any hope for extending to human-scale intelligence.

Surprisingly, we are aware of only one article (Schwarz et al. 2018) that jointly considers forgetting and meta-learning through their scaling properties. However, Schwarz et al.’s research, like that in the catastrophic-forgetting and meta-learning literatures, suggests that specialized mechanisms are required for neural networks to operate in a lifelong learning setting. The punch line of our article is that a standard neural network architecture trained sequentially to acquire and maintain mastery of multiple tasks exhibits faster acquisition of new knowledge and less disruption of previously acquired knowledge as domain expertise expands. We also argue that the net’s learning and forgetting characteristics have an intriguing correspondence to the human and animal behavioral literature.

Related research

To overcome catastrophic forgetting, standard techniques such as drop out have been suggested (Goodfellow et al. 2015), but most propose augmenting models with specialized mechanisms (Paris et al. 2019 for review). Kirkpatrick et al. (2017) introduce elastic weight consolidation, which adds a penalty to the model loss that encourages stability of weights that most contribute to performance on previously trained tasks. Lopez-Paz and Ranzato (2017) describe Gradient Episodic Memory, which retains examples of previous tasks and minimizes the aforementioned negative backward transfer. Kenner and Kanani (2018) devise FearNet, a neurally-inspired model with dual-memory design, using consolidation mechanisms modeled after mammalian sleep consolidation. Zenke, Poole, and Ganguli (2017) are similarly biologically-inspired, motivating intelligent synapses which track their relevance to particular tasks. Kamra, Gupta, and Liu (2017) offer another dual-memory model, augmenting with a generative re-play model able to recreate past experiences and improve performance on previously learned tasks.

To facilitate metalearning, mechanisms have been offered to encourage inter-task transfer, such as MAML (Finn, Abbeel, and Levine 2017) and SNAIL (Mishra et al. 2018). Other approaches employ recurrence to modify the learning procedure itself (Andrychowicz et al. 2016; Wang et al. 2017). Schwarz et al. (2018) construct a dual-component model consisting of a knowledge store of previously learned tasks and an active component that is used to efficiently learn the current task. A consolidation procedure then transfers knowledge from short- to long-term stores.

Despite the creativity of this assortment of methods, our concern centers on the fact that researchers assume the inadequacy of standard methods, and no attempt has been made to understand properties of a standard architecture as it is trained sequentially on a series of tasks, and to characterize the extent of forgetting and transfer as more tasks are learned.

Figure 1: Example training images

Methodology

The tasks we train are defined over images consisting of multiple synthetic shapes having different colors and textures (Figure 1). The tasks involve yes/no responses to questions about whether an image contains certain objects or properties, such as “is there a red object?” or “is there a spherical object?” We generate a series consisting of 10 episodes; in each episode, a new task is introduced (more details to follow on the tasks). A model is trained de novo on episode 1, and then continues training for the remaining episodes. In episode $i$, training involves a mix of examples drawn from tasks 1 to $i$ until an accuracy criterion of 95% is attained on a hold-out set for all tasks. To balance training on the newest task (task $i$ in episode $i$) and retraining on previous tasks, we adapt the methodology of Nguyen et al. (2018): half the training set consists of examples from the newest task, and the other half consists of an equal number of examples from each of the previous tasks 1 through $i−1$. (In episode 1, only the single task is trained.) The same set of training images is repeated each epoch of training, but they are randomly reassigned to different tasks from epoch to epoch. In each epoch, we roughly balance the number of yes and no target responses for each task. We turn now to details of the images, tasks, and architecture.

Image generation. We leverage the CLEVR (Johnson et al. 2017) codebase to generate 160 × 120 pixel color images each with 4 or 5 objects that vary along three dimensions: shape, color, and texture. We introduce additional features on each dimension to ensure 10 feature values per dimension. (See supplementary material for details.) We synthesized 45,000 images for a training set, roughly balancing the count of each feature across images. An additional 5,000 images were generated for a hold-out set. Each image could be used for any task. Each epoch of training involved one pass through all images, with a random assignment of images to task each epoch to satisfy the constraint on the distribution of tasks.

