Convolutional Neural Networks Can (Meta-)Learn the Same-Different Relation

Max Gupta[1], **Sunayana Rane**[1], **R. Thomas McCoy**[2], **Thomas L. Griffiths**[1,3]
mg7411@princeton.edu, srane@princeton.edu, tom.mccoy@yale.edu, tomg@princeton.edu
[1]Department of Computer Science, Princeton University
[2]Department of Linguistics and Wu Tsai Institute, Yale University
[3]Department of Psychology, Princeton University

Abstract

While convolutional neural networks (CNNs) have come to match and exceed human performance in many settings, the tasks these models optimize for are largely constrained to the level of individual objects, such as classification and captioning. Humans remain vastly superior to CNNs in visual tasks involving relations, including the ability to identify two objects as 'same' or 'different'. A number of studies have shown that while CNNs can be coaxed into learning the same-different relation in some settings, they tend to generalize poorly to other instances of this relation. In this work we show that the same CNN architectures that fail to generalize the same-different relation with conventional training are able to succeed when trained via meta-learning, which explicitly encourages abstraction and generalization across tasks.

Keywords: Meta-Learning; Relational Reasoning; Similarity

Introduction

Debates about what aspects of human learning can be captured by artificial neural networks have played a prominent role in the history of cognitive science (Minsky & Papert, 1969; Rumelhart & McClelland, 1986; Pinker & Prince, 1988; Fodor & Pylyshyn, 1988). One recent manifestation of this question has focused on the learning capacities of convolutional neural networks (CNNs), which are widely used in computer vision and have been used to capture aspects of human behavioral and neural responses in object recognition tasks (Kubilius et al., 2019; Peterson, Abbott, & Griffiths, 2018). Despite their strong performance in capturing the features of objects, these models seem to fall short when tasked with learning about relations between objects. A growing body of research highlights these limitations, particularly in tasks requiring abstract relational reasoning, such as same-different classification (Fleuret et al., 2011; Kim, Ricci, & Serre, 2018; Puebla & Bowers, 2022).

The human capacity for abstraction is based on understanding the laws that govern relations. The most basic of these abilities, arguably a precursor to more complex abstract reasoning, is the same-different relation: the ability to tell if two objects are the same or not. Extensive work in cognitive science dating back to the 1980's (Premack, 1983) has shown that this ability develops early on in human childhood (Blöte, Resing, Mazer, & Van Noort, 1999), is associated with the learning of language (Lupker, Nakayama, & Perea, 2015), and extends far and wide in the animal kingdom, from bees to ducklings and chimpanzees (Gentner, Shao, Simms, & Hespos, 2021). However, learning the same-different relation has proven surprisingly difficult for artificial neural networks. Early convolutional neural networks were shown to have difficulty on the same-different task in Kim et al. (2018). Subsequent work has suggested that CNNs can learn some forms of the same-different relation given a well-structured training regime and near-distribution testing regime (Puebla & Bowers, 2022; Geiger, Carstensen, Frank, & Potts, 2020), though they still struggle with true out-of-distribution generalization. These results have led to the tentative conclusion that CNNs may lack the inductive biases needed to learn abstract relational information.

These negative results do not mean that neural networks are incapable of representing the same-different relation. As models have become larger and more advanced, we have very recently begun to see generalizable understanding of same-different relations emerging in state of the art vision-transformer models pre-trained on ImageNet with methods such as contrastive learning (Dosovitskiy, 2020; Tartaglini et al., 2023). The observation that some very large neural networks are able to learn this relation motivates re-investigating whether shallower CNNs have this capacity.

Previous attempts to train CNNs on the same-different relation have used standard techniques for training neural networks, in which a task (or set of tasks) is defined and

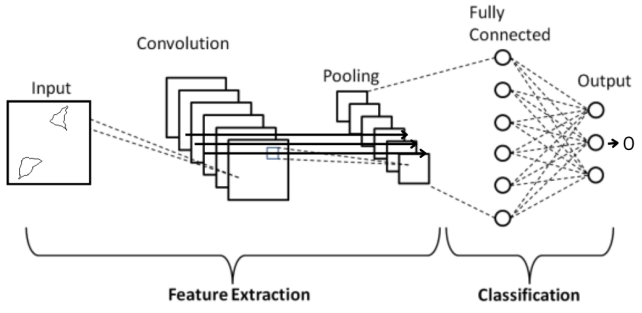

Figure 1: Example of a same-different task as it is posed to a convolutional neural network (CNN) at test-time. Given an image containing two objects, the CNN should return a label of 1 if the two objects are the same or a label of 0 if they are different (as in the example input to the left of the figure)

