Scaling Characteristics of Sequential Multitask Learning: Networks Naturally Learn to Learn

Anonymous Authors[1]

Abstract

We explore the behavior of a standard convolutional neural net in a 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, for example, as an individual reads a textbook chapter-by-chapter. Through simulations involving sequences of 10 related tasks, we find reason for optimism that nets will scale well as they advance from having a single skill to becoming domain experts. We observed 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. Forward facilitation is the goal of research on metalearning, and reduced backward interference is the goal of research on ameliorating catastrophic forgetting. We find that both of these goals are attained simply through broader exposure to a domain.

In a standard supervised learning setting, neural networks are trained to perform a single task, such as classification, defined in terms of a discriminative distribution $p(y \,|\, x, \mathcal{D})$ for labels y conditioned on input x given a data set $\mathcal{D}$. Although such models are useful in engineering applications, they do not reflect the breadth of human intelligence, which depends on the capability to perform arbitrary tasks in a context-dependent manner. *Multitask learning* (Caruana, 1997) is concerned with performing any one of n tasks, usually by having multiple heads on a neural network to produce outputs appropriate for each task, cast formally in terms of the distribution $p(y_i \,|\, x, \mathcal{D}_1, \ldots, \mathcal{D}_n)$, where the subscript denotes a task index and $i \in \{1, \ldots, n\}$ is an arbitrary task. When related, multiple tasks can provide a useful inductive bias to extract shared structure (Caruana, 1993), and as a regularization method to guide toward solutions helpful on a variety of problems (Ruder, 2017).

Multitask learning is typically framed in terms of simultaneous training on all tasks, but humans and artificial agents operating in naturalistic settings more typically tackle tasks sequentially and need to maintain mastery of previously learned tasks as they acquire a new one. Consider students reading a calculus text in which each chapter presents a different method. Early on, engaging with a chapter and its associated exercises will lead to forgetting of the material they had previously mastered. However, as more knowledge is acquired, students learn to effectively scaffold knowledge and eventually are able to leverage prior experience 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. In this article, we study the machine-learning analog of our hypothetical students. *The punch line of the article is that a generic neural network trained sequentially to acquire and maintain mastery of multiple tasks behaves similarly to human learners, exhibiting faster acquisition of new knowledge and less disruption of previously acquired knowledge with diverse domain experience.*

1. Sequential multitask learning

Early research investigating sequential training observed *catastrophic forgetting* (McCloskey & Cohen, 1989), characterized by a dramatic drop in task 1 performance following training on task 2, i.e., the accuracy of the model $p(y_1 \,|\, x, \mathcal{D}_1 \rightarrow \mathcal{D}_2)$ is significantly lower than accuracy of the model $p(y_1 \,|\, x, \mathcal{D}_1)$, where the arrow denotes training sequence. Parisi et al. (2019) review efforts to quantify and reduce catastrophic forgetting, including specialized mechanisms that aim to facilitate sequential learning.

A second line of research exploring sequential training is the active topic of *metalearning*, or learning to learn

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2019 Workshop "Identifying and Understanding Deep Learning Phenomena". Do not distribute.

(Schmidhuber, 1987; Bengio et al., 1991; Thrun, 1996). Metalearning assesses facilitation that arises 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, $p(y_n|x, \mathcal{D}_1 \to \ldots \to \mathcal{D}_n)$, relative to the first task, $p(y_1 \,|\, x, \mathcal{D}_1)$. Some metalearning approaches, such as MAML (Finn et al., 2017) or SNAIL (Mishra et al., 2018) offer mechanisms to encourage transfer between tasks, while other approaches employ recurrence to modify the learning procedure itself (Andrychowicz et al., 2016; Wang et al., 2017).

Catastrophic forgetting and metalearning 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.[1] Whereas catastrophic forgetting has focused on the first task learned, metalearning has focused on the last task learned. We thus view these two topics as endpoints of a continuum. Surprisingly, we are not aware of any work that systematically examines these two topics in conjunction with one another.