Tasks. For each replication of our simulation, we select one of the three dimensions and randomize the order of the ten within-dimension tasks. To reduce sensitivity of the results to order, we performed replications using a Latin square design (Bailey 2008 ch. 9), guaranteeing that within a block of ten replications, each task will appear in each ordinal position exactly once. We constructed six such Latin square blocks for each of the three dimensions, resulting in 180 total simulation repli-
cations. Because we observed no meaningful differences across task dimensions (see supplementary material), the results we report below collapse across dimension.

**Architecture.** Our experiments use a basic vision architecture with four convolutional layers followed by four fully connected layers (Figure 2). The convolutional layers—with 16, 32, 48, and 64 filters successively—each have 3x3 kernels with stride 1 and padding 1, followed by ReLU nonlinearities, batch normalization, and 2x2 max pooling. The fully-connected layers have 512 units in each, also with ReLU nonlinearities. All models were implemented in PyTorch [Paszke et al. 2017] and trained with ADAM [Kingma and Ba 2015] using a learning rate of 0.0005 and weight decay of 0.0001. Note that our model is generic and is not specialized for metalearning or for preventing catastrophic forgetting. Instead of having one output head for each task, we specify the task as a component of the input. Similar to Sort-of-CLEVR [Santoro et al. 2017], we code the task as a one-hot input vector. We concatenate the task representation to the output of the last convolutional layer before passing it to the first fully-connected layer.

### Results

**Metalearning**

Figure 3 depicts hold-out accuracy for a newly introduced task as a function of the number of training trials. Curve colors indicate the task’s ordinal position in the series of episodes, with cyan being the first and magenta being the tenth. Not surprisingly, task accuracy improves monotonically over training trials. But notably, metalearning is evidenced because the accuracy of task $i+1$ is strictly higher than the accuracy of task $i$ for $i > 2$.

To analyze our simulations more systematically, we remind the reader that the simulation sequence presents fifty-five opportunities to assess learning: the task intro-

![Figure 2: Model architecture. Input consists of image and task representation. Dashed line from task representation to Conv-1 indicates optional task modulated visual processing, described under “Task-modulated visual processing.”](image)

![Figure 3: (a) Hold-out set accuracy as a function of training trials (log scale) for a newly introduced task. Colored lines indicate task ordinal position (cyan = introduced in episode 1; magenta = introduced in episode 10). In all panels, the shaded region represents ±1 standard error of the mean. (b) Hold-out accuracy of the task introduced in episode 1 by number of times it is retrained (black = 1 time, copper = 10 times). (c) Number of trials required to reach the accuracy criterion (log scale) as a function of the number of times a given task is trained (also log scale). As in (a), the colors indicate task ordinal position (the episode in which a task is introduced). (d) Similar to (c) but plotting only the new task introduced in a given episode. (e) Hold-out accuracy attained after a fixed amount of training (22.5k trials) of a given task, graphed as a function of number of times a given task is trained. As in (a), the colors indicate the episode in which a task is introduced. (f) Similar to (e) but plotting only the new task introduced in a given episode.](image)
duced in episode 1 (i.e., ordinal position 1) is trained ten times, the task introduced in episode 2 is trained nine times, and so forth, until the task introduced in episode 10, which is trained only once. Figure 3b indicates, with one line per task, the training required in a given episode to reach a hold-out accuracy of 95%—the dashed line in Figure 3. Training required per episode is plotted as a function of the number of times the task is retrained. The downward shifting intercept of the curves for later tasks in the sequence indicates significantly easier learning and relearning. Figure 3e shows an alternative view of difficulty-of-learning by plotting accuracy after a fixed amount of (re)training. The conditions that require the least number of trials to criterion (Figure 3e) also achieve the highest accuracy after a small amount of training (Figure 3).

**Catastrophic forgetting**

Figure 3b shows the accuracy of the task introduced in the first episode ($y_1$) as it is retrained each episode. The fact that performance in a new episode drops below criterion (the dashed line) indicates backward interference. However, there is a relearning savings: the amount of interference diminishes monotonically with the number of times trained. Notably, catastrophic forgetting of task 1 is essentially eliminated by the last few episodes. Figure 3c shows very similar relearning savings for tasks 2-10 as for task 1. The roughly log-log linear curves offer evidence of power-law decrease in the retraining effort required to reach criterion.