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

the weights of the network are optimized to perform that task (or set of tasks). Recently, researchers have begun to explore a different technique for training neural networks, known as meta-learning (e.g., Finn, Abbeel, & Levine, 2017). Using this technique, a set of neural networks are each trained to perform a different task, and the shared initial weights of those neural networks are optimized to increase the performance of all networks across all tasks. The learned initial weights encode the shared structure of the different tasks, making it easier for the individual networks to learn to perform those tasks. Meta-learning has been shown to allow simple neural networks to quickly learn to perform tasks that previously were assumed to require symbolic representations, such as learning formal languages (e.g., McCoy & Griffiths, 2023). In this paper, we explore whether this tendency to find generalizable abstractions is sufficient to allow CNNs to learn the same-different relation.

Background

Convolutional neural networks

Convolutional neural networks are a type of multi-layered artificial neural network that takes pixel-level visual data as input (LeCun, Bengio, & Hinton, 2015). Key components of the CNN architecture take inspiration from biological visual systems (Hubel, Wiesel, et al., 1959). The initial layers of the network learn filters that are applied across an image, and their outputs are spatially pooled to form representations that are translation-invariant and expressed at different scales. The learned filters detect features, such as edges, which are useful for image classification and related tasks. CNNs first came to prominence for their remarkable image classification ability (Krizhevsky, Sutskever, & Hinton, 2012), and have since matched or surpassed human performance on a variety of visual tasks (Alzubaidi et al., 2021).

Learning the same-different relation

Various forms of CNNs have been trained and tested on relational visual tasks. A common dataset used for training and evaluation is the Synthetic Visual Reasoning Test (SVRT) dataset, a battery of 23 different visual-relation tasks (Fleuret et al., 2011). In early experiments, CNN architectures that were very successful in image classification tasks were largely unsuccessful on the visual-relation tasks in the SVRT dataset (Stabinger, Rodríguez-Sánchez, & Piater, 2016). This raised the question of whether CNNs lack the human-like inductive biases necessary for relational reasoning.

As CNN architectures improved and showed heightened performance on computer vision problems, further studies investigated whether these more sophisticated architectures were able to solve visual relation tasks such as same-different. A study using a CNN architecture with increased multi-layered attention mechanisms, for example, showed that the more sophisticated architecture significantly improved performance on a range of relational classification tasks (Wang, Cao, De Melo, & Liu, 2016). This suggested that, under certain circumstances, forms of the CNN architecture might be capable of achieving strong performance on relational tasks. However, further studies on same-different tasks in particular yielded mixed results. A recent study investigated various CNN architectures on same-different tasks and found that while the networks could perform well on same-different tasks that were similar to the tasks in their training data, their performance dropped significantly when tested on another family of same-different tasks that were substantially different from those in the training data (although the abstract visual relation tested was, of course, the same) (Puebla & Bowers, 2022). This outcome was also true for larger, deeper, more sophisticated CNN architectures such as ResNets, leading to the conclusion that abstract same-different relations were difficult or impossible for CNNs to learn in a generalizable manner.

Meta-learning

While previous work has suggested that these results indicate that CNNs may not have the architectural inductive biases needed to robustly learn abstract visual relations such as same-different, another source of relevant inductive biases is the training paradigm that is used. All evaluations of CNNs in previous work have learned a set of weights by training models to perform a single task or set of tasks simultaneously. In this paper we use a different approach: meta-learning. In particular, we focus on the Model-Agnostic Meta-Learning (MAML) algorithm (Finn et al., 2017), which is designed to find the optimal starting point in weight space for a set of related tasks, such that the model can rapidly generalize to new, unseen tasks.