To unify the topics, this article examines the continuum from the first task to the n'th. We devised a setting in which we train a model 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 $1 \ldots i-1$. 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 home for extend to human-scale intelligence.

2. 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

[1]In the psychology literature, backward interference is referred to as retroactive interference (Osgood, 1948; Postman, 1961). In the machine learning literature, the more general terms backward and forward transfer are sometimes used (Lopez-Paz & Ranzato, 2017).

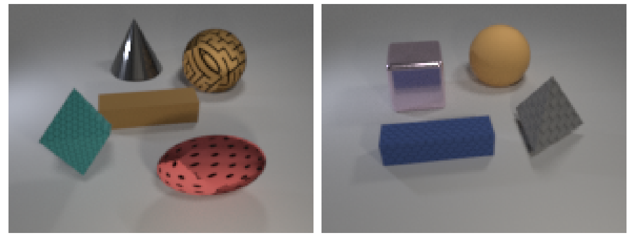

Figure 1: Example training images

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–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. Each epoch of training consists of one pass through each of the training images. These images can be assigned to arbitrary tasks. In each epoch, we roughly balance the number of yes and no target responses for each task. We turn now from this overview to details of the images, tasks, and architecture.

Image generation. We leverage the CLEVR (Johnson et al., 2017) image generation codebase to produce 160×120 pixel color images each with 4 or 5 objects that varied along three *dimensions*: shape, color, and texture. To balance the dimensions, we introduced additional *features* in each dimension to ensure 10 feature values per dimension. 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 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 replications. Because we observed no meaningful differences across task dimensions (see Appendix), the results we report below collapse across dimension.

Architecture. We report experiments using a basic vision architecture with four convolutional layers followed by four fully connected layers. The convolutional layers—