Figure 3 also reveals that the first two episodes are anomalous. Strong backward interference on task 1 is exhibited when task 2 is introduced (the crossover of the cyan curve in Figure 3c), a phenomenon that does not occur for subsequent tasks. Similarly, strong forward interference on task 2 of task 1 is evident (slower learning for task 2 than for task 1 in Figure 3a), but tasks 3-10 are increasingly facilitated by previous learning. These findings suggest that to understand properties of neural nets, we must look beyond training on just two tasks, which is often the focus of research in transfer learning and catastrophic forgetting.

**Resilience to forgetting**

The fact that old tasks need to be retrained each episode suggests that training on a new task induces forgetting of the old. However, because we trained simultaneously on the old and new tasks, we have no opportunity to examine forgetting explicitly. However, we can clone weights at any point in the simulation and examine a different training trajectory moving forward. We took the network weights at the start of each episode $i$, at which point the network is at criterion on tasks 1 through $i-1$. Then, instead of retraining on all $i$ tasks, we train only on task $i$. We probe the network regularly to evaluate performance on old tasks.

Figure 4a depicts the time course of forgetting of the task introduced in episode 1 on each subsequent episode. The black curve corresponds to episode 2 (task 1 has been trained only once previously) and the copper curve corresponds to episode 10 (task 1 has been trained 9 times previously). Task 1 becomes more robust to backward interference from the new task in later episodes. In episode $i$, task 1 has been (re)trained $i-1$ times previously, yielding a sort of spaced practice that appears to cause the memory to be more robust. This result is suggestive of the finding in human memory that interleaved, temporally distributed practice yields more robust and durable memory (Kang et al. 2014; Cepeda et al. 2008).

Figure 4a depicts only some of the forty-five opportunities we have to assess forgetting: we have one after the model learns a single task, two after the model learns two, up to nine after the model learns the ninth task (for which we examine forgetting by training on the tenth and final task in the order). To conduct a more systematic analysis, we fit the forgetting curves for each task $i$ in each episode $e > i$. The forgetting curve characterizes accuracy $a$ after $t$ training batches of 1500 trials. Accuracy must be adjusted for guessing: because our tasks have a baseline correct-guessing rate of 0.5, we define $a = 0.5 + 0.5m$, to be the observed accuracy when memory strength $m$ lies between 0 (no task memory) and 1 (complete and accurate task memory). We explore two characterizations of memory strength. The first is of exponential decay, $m = a\exp(-\beta t)$, where $a$ is the initial accuracy, $\beta$ is a decay rate, and $t$ is the number of intervening training batches. The second is of power-law decay, $m = a(1 + \gamma t)^{-\beta}$, where $\gamma$ serves as a timescale variable. This power-law decay curve is common in the psychological literature on forgetting (Wixted and Carpenter 2007) and has the virtue over $m = at^{-\gamma}$ that it can characterize strength at $t = 0$.

We fit the exponential and power-law functions separately to the data from each of the 45 model training points across 67 replications of our experiment. Following Clauset, Shalizi, and Newman (2009), we fit each
from all three dimensions. Each replication utilizes a single permutation of the three dimensions and samples the ten tasks cycling between the dimensions (four from the first, three from the other two). We employed a similar Latin square design to balance between the permutations, such that each block of six replications includes each permutation once.

Figure 5b presents the results of 114 replications of the heterogeneous sequences, nineteen using each of the six task permutations. To facilitate the comparison to the homogeneous sequence results (Figure 3f), we plot in Figure 5b the increase in number of trials to criterion with homogeneous sequences compared to heterogeneous as a baseline. With several exception points, the differences are not notable, suggesting that inter-task effects with heterogeneous sequences are similar to those with homogeneous sequences. Thus, inter-task effects appear to be primarily due to learning to process visual images in general, rather than the specific task-relevant dimensions. The two outlier points in Figure 5b concern the first two episodes: With heterogeneous training, the interference between tasks 1 and 2 nearly vanishes, perhaps because the resources and representations required to perform the two tasks overlap less. One might have predicted just the opposite result, but apparently, extracting information relevant for one dimension does not preclude constructing representations suitable for other dimensions. In fact, the result appears consistent with a finding from human memory—that reducing (semantic) similarity of items reduces interference among them (Baddeley and Dale 1966).