Given a set of tasks $\mathcal{T}$ such that each task $t \in \mathcal{T}$ has an associated loss function L_t, conventional neural network training seeks a set of weights for a neural network ϕ that minimizes the loss function

$$\mathcal{L}_{\text{conventional}} = \sum_{t \in \mathcal{T}} L_t(\phi) \tag{1}$$

which is simply the sum of the losses across different tasks. By contrast, MAML seeks to find the initial weights θ that minimize the loss function

$$\mathcal{L}_{\text{MAML}} = \sum_{t \in \mathcal{T}} L_t(\phi_t) \qquad \text{for } \phi_t = \theta - \alpha \nabla L_t(\theta) \tag{2}$$

where ϕ_t are a set of weights adapted for performing task t via gradient descent applied to the loss L_t of task t starting at the initial weights θ (with α being a learning rate). The resulting θ should capture the regularities shared by the tasks in $\mathcal{T}$, supporting abstraction and generalization (see Figure 2).

In this work, we explore whether meta-learning allows convolutional neural networks to form generalizable representations of the same-different relation, challenging previous accounts suggesting certain neural architectures lack the capacity to capture same-different reasoning (Kim et al., 2018; Puebla & Bowers, 2022). We replicate previous

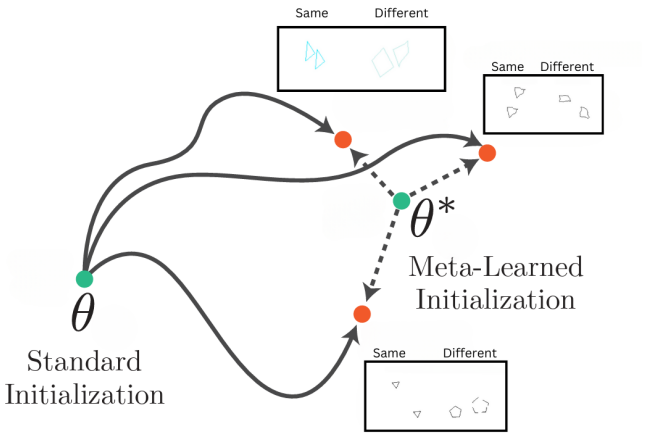

Figure 2: Meta-learning initial weights for generalization. A neural network with a standard initialization (θ) typically requires a large amount of training to learn a specific task. Meta-learning optimizes the network's initialization to create a meta-learned initialization θ^* from which a range of different tasks can be learned with a small amount of training. In our setting, we use meta-learning not for its typical purpose of enabling the rapid learning of many tasks but rather as a way to encourage abstraction.

studies testing CNNs of varying depths and switch the conventional training regime to one based on meta-learning, looking for the emergence of a reliable generalization of same-different understanding across novel stimuli and in novel tasks. Importantly, the networks that we train are exposed to exactly the same training data regardless of whether the networks are optimized using conventional training or meta-learning; the only difference is the type of optimization algorithm that is applied to the data. Matching the data allows us to isolate the effect of this algorithm.

Replicating Previous Work

Throughout all experiments, we replicate the same CNN architectures evaluated by Kim et al. (2018), who tested performance on the SVRT dataset for CNNs of varying depths and convolutional filter sizes. In addition to the original Problem 1 from the SVRT challenge (pictured in Figure 1) we also train on 9 same-different tasks created by Puebla and Bowers (2022), which augment the standard SVRT dataset with new shapes such as arrows, irregular polygons, and shapes with random colors (the full set is shown in Figure 3). In the standard learning setting, same-different tasks are processed as individual input/label pairs: an image and a corresponding 0/1 label for different/same (Figure 1 illustrates this at test-time, where the label is withheld). Note that, in these datasets, what constitutes a distinct "task" is a particular type of shape over which same-different judgments must be made. For instance, one task is based on irregular polygons while another is based on regular polygons. Thus, all tasks target the same abstract relation (same-different), but they instantiate this relation with different types of shapes.

To establish a baseline, we first evaluate the performance of three CNN architectures in this setting, training each model end-to-end on all 10 distinct tasks, with equal frequency for each task and for the categories of *same* and *different*. We test the CNN architectures used by Kim et al. with 2, 4, and 6 convolutional layers using max pooling, batch normalization, and ReLU activation, followed by 3 fully connected layers of 1024 units each and a 2-dimensional classification layer. All models are trained with Adam optimization (Kingma & Ba, 2015) and a base learning rate of 1e-3.

