
One-Shot Imitation Learning

**Yan Duan^{†§}, Marcin Andrychowicz[†], Bradly Stadie^{††}, Jonathan Ho^{†§},
Jonas Schneider[†], Ilya Sutskever[†], Pieter Abbeel^{†§}, Wojciech Zaremba[†]**

[†]Berkeley AI Research Lab, [‡]OpenAI

[§]Work done while at OpenAI

{rockyduan, jonathanho, pabbeel}@eecs.berkeley.edu
{marcin, bstadie, jonas, ilyasu, woj}@openai.com

Abstract

Imitation learning has been commonly applied to solve different tasks in isolation. This usually requires either careful feature engineering, or a significant number of samples. This is far from what we desire: ideally, robots should be able to learn from very few demonstrations of any given task, and instantly generalize to new situations of the same task, without requiring task-specific engineering. In this paper, we propose a meta-learning framework for achieving such capability, which we call *one-shot imitation learning*.

Specifically, we consider the setting where there is a very large (maybe infinite) set of tasks, and each task has many instantiations. For example, a task could be to stack all blocks on a table into a single tower, another task could be to place all blocks on a table into two-block towers, etc. In each case, different instances of the task would consist of different sets of blocks with different initial states. At training time, our algorithm is presented with pairs of demonstrations for a subset of all tasks. A neural net is trained such that when it takes as input the first demonstration demonstration and a state sampled from the second demonstration, it should predict the action corresponding to the sampled state. At test time, a full demonstration of a single instance of a new task is presented, and the neural net is expected to perform well on new instances of this new task. Our experiments show that the use of soft attention allows the model to generalize to conditions and tasks unseen in the training data. We anticipate that by training this model on a much greater variety of tasks and settings, we will obtain a general system that can turn any demonstrations into robust policies that can accomplish an overwhelming variety of tasks.

1 Introduction

We are interested in robotic systems that are able to perform a variety of complex useful tasks, e.g. tidying up a home or preparing a meal. The robot should be able to learn new tasks without long system interaction time. To accomplish this, we must solve two broad problems. The first problem is that of dexterity: robots should learn how to approach, grasp and pick up complex objects, and how to place or arrange them into a desired configuration. The second problem is that of communication: how to communicate the *intent* of the task at hand, so that the robot can replicate it in a broader set of initial conditions.

Demonstrations are an extremely convenient form of information we can use to teach robots to overcome these two challenges. Using demonstrations, we can unambiguously communicate essentially any manipulation task, and simultaneously provide clues about the specific motor skills required to perform the task. We can compare this with an alternative form of communication, namely natural language. Although language is highly versatile, effective, and efficient, natural language processing

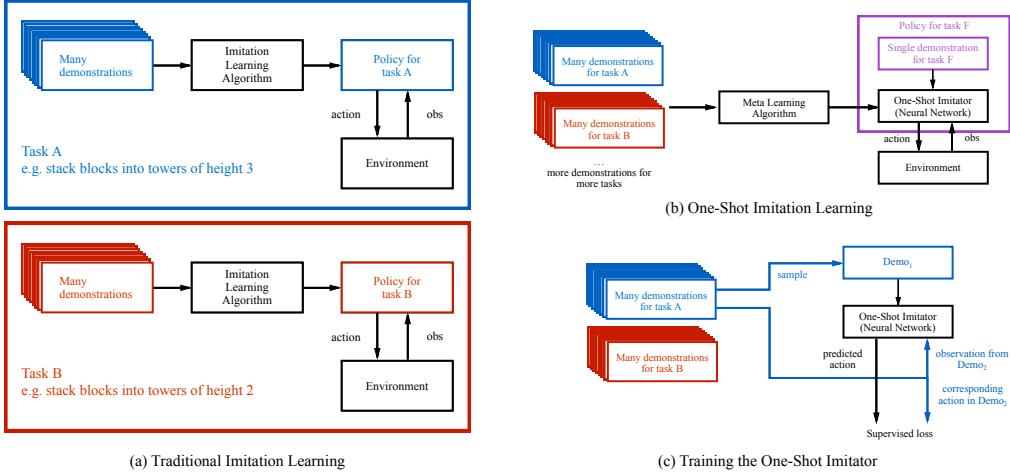


Figure 1: (a) Traditionally, policies are task-specific. For example, a policy might have been trained through an imitation learning algorithm to stack blocks into towers of height 3, and then another policy would be trained to stack blocks into towers of height 2, etc. (b) In this paper, we are interested in training networks that are *not* specific to one task, but rather can be told (through a single demonstration) what the current new task is, and be successful at this new task. For example, when it is conditioned on a single demonstration for task F, it should behave like a good policy for task F. (c) We can phrase this as a supervised learning problem, where we train this network on a set of training tasks, and with enough examples it should generalize to unseen, but related tasks. To train this network, in each iteration we sample a demonstration from one of the training tasks, and feed it to the network. Then, we sample another pair of observation and action from a second demonstration of the same task. When conditioned on both the first demonstration and this observation, the network is trained to output the corresponding action.