**Task-modulated visual processing**

The architecture that we have experimented with thus far treats the convolutional layers as visual feature extractors, trained end-to-end on task sequences, but the convolutional layers have no explicit information about task; task input is provided only to the final layers of the net. In contrast, processing in human visual cortex can be task modulated (Fias et al. 2002). Perhaps modifying the architecture to provide task information to convolutional layers would reduce inter-task interference. Along the lines of Mozer and Fan (2008), we investigated a modified model using task-modulated visual processing, adopting a simpler approach than most existing architectures for conditional normalization or gated processing (Perez et al. 2018; Chen et al. 2018). We consider task modulation via a task-specific learned bias for each channel in a convolutional layer. As before, task is coded as a one-hot vector. We incorporate connections from the task representation to a convolutional layer (Figure 3), with one bias parameter for the Cartesian product of tasks and channels. This bias parameter is added to the output of each filter in a channel before applying the layer nonlinearity.

We investigated task modulation at each of the four convolutional layers in our model. Because the results of task modulation at the different layers are quite similar (see supplementary material), we report the results of
modulating processing at the first convolutional layer. Figure 6 depicts the results of three Latin square replications, yielding thirty simulations for each dimension, or ninety in total. Introducing task-based modulation allows the model to avoid catastrophic forgetting previously observed from learning the second task on the first, and to a lesser effect, improves performance in the third episode as well. As the model learns additional tasks, and continues retraining on the same tasks, the benefits of task-modulation diminish rapidly (Figure 6b), suggesting the primary benefit is in aiding early learning. We hypothesize that modulating visual processing with the task representation allows the model to learn flexible visual representations that produce less interference.

Comparison to MAML

The results we have presented thus far serve as a baseline against which one can compare any method specialized to reduce forgetting or boost transfer. We conducted comparisons to several such methods, and in this section we report on experiments with model-agnostic meta-learning or MAML (Finn, Abbeel, and Levine 2017). MAML is designed to perform metalearning on a sequence of tasks in order to learn the next task in the sequence more efficiently. However, it is not designed for our continual-learning paradigm, which requires preservation of mastery for previous tasks. We explored two variants of MAML adapted to our paradigm. We report here on the more successful of the two (see supplementary material for details).

Our paradigm is based on a series of 10 episodes where tasks accumulate across episodes. MAML is also trained over a series of episodes, but we make a correspondence between one episode of MAML—the outer loop of the algorithm—and what we will refer to as a micro-episode of our paradigm, which corresponds to a single batch in our original training procedure. Each micro-episode starts with network weights \( w \), and we draw a half-batch of 750 examples (compared to 1500 in the original setting) of which 50% are from the newest task, and the remainder are split evenly across the previous tasks. (For task 1, all examples are from task 1.) From \( w \), we compute a gradient step based on the examples for each task, and apply this step separately to \( w \), yielding \( i \) copies of the weights in episode \( i \), \( \{w_1, ..., w_i\} \), each specialized for its corresponding task. We then draw a second half-batch of 750 examples and perform a metatraining step, as described in MAML. Metatraining involves computing the gradient with respect to \( w \) for the new examples of each task \( k \) based on the weights \( w_k \). Following MAML, we then update \( w \), and proceed to the next micro-episode until our training criterion is attained. Having halved the batch size, we doubled the learning rate from 0.0005 in the original setting to 0.001 for both of MAML’s learning rates. Model details are otherwise identical to the base model.

Over 90 replications (30 per dimensions), we find that the performance of our MAML variant is qualitatively similar to that of our base model (compare Figure 7a and Figure 3a). However, quantitatively, the MAML-based method requires more trials to reach criterion on expectation: Figure 7b shows the relative number of trials to criterion, where negative indicates that MAML is worse than our base model. Apparently the cost of splitting the data and devoting half to meta-training does not outweigh the benefit of meta-training.