We then test each model on unseen same-different examples from the 10 tasks the model has been trained on, as a basic in-distribution test of learning these tasks. Averaging over 10 seeds run to convergence, we find performance stabilizing almost exactly at the level of random guessing for all three model depths (Figure 4, left); since there are two possible labels (*same* and *different*), random guessing would yield an accuracy of 50%. As an exception to this general trend, some conditions do manage to reliably converge to high-accuracy solutions (in particular, a 2-layer CNN achieves 99 percent test-time accuracy on the scrambled task and 80 percent accuracy on the lines task). A potential explanation is that the scrambled and lines tasks are the only two tasks in the dataset that feature only straight, right-angled lines, a low-level feature that develops earlier in shallower networks, but may be overlooked in deeper networks.

Meta-learning Same-Different In-Distribution

Having established the performance of a class of CNNs on the augmented same-different dataset, we move to formalizing the meta-learning setup of the task. Without changing the model architecture or the content or quantity of data, we change the learning algorithm from standard stochastic gradient descent to MAML in order to explore the impact of using meta-learning.

In the meta-learning setting, we generate 'episodes' from each task consisting of labeled support sets for task-specific adaptation and query sets with held-out labels for evaluation. A set of examples is sampled and then randomly partitioned into a support set and a query set. The meta-learner then has a chance to 'practice' on the episode's support set before outputting predictions for each query image. Crucially, each episode contains examples from exactly one task. Within-episode learning constitutes the 'inner loop' of standard learning from the support set – the gradient step taken from the initial weights θ – for which we provide a 'fast' inner learning rate of 1e-2. To aid the model in learning from different episode structures, we use variable support set sizes of evenly distributed same/different examples (4, 6, 8, and 10 examples) and fixed query sizes (3 examples, always).

The outer loop consists of the transfer of learning across task-specific episodes and is associated with the more gradual transfer of knowledge between tasks, for which we assign a

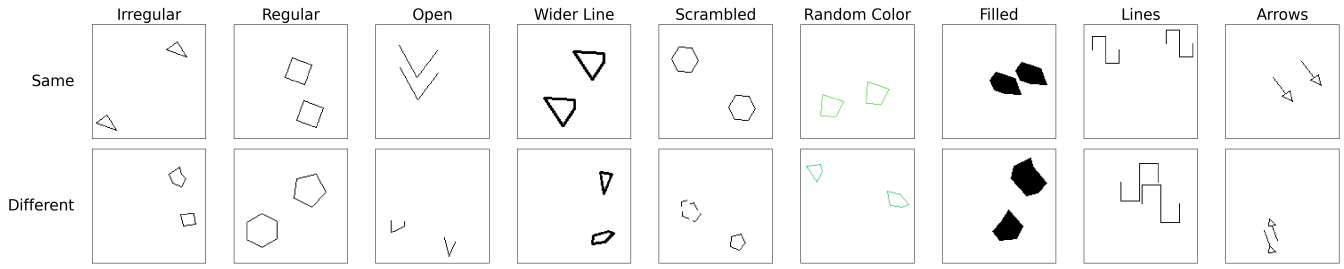

Figure 3: The Same-Different dataset from Puebla and Bowers (2022). Each column shows one of the nine tasks in this dataset, where all tasks are based around the same-different relation but use different types of shapes to instantiate that relation. Each task has a stochastic function generator that ensures each example is unique within any given dataset.

'slower' learning rate of 1e-4 and use Adam optimization.

By using this outer-loop update which optimizes for generalization across examples, we intend to create pressure for the learner to go beyond example-specific properties toward a more generalizable notion of sameness/difference (see Discussion). This generality is further encouraged by sampling episodes alternating across task types and support sizes. We ensure that our meta-learning models (described in this section) and our vanilla models (described in the previous section) receive exactly the same training data as follows: We first generate the dataset for the meta-learning setting as described in the previous paragraphs, sampling a support set and a query set for each of the 1000 episodes that we produce for each of the 10 tasks. The training data used for the vanilla models in the previous section are then created by 'flattening' this meta-learning dataset – that is, the vanilla training set is the concatenation of all support sets and all query sets from all episodes in the meta-learning training set.