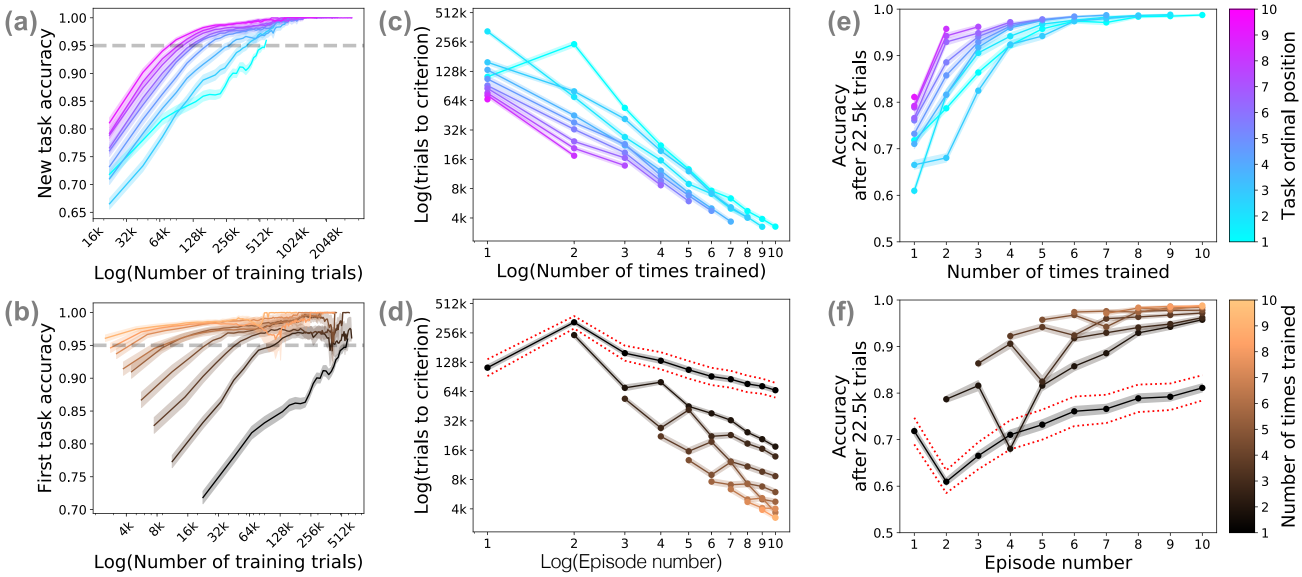

Figure 2: **(a)** Hold-out set accuracy as a function of training trials (log scale) for a newly introduced task. Colored lines indicate task ordinal position (cyan = introduced in episode 1; magenta = introduced in episode 10). In all panels, the shaded region represents ± 1 standard error of the mean. **(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 graphed as a function of episode number with the line colors indicating—as in (b)—the number of times a task is retrained. **(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 graphed as a function of episode number with the line colors indicating—as in (b)—the number of times a task is retrained.

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. 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, task is specified as a component of the input. Similar to Sort-of-CLEVR (Santoro et al., 2017), task is coded as a one-hot input vector. Task representation is concatenated to the output of the last convolutional layer before passing it to the first fully-connected layer.

3. Results

Figure 2a 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 $i > 2$. Figure 2b shows the accuracy of the task introduced in the first episode (y_1) as it is retrained each episode.[2] Not surprisingly, task accuracy improves monotonically with the number of times trained, indicating a relearning savings. But notably, the catastrophic forgetting present in early episodes vanishes by the tenth episode.

To analyze our simulations more systematically, we remind the reader that the simulation sequence presents fifty-five opportunities to assess learning: the task introduced 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. Figures 2c,d provide two views on the amount of training to reach an accuracy criterion of 95%—the dashed line in Figures 2a,b. The data are plotted either as a function of the number of times a task is retrained (Figure 2c) or as a function of the episode number (Figure 2d), with the curves color coded as in Figures 2a,b. The roughly log-log linear curves offer evidence of power-law decrease in the retraining effort required to reach criterion. (We discuss the exception points shortly.) Backward interference diminishes both as a function of the number of times a task is relearned (Figure 2c) and the amount of domain experi-

[2] The misalignment of the first point is due to the fact that the accuracy is assessed at the end of a training epoch, and each successive episode has fewer trials of task y_1 per epoch.

ence, as indexed by the episode number (Figure 2d). Figures 2e,f show an alternative view of backward interference by plotting accuracy after a fixed amount of retraining. The conditions that require the least number of trials to criterion (Figures 2c,d) also achieve the highest accuracy after a small amount of training (Figures 2e,f).

To examine forward facilitation, we focus on the newest task introduced, the highlighted curve in Figures 2d,f. Starting at the third episode, we observe forward facilitation, evidenced by both a reduced number of examples required to learn the new task, as well as higher accuracy after a fixed amount of training. Similar forward facilitation occurs not just for the newest tasks, but even for relearning older tasks, as reflected in the black-to-copper curves.

Figure 2 reveals an anomaly in the second episode. No forward facilitation is observed for the new task—as indicated by the rise in the highlighted curve in Figure 2d—and strong backward interference is observed for the old task—as indicated by the crossover of the cyan curve in Figures 2c,e. This finding suggests 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.

4. Discussion

We explored the behavior of a standard convolutional neural net for classification tasks in a 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 face as they become experts in a domain, for example, as they read a textbook chapter by chapter. Our network exhibits six interesting properties:

1. Forward facilitation is observed once the net has acquired sufficient expertise in the domain, as evidenced by requiring less training to learn new tasks as a function of the number of related tasks learned (see highlighted black curve in Figures 2d,f).
2. Backward interference is reduced as a function of the number of related tasks previously learned (compare magenta-to-cyan curves in Figures 2c,e for a given position on the abscissa).
3. Forward facilitation occurs and backward interference is reduced only after two or more tasks have been learned. This pattern can be seen by the nonmonotonicities in the highlighted curves of Figures 2d,f and in the crossover of curves in Figures 2c,e.
4. Backward interference is also reduced as a function of the number of times a task is relearned, controlling for the total number of tasks learned. This phenomenon is demonstrated by the ordering of the black-to-copper curves in Figures 2d,f for a given position along the abscissa. This reduction in backward interference has long been identified in human learning, where it is known as the *saving effect* (Ebbinghaus, 1908/1973).
5. Training performance improves according to a power function of the number of tasks learned, controlling for experience on a task (the slope of the curves in Figure 2d), and also according to a power function of the amount of training a given task has received, controlling for number of tasks learned (the slope of the curves in Figure 2c). Power-law learning is a robust characteristic of human skill acquisition, observed on a range of behavioral measures (Newell & Rosenbloom, 1980; Donner & Hardy, 2015).
6. Catastrophic forgetting is evidenced primarily for task 1 when task 2 is learned—the canonical case studied in the literature. However, the model becomes more robust as it acquires sufficient domain experience, and eventually the relearning effort becomes negligible (see copper curves in Figures 2b,d,f). The anomalous behavior of task 2 is noteworthy, yielding a transition behavior that is perhaps analogous to the “zero one infinity” rule coined by Willem van der Poel.

We are able to identify these interesting 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 entire continuum from the first task to the n'th is quite revealing.

We found strong evidence for improved learning performance with broader domain expertise, and further investigation is merited. We are beginning investigations that examine how similar tasks must be to facilitate one another: how does scaling behavior change when the tasks dimensions switch across successive episodes (e.g., from color to shape to texture)? Our preliminary results suggest that the domain knowledge acquired is quite general and extends to other dimensions of the images. We are also examining the scaling properties of metalearning methods that are explicitly designed to facilitate transfer. The results presented in this article can serve as a baseline to measure the magnitude of facilitation that the specialized methods offer. A holy grail of sorts would be to identify methods that demonstrate *backward facilitation*, where training on later tasks improves performance on earlier tasks, and *compositional generalization* (Fodor & Pylyshyn, 1988; Fodor & Lepore, 2002; Lake & Baroni, 2018; Loula et al., 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 et al., 1957; Jacoby et al., 2015); the latter is common in human behavior, as when individuals are able to perform a task immediately from instruction.

References

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and de Freitas, N. Learning to learn by gradient descent by gradient descent. *Advances in Neural Information Processing Systems 29*, pp. 3981—-3989, jun 2016. URL `http://arxiv.org/abs/1606.04474`.

Ausubel, D. P., Robbins, L. C., and Blake, Elias, J. Retroactive inhibition and facilitation in the learning of school materials. *Journal of Educational Psychology*, 48(6):334–343, 1957. ISSN 0022-0663. doi: 10.1037/h0043524. URL `http://content.apa.org/journals/edu/48/6/334`.

Bailey, R. *Design of comparative experiments*. Cambridge University Press, 1st edition, 2008. ISBN 9780511611483.

Bengio, Y., Bengio, S., and Cloutier, J. Learning a synaptic learning rule. In *Seattle International Joint Conference on Neural Networks*, volume ii, pp. 969. IEEE, 1991. ISBN 0-7803-0164-1. doi: 10.1109/IJCNN.1991.155621. URL `http://ieeexplore.ieee.org/document/155621/`.

Caruana, R. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 41–48. Morgan Kaufmann, 1993.

Caruana, R. Multitask Learning. *Machine Learning*, 28: 41–75, 1997.

Donner, Y. and Hardy, J. L. Piecewise power laws in individual learning curves. *Psychonomic Bulletin & Review*, 22(5):1308–1319, oct 2015. ISSN 1069-9384. doi: 10.3758/s13423-015-0811-x. URL `http://link.springer.com/10.3758/s13423-015-0811-x`.

Ebbinghaus, H. *Psychology: An elementary textbook*. Arno Press, 1908/1973.

Finn, C., Abbeel, P., and Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *Proceedings of the 34th International Conference on Machine Learning,*, 70, 2017. URL `https://arxiv.org/pdf/1703.03400.pdf`.