Discussion

We explored the behavior of a standard convolutional neural net for classification in a continual-learning setting that introduces tasks sequentially and requires the net to master new tasks while preserving mastery of previously learned tasks. This setting corresponds to that which human learners naturally face as they become domain experts. For example, consider students reading a calculus text chapter by chapter. Early on, engaging with a chapter and its associated exercises results in forgetting of previously mastered material. However, as more knowledge is acquired, students begin to scaffold and link knowledge and eventually are able to integrate the new material with the old. As the final chapters are studied, students have built a strong conceptual framework which facilitates the integration of new material with little disruption of the old. These hypothetical students behave much like the net we studied in this
Training efficiency improves according to a power law. Metalearning (forward facilitation) is observed once the net has acquired sufficient domain expertise. In our paradigm, ‘sufficient expertise’ means having mastered two tasks previously. Metalearning is demonstrated when training efficiency—the number of trials needed to reach criterion—improves with each successive task (Figures 3d,f). Metalearning occurs naturally in the model and does not require specialized mechanisms. Indeed, incorporating a specialized mechanism, MAML, fails to enhance metalearning in our continual-learning paradigm. 

1. Catastrophic forgetting (backward interference) is reduced as the net acquires increasing domain expertise (i.e., as more related tasks are learned). In Figure 3c, compare tasks introduced early (cyan) and late (magenta) in the sequence, matched for number of times they have been trained (position on the abscissa). Retraining efficiency improves for tasks introduced later in the task sequence, indicating a mitigation of forgetting. Note that the number of trials to relearn a skill is less than the number of trials required to initially learn a skill (the exception being task 1 in episode 2). This relearning savings effect has long been identified as a characteristic of human memory (Ebbinghaus 1908), as, of course, has the ubiquity of forgetting, whether due to the passage of time (Lindsey et al. 2014) or to backward interference from new knowledge (Osgood 1948).

2. The potential for catastrophic forgetting (backward interference) is also reduced each time a task is relearned, as indicated by the monotonically decreasing curves in Figure 4, and by the change in forgetting rates in Figure 5. A task that is practiced over multiple episodes receives distributed practice that is interleaved with other tasks. The durability of memory with distributed, interleaved practice is one of the most well studied phenomena in cognitive psychology (Kang et al. 2014; Cepeda et al. 2008; Taylor and Rohrer 2010; Birnbaum et al. 2013).

3. Training efficiency improves according to a power function of the number of tasks learned, controlling for experience on a task (indicated by the linear curve in Figure 3l, plotted in log-log coordinates), and also according to a power function of the amount of training a given task has received, controlling for number of tasks learned (indicated by the linear curves in Figure 3). Power-law learning is a robust characteristic of human skill acquisition, observed on a range of behavioral measures (Newell and Rosenbloom 1980; Donner and Hardy 2015).

4. Forward facilitation and reduction in backward interference is observed only after two or more tasks have been learned. This pattern can be seen by the non-monotonocities in the curves of Figures 3l,f and in the crossover of curves in Figures 3e. Catastrophic forgetting is evidenced primarily for task 1 when task 2 is learned—the canonical case studied in the literature. However, the net becomes more robust as it acquires domain expertise, and eventually the relearning effort becomes negligible (e.g., copper curves in Figure 3b). The anomalous behavior of task 2 is noteworthy, yielding a transition behavior perhaps analogous to the “zero-one-infinity” principle (MacLennan 1999).

5. Catastrophic forgetting in the second episode can be mitigated in two different ways: first, by choosing tasks that rely on different dimensions (Figure 5); and second, by introducing task-based modulation of visual processing (Figure 4). We conjecture that both of these manipulations can be characterized in terms of reducing the similarity of the tasks. In human learning, reducing (semantic) similarity reduces interference (Baddeley and Dale 1966).