As a first step, we test the three previously described CNN architectures in our meta-learning setup by meta-training on 1000 episodes from each of the 10 tasks and testing on unseen episodes from the same tasks. A 2-layer CNN performs at chance on the majority of tasks seen in-distribution, but we see a striking increase in performance as we increase convolutional depth (Figure 4, right). A CNN with 6 convolutional layers performs at almost perfect accuracy across all tasks it has been meta-trained on, and its accuracy is consistent across seeds.[1]

This suggests that these deeper networks are better able to respond to the pressure for abstraction that we intend to create via meta-learning. The bias-variance tradeoff formalizes this observation: deeper networks have more tunable parameters, resulting in less bias and an increased ability to represent variance in the dataset. However, if a neural network can perform well on data sampled in-distribution (as shown here), one possible explanation is that the network has simply learned to memorize shallow properties of its data distribution, and not to generalize the abstract notion of sameness and difference—a concern that motivates the analyses in the next section.

Meta-learning Same-Different Out-of-Distribution: Leave-One-Out

In this section, we aim to further investigate the ability of meta-trained CNNs to capture a generalizable notion of sameness/difference by testing on unseen, out-of-distribution tasks. To set this up, we perform a leave-one-out test, training the previously highest performing model (the 6-layer CNN) on the same battery of same-different classification problems, but crucially holding out one task from training for testing. In this way, we systematically test the model on out-of-distribution tasks it has never seen during training. We do this for all tasks and meta-train to convergence with MAML, using the same parameters described above.

The results are shown in Figure 5. We intentionally replicate the structure of this experiment from Puebla and Bowers (2022), who found that even much larger, pre-trained ResNet architectures were unable to reliably generalize out-of-distribution on this set of tasks using standard learning techniques. By using this setup in a meta-learning context, we see that even much shallower CNNs without pre-training can reliably generalize to even the most challenging OOD tasks.

Classification accuracies were at or about 95 percent in all tasks except for three (arrows, lines, and scrambled). These three tasks were also found to be hardest for ResNets in the analysis performed by Puebla and Bowers (2022). In our case, however, meta-learning allows even these shallower CNNs to outperform ResNets in each of these 'failure' cases by significant margins, and with much less data. For example in the worst performing task ('lines'), previous results showed ResNet50 and ResNet152 performing at chance accuracy, whereas we show a meta-trained CNN consistently performs above chance (where chance level is 50%). [2]

[1]For full results, see Appendix A here: https://arxiv.org/abs/2503.23212

[2]Although Puebla and Bowers (2022) did not release exact details on data quantity during training, ImageNet pretraining allows the models to see over 1 million images prior to fine-tuning on the same-different task.