Fodor, J. A. and Lepore, E. *Compositionality Papers*. Oxford University Press UK, 2002.

Fodor, J. A. and Pylyshyn, Z. W. Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28 (1-2):3–71, mar 1988. ISSN 0010-0277. URL `http://www.ncbi.nlm.nih.gov/pubmed/2450716`.

Jacoby, L. L., Wahlheim, C. N., and Kelley, C. M. Memory consequences of looking back to notice change: Retroactive and proactive facilitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(5):1282–1297, sep 2015. ISSN 1939-1285. doi: 10.1037/xlm0000123. URL `http://www.ncbi.nlm.nih.gov/pubmed/26010827http://doi.apa.org/getdoi.cfm?doi=10.1037/xlm0000123`.

Johnson, J., Fei-Fei, L., Hariharan, B., Zitnick, C. L., Van Der Maaten, L., and Girshick, R. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. ISBN 1612.06890v1. URL `https://arxiv.org/pdf/1612.06890.pdf`.

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/lake18a.html`.

Lopez-Paz, D. and Ranzato, M. . A. Gradient Episodic Memory for Continual Learning. In *NIPS*, 2017. URL `https://github.com/`.

Loula, J., Baroni, M., and Lake, B. Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks. jul 2018. URL `https://arxiv.org/abs/1807.07545`.

McCloskey, M. and Cohen, N. J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, 24:109–165, jan 1989. ISSN 0079-7421. doi: 10.1016/S0079-7421(08)60536-8. URL `https://www-sciencedirect-com.ccl.idm.oclc.org/science/article/pii/S0079742108605368`.

Mishra, N., Rohaninejad, M., Chen, X., and Abbeel, P. A Simple Neural Attentive Meta-Learner. In *International Conference on Learning Representations*, 2018. URL `https://arxiv.org/pdf/1707.03141.pdf`.

Newell, A. and Rosenbloom, P. S. Mechanisms of Skill Acquisition and the Law of Practice. In Anderson, J. R. (ed.), *Cognitive Skills and their Acquisition*. Erlbaum, hillsdale, edition, 1980. URL `https://apps.dtic.mil/docs/citations/ADA096527`.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational Continual Learning. *ICLR*, oct 2018. URL http://arxiv.org/abs/1710.10628.

Osgood, C. E. An investigation into the causes of retroactive interference. *Journal of Experimental Psychology*, 38(2):132–154, 1948. ISSN 0022-1015. doi: 10.1037/h0055753. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0055753.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, may 2019. doi: 10.1016/J.NEUNET.2019.01.012. URL https://www-sciencedirect-com.ccl.idm.oclc.org/science/article/pii/S0893608019300231https://arxiv.org/pdf/1802.07569.pdf.

Postman, L. The Present Status of Interference Theory. In *Verbal learning and verbal behavior.*, pp. 152–196. McGraw-Hill Book Company, New York, 1961. doi: 10.1037/11182-007. URL http://content.apa.org/books/11182-007.

Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. jun 2017. URL https://arxiv.org/abs/1807.07545.

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T., and London, D. A simple neural network module for relational reasoning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. ., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4967—-4976. Curran Associates, Inc., 2017. URL https://arxiv.org/pdf/1706.01427.pdf.

Schmidhuber, J. *Evolutionary Principles in Self-Referential Learning*. PhD thesis, TU Munich, 1987. URL http://people.idsia.ch/{~}juergen/diploma1987ocr.pdf.

Thrun, S. Is Learning The n-th Thing Any Easier Than Learning The First? In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E. (eds.), *Advances in Neural Information Processing Systems 8*, pp. 640—-646. MIT Press, 1996. URL https://papers.nips.cc/paper/1034-is-learning-the-n-th-thing-any-easier-than-learning-the-first.pdf.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. Learning to Reinforcement Learn. In *CogSci*. DeepMind, 2017. URL https://arxiv.org/pdf/1611.05763v3.pdf.