We are able to identify these intriguing phenomena because our simulations examined scaling behavior and not just effects of one task on a second—the typical case for studying catastrophic forgetting—or the effects of many tasks on a subsequent task—the typical case for metalearning and few-shot learning. Studying the continuum from the first task to the nth is quite revealing.

We find that learning efficiency improves as more tasks are learned. Although MAML produces no benefit over the standard architecture that served as our baseline, we have yet to explore other methods that are explicitly designed to facilitate transfer and suppress interference (Mishra et al. 2018; Kirkpatrick et al. 2017; Lopez-Paz and Ranzato 2017). The results presented in this article serve as a baseline to assess the benefits of specialized methods. A holy grail of sorts would be to identify methods that achieve backward facilitation, where training on later tasks improves performance on earlier tasks, and compositional generalization (Fodor and Pylyshyn 1988; Fodor and Lepore 2002; Lake and Baroni 2018; Loula, Baroni, and Lake 2018), where learning the interrelationship among earlier tasks allows new tasks to be performed on the first trial. Humans demonstrate the former under rare conditions (Ausubel, Robbins, and Blake 1957; Jacoby, Wahlheim, and Kelley 2015); the latter is common in human behavior, as when individuals are able to perform a task immediately from instruction.

An exciting direction for future research concerns optimizing curricula for continual learning. Our initial approach was inspired by best practices of the science of learning literature (Weinstein, Madan, and Sumeracki 2018). Our hope is that investigations of networks may in turn provide helpful guidance for improving curricula for human learners. Toward this goal, it is encouraging that we observed more than superficial similarities between human and network continual learning.
Supplementary materials

Feature values
The ten colors used in our experiments are: gray, red, blue, green, brown, purple, magenta, yellow, orange, pink. The ten shapes are: cube, sphere, cylinder, pyramid, cone, torus, rectangular box, ellipsoid, octahedron, dodecahedron. And the ten textures are: metal, rubber, chainmail, marble, maze, metal weave, polka dots, rug, bathroom tiles, wooden planks. See additional example images below:

Figure 8: Additional example training images
Results by dimension

To justify our collapsing of the results across dimensions, we provide the results broken down for each individual dimension below. Figure 9 depicts the trials required to reach the accuracy criterion, Figure 9g,h reproducing Figure 3c,d, and the rest of the subfigures offering the results for replications within each dimension. While colors are easier to learn than shapes or textures, simulations in all three dimensions show the same qualitative features. Similarly, Figure 10 depicts the accuracy after a fixed small amount of training, with Figure 10g,h reproducing Figure 3e,f. These results provide further evidence for the ease of learning color compared to the other two dimensions, but the qualitative similarity remains striking.
Figure 9: (a, c, e, g): Number of trials required to reach the accuracy criterion (log scale) as a function of the number of times a given task is trained (also log scale). The colored lines indicate task ordinal position (cyan = introduced in episode 1; magenta = introduced in episode 10). (b, d, f, h): Number of trials required to reach the accuracy criterion (log scale) as a function of the episode number. The colored lines indicate the number of times a task was retrained on (black = 1 time, copper = 10 times). In all panels, the shaded region represents ±1 standard error of the mean.

Figure 10: (a, c, e, g): Accuracy after a fixed amount of training (22,500 trials) as a function of the number of times a given task is trained (log scale). The colored lines indicate task ordinal position (cyan = introduced in episode 1; magenta = introduced in episode 10). (b, d, f, h): Accuracy after the same fixed amount of training as a function of the episode number. The colored lines indicate the number of times a task was retrained on (black = 1 time, copper = 10 times). In all panels, the shaded region represents ±1 standard error of the mean.
Task-modulated processing at different levels