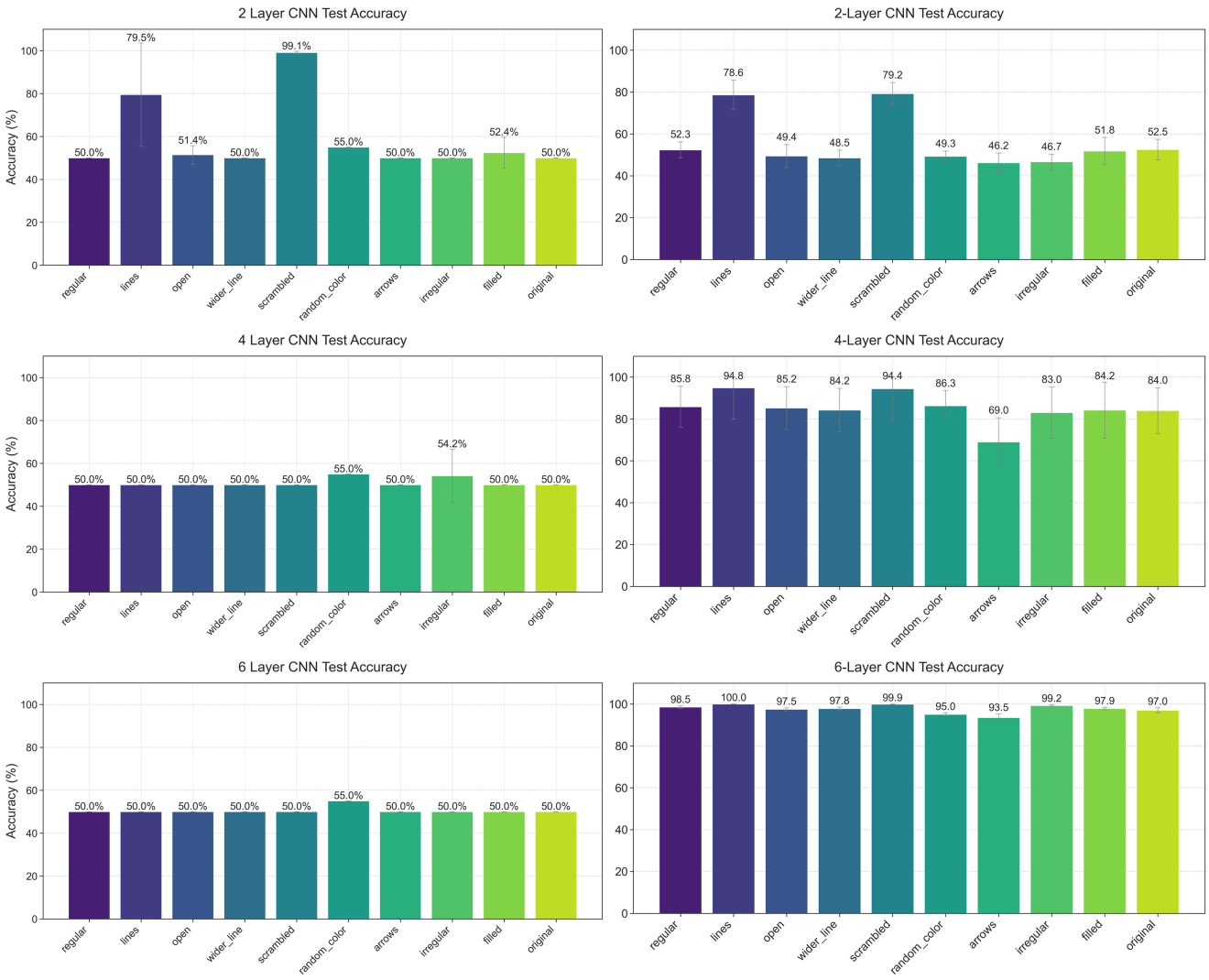

Figure 4: In-distribution same-different classification accuracy for a vanilla-learner (left) versus a meta-learner (right) by task and architecture. The vanilla learner is trained using standard gradient descent, while the meta-learner is trained with the algorithm MAML in a way that is intended to encourage abstraction. Each bar is one version of the same-different task, where the task versions differ in terms of what types of shapes are used to illustrate same-different relations. These evaluations are in-distribution because the models' training data always contained examples of the same type being evaluated on. "Original" represents task #1 from the SVRT dataset. Error bars represent standard deviations from mean accuracy across 10 randomly initialized seeds.

Discussion

By using a training algorithm based on meta-learning, we have shown that CNNs can successfully learn the same-different relation even though they struggle to do so when they are trained with more standard optimization approaches. CNNs trained via meta-learning can perform same-different classification with high accuracy (exceeding 95% in most conditions) even when the input is based on types of images that never appeared in the meta-training data, indicating that these networks have internalized a version of same-different relations that is abstract enough to generalize to new types of shapes. While much work has recently focused on novel architectures dedicated to improving relational reasoning in vision models (Webb, Sinha, & Cohen, 2021; Kerg et al., 2022; Altabaa, Webb, Cohen, & Lafferty, 2024; Webb et al., 2024), our results demonstrate that an alternative pathway toward enhanced relational reasoning is via the nature of the training algorithm: optimizing a standard CNN (without architectural modifications) using meta-learning rather than standard learning.

What is it about meta-learning that leads to the enhanced same-different reasoning we have observed? One possible explanation is that it may provide an incentive for abstraction. To understand this point, it is useful to compare meta-learning

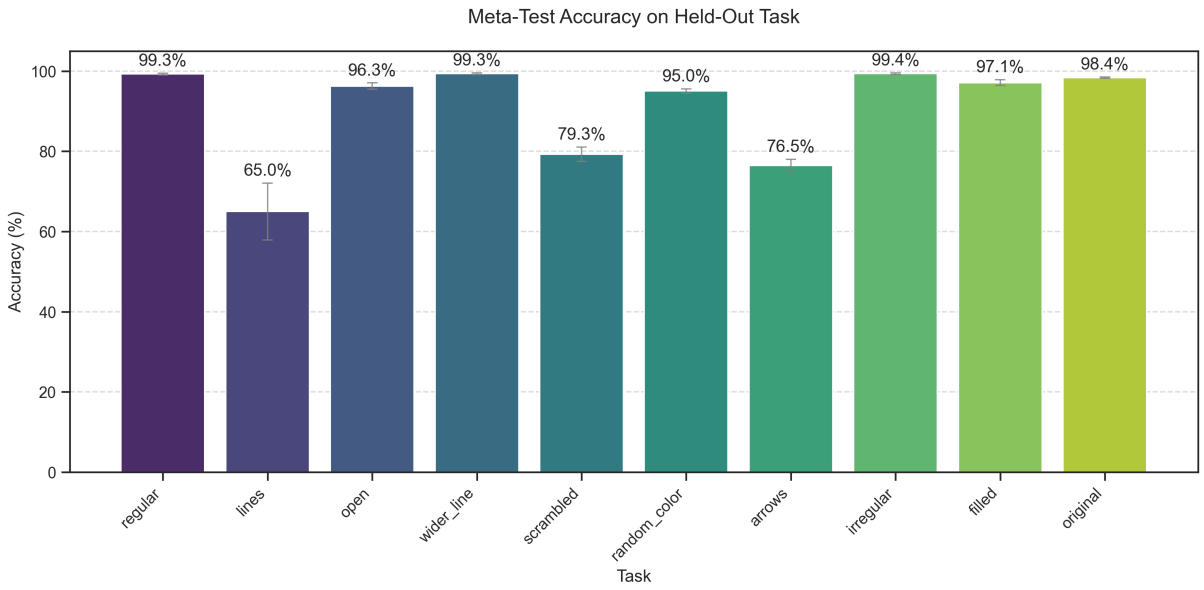

Figure 5: Out-of-distribution same-different classification accuracy for a CNN trained using meta-learning. Each bar shows one version of the same-different task, where the versions differ in terms of the types of shapes used to illustrate same-different relations. A separate CNN was meta-trained for each bar, where the meta-training process included all tasks except the one to be evaluated on, and then evaluated on that withheld task. Thus, the task being evaluated on was always out-of-distribution.

to standard learning. In standard learning, a network is shown a batch of examples, and its weights are then adjusted such that, if it were to process those same examples again, it would achieve a smaller error on them. In contrast, in meta-learning, each batch of examples (i.e., each episode) has two parts: the support set and the query set. A copy of the network is trained on the support set and evaluated on the query set, and the original network's weights are then adjusted such that, if it were trained again on the support set, its performance on the query set would improve. Thus, meta-learning incentivizes the model to be able to learn from one set of examples (the support set) in a way that is useful for processing a different set of examples (the query set). This pressure for generalizing from one set of examples to another might facilitate abstraction because low-level features (e.g., the angles of particular shapes) are unlikely to be broadly useful, whereas more abstract features (e.g., same-different information) will have more general utility.

These results add to a growing body of evidence that meta-learning can enable neural networks to overcome some of their most notorious limitations (Irie & Lake, 2024). For instance, meta-learning can increase neural network abilities in few-shot learning (Hochreiter, Younger, & Conwell, 2001; McCoy & Griffiths, 2023), compositionality (Lake & Baroni, 2023), and out-of-distribution generalization (Finn et al., 2017). However, the nature of the advantage that meta-learning provides is different in our work than in previous work. Traditionally, meta-learning serves as a targeted weight-initialization method: it is used to identify a starting point from which a network can efficiently learn many different tasks. This starting point encodes inductive biases that have been acquired through the meta-learning process and that enable the network to subsequently learn and generalize more effectively. In contrast, our use of meta-learning functions more as a training algorithm than as a weight initialization algorithm. Rather than having a network meta-learn from episodes that each instantiate a different task, we have networks meta-learn from episodes that all instantiate the same task (same-different classification) but with variation in the types of inputs that are used. Therefore, rather than instilling the ability to learn many different tasks (as is more typical in meta-learning), our usage modifies the way in which the network learns a single task. Specifically, as argued in the previous paragraph, the goal in our usage of meta-learning is to optimize for cross-example generalization in a way that facilitates abstraction. Our results suggest that the approach has indeed had this effect.