All figures reported below are combined over replications in all three dimensions, where for each modulation level we performed thirty simulations in each dimension, yielding ninety simulations in total for each modulation level. In Figure 11, we provide the results plotted in Figure 5a-b for task-modulation at each convolutional layer (separately). In Figure 12 we provide equivalent plots to Figure 2e-f for the task-modulated models. In Figure 13 we provide equivalent plots to Figure 5c-d for the task-modulated models. The only anomaly we observe is in Figure 13 for task-modulation at the second convolutional layer, where the eight and ninth tasks appear easier to learn for the first time without task-modulation. Save for this anomaly, we observed remarkably consistent results between the different modulation levels, and hence we reported a single one, rather than expanding about all four.
Figure 11: Top panels: Number of trials required to reach the accuracy criterion (log scale) as a function of the number of times a given task is trained (also log scale). The colors indicate task ordinal position (the episode in which a task is introduced; cyan = introduced in episode 1; magenta = introduced in episode 10). Bottom panels: Similar to the top panels, but graphed as a function of episode number with the line colors indicating the number of times a task is retrained (black = 1 time, copper = 10 times).

Figure 12: Top panels: Hold-out accuracy attained after a fixed amount of training (22.5k trials) of a given task, graphed as a function of number of times a given task is trained. As in Figure 11 the colors indicate task ordinal position (the episode in which a task is introduced; cyan = introduced in episode 1; magenta = introduced in episode 10). Bottom panels: Similar to the top panels, but graphed as a function of episode number with the line colors indicating the number of times a task is retrained (black = 1 time, copper = 10 times).
Figure 13: Top panels: Increase in number of trials required to reach accuracy criterion for non-task-modulated versus task-modulated architectures as a function of the number of times a given task is trained (also log scale). The colors indicate task ordinal position (the episode in which a task is introduced; cyan = introduced in episode 1; magenta = introduced in episode 10). Bottom panels: Similar to the top panels, but graphed as a function of episode number with the line colors indicating the number of times a task is retrained (black = 1 time, copper = 10 times).
MAML comparison supplement

We compared our baseline model to two versions of MAML. Both utilized the training procedure we describe under the ‘Comparison to MAML’ section. The first, reported in the middle column below, only utilized this procedure in training, and was tested without the meta-testing step. In other words, this model was tested exactly as our baseline model was tested, to see if MAML manages to learn representations that allow it to answer questions on unseen images without further adaptation. The second version, which we ended up reporting, also utilizes the micro-episode procedure at test time, making train and test identical. The results below demonstrate similar qualitative behavior between our baseline and both versions. However, as the second version, using the meta-testing procedure, fares better, we opt to report it in the submission.

![Graphs comparing baseline, MAML (no meta-test), and MAML (reported)](image)

**Figure 14:** MAML comparison. The left column plots result from our baseline condition. The middle column offers results from a version of MAML which did not follow the micro-episode procedure at test that, that is, did not meta-test. The right column, corresponding to the model reported in the submission, follows the micro-episode procedure we describe at both train and test.
Comparison to simultaneous learning

We performed a systematic comparison between our sequential method of training and the standard supervised learning approach of training on all tasks simultaneously. We know that sequential training is beneficial to humans—every course covers one topic at a time, rather than throwing the entire textbook and mixing all topics from day one. There is also ample evidence for the value of curricular approaches in machine learning, going as far back as [Elman 1993]. However, curricula in machine learning usually attempt to scaffold tasks from smaller to larger, or easier to harder, following some difficulty gradient. Our results in Figure 15 suggest, surprisingly, that randomly chosen sequential curriculum (that is, random task introduction orderings) can significantly speed up learning in some cases. We find, interestingly, that this effect varies by dimension. While in the shape condition the simultaneous learning is competitive with sequential training, we find that in both texture and color sequential training proceeds much faster. In those cases, the number of training trials required to learn each task when trained sequentially (the cyan-to-magenta curves) is far less than the number of trials required to learn each task when trained simultaneously (the red curve). That is, task $n+1$ is learned far faster following tasks $1$–$n$ than simultaneously with tasks $1$–$10$. The long plateau in the color and texture cases appears to suggest some form of initial representation learning which is made more efficient by learning sequentially, rather than simultaneously.

![Figure 15: Simultaneous vs. sequential training. The cyan (first) to magenta (last) colored lines plot the accuracy after some number of training trials for each task the model learned. The average accuracy over all ten tasks, when learned simultaneously, is plotted in red. To make the comparisons valid, the simultaneous training is in the number of training trials for each task, rather than combined for all tasks.](image-url)