Modern computer vision has advanced largely through optimizing classification objectives for individual objects, but what can be left out in the process is the rich 'invisible' space between objects that humans seem to make sense of effortlessly: how individual objects relate to one another. By using meta-learning to train a neural network on same-different tasks, this work provides a path to imbue stronger pressures for neural networks to reason relationally across novel stimuli, using the higher-order nature of the gradient updates in meta-learning to encourage the development of more abstract relational information inside neural networks.

Acknowledgments

MG acknowledges support from the Princeton AI Teaching Fellowship. We also gratefully acknowledge the computational resources provided by the Della high-performance computing cluster at Princeton University, which were essential for meta-training the models presented in this study. This work was supported by grant N00014-23-1-2510 from the Office of Naval Research.

References

Altabaa, A., Webb, T. W., Cohen, J. D., & Lafferty, J. (2024). Abstractors and relational cross-attention: An inductive bias for explicit relational reasoning in transformers. In *The Twelfth International Conference on Learning Representations.*

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*, 1–74.

Blöte, A. W., Resing, W. C., Mazer, P., & Van Noort, D. A. (1999). Young children's organizational strategies on a same–different task: A microgenetic study and a training study. *Journal of Experimental Child Psychology*, *74*(1), 21–43.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dosovitskiy, A. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning* (pp. 1126–1135).

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, *108*(43), 17621–17625.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2020). Relational reasoning and generalization using non-symbolic neural networks. *arXiv preprint arXiv:2006.07968*. Retrieved from `https://arxiv.org/abs/2006.07968`

Gentner, D., Shao, R., Simms, N., & Hespos, S. (2021). Learning same and different relations: cross-species comparisons. *Current Opinion in Behavioral Sciences*, *37*, 84–89.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Hochreiter, S., Younger, A. S., & Conwell, P. R. (2001). Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks (ICANN)* (pp. 87–94).

Hubel, D. H., Wiesel, T. N., et al. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*(3), 574–591.

Irie, K., & Lake, B. M. (2024). Neural networks that overcome classic challenges through practice. *arXiv preprint arXiv:2410.10596*.

Kerg, G., Mittal, S., Rolnick, D., Bengio, Y., Richards, B. A., & Lajoie, G. (2022). Inductive biases for relational tasks. In *ICLR 2022 workshop on the elements of reasoning: Objects, structure and causality.*

Kim, J., Ricci, M., & Serre, T. (2018). Not-so-CLEVR: learning same–different relations strains feedforward neural networks. *Interface focus*, *8*(4), 20180011.

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... others (2019). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, *32*.

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lupker, S. J., Nakayama, M., & Perea, M. (2015). Is there phonologically based priming in the same-different task? Evidence from Japanese-English bilinguals. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(5), 1281.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

McCoy, R. T., & Griffiths, T. L. (2023). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *arXiv preprint arXiv:2305.14701*.

Minsky, M. L., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2).

Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193.

Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences*, *6*(1), 125–136.

Puebla, G., & Bowers, J. S. (2022). Can deep convolutional neural networks support relational reasoning in the same-different task? *Journal of Vision*, *22*(10), 11–11.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. *Psycholinguistics: Critical Concepts in Psychology*, *4*, 216–271.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Stabinger, S., Rodríguez-Sánchez, A., & Piater, J. (2016). 25 years of CNNs: Can we compare to human abstraction capabilities? In *International Conference on Artificial Neural Networks (ICANN)* (pp. 380–387).

Tartaglini, A. R., Feucht, S., Lepori, M. A., Vong, W. K., Lovering, C., Lake, B. M., & Pavlick, E. (2023). Deep neural networks can learn generalizable same-different visual relations. *arXiv preprint arXiv:2310.09612*.

Wang, L., Cao, Z., De Melo, G., & Liu, Z. (2016). Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1298–1307).

Webb, T. W., Frankland, S. M., Altabaa, A., Segert, S., Krishnamurthy, K., Campbell, D., ... Cohen, J. D. (2024). The relational bottleneck as an inductive bias for efficient abstraction. *Trends in Cognitive Sciences*.

Webb, T. W., Sinha, I., & Cohen, J. (2021). Emergent symbols through binding in external memory. In *International Conference on Learning Representations.*