systems are not yet at a level where we could easily use language to precisely describe a complex task to a robot. Compared to language, using demonstrations has two fundamental advantages: first, it does not require the knowledge of language, as it is possible to communicate complex tasks to humans that don't speak one's language. And second, there are many tasks that are extremely difficult to explain in words, even if we assume perfect linguistic abilities; for example, explaining how to swim without demonstration and experience seems to be, at the very least, an extremely challenging task.

Indeed, learning from demonstrations have had many successful applications . However, so far these applications have either required careful feature engineering, or a significant amount of system interaction time. This is far from what we desire: ideally, we hope to demonstrate a certain task only once or a few times to the robot, and have it instantly generalize to new situations of the same task, without long system interaction time or domain knowledge about individual tasks.

In this paper we explore the one-shot imitation learning setting illustrated in Fig. 1, where the objective is to maximize the expected performance of the learned policy when faced with a new, previously unseen, task, and having received as input only one demonstration of that task. For the tasks we consider, the policy is expected to achieve good performance without any additional system interaction, once it has received the demonstration.

We train a policy on a broad distribution over tasks, where the number of tasks is potentially infinite. For each training task we assume the availability of a set of successful demonstrations. Our learned policy takes as input: (i) the current observation, and (ii) one demonstration that successfully solves a different instance of the same task (this demonstration is fixed for the duration of the episode). The policy outputs the current controls. We note that any pair of demonstrations for the same task provides a supervised training example for the neural net policy, where one demonstration is treated as the input, while the other as the output.

To make this model work, we made essential use of soft attention [6] for processing both the (potentially long) sequence of states and action that correspond to the demonstration, and for processing the components of the vector specifying the locations of the various blocks in our environment. The use of soft attention over both types of inputs made strong generalization possible. In particular, on a family of block stacking tasks, our neural network policy was able to perform well on novel block configurations which were not present in any training data. Videos of our experiments are available at <http://bit.ly/nips2017-oneshot>.

2 Related Work

Imitation learning considers the problem of acquiring skills from observing demonstrations. Survey articles include [48, 11, 3].

Two main lines of work within imitation learning are behavioral cloning, which performs supervised learning from observations to actions (e.g., [41, 44]); and inverse reinforcement learning [37], where a reward function [1, 66, 29, 18, 22] is estimated that explains the demonstrations as (near) optimal behavior. While this past work has led to a wide range of impressive robotics results, it considers each skill separately, and having learned to imitate one skill does not accelerate learning to imitate the next skill.

One-shot and few-shot learning has been studied for image recognition [61, 26, 47, 42], generative modeling [17, 43], and learning “fast” reinforcement learning agents with recurrent policies [16, 62]. Fast adaptation has also been achieved through fast-weights [5]. Like our algorithm, many of the aforementioned approaches are a form of meta-learning [58, 49, 36], where the algorithm itself is being learned. Meta-learning has also been studied to discover neural network weight optimization algorithms [8, 9, 23, 50, 2, 31]. This prior work on one-shot learning and meta-learning, however, is tailored to respective domains (image recognition, generative models, reinforcement learning, optimization) and not directly applicable in the imitation learning setting. Recently, [19] propose a generic framework for meta learning across several aforementioned domains. However they do not consider the imitation learning setting.

Reinforcement learning [56, 10] provides an alternative route to skill acquisition, by learning through trial and error. Reinforcement learning has had many successes, including Backgammon [57], helicopter control [39], Atari [35], Go [52], continuous control in simulation [51, 21, 32] and on real robots [40, 30]. However, reinforcement learning tends to require a large number of trials and requires specifying a reward function to define the task at hand. The former can be time-consuming and the latter can often be significantly more difficult than providing a demonstration [37].

Multi-task and transfer learning considers the problem of learning policies with applicability and re-use beyond a single task. Success stories include domain adaptation in computer vision [64, 34, 28, 4, 15, 24, 33, 59, 14] and control [60, 45, 46, 20, 54]. However, while acquiring a multitude of skills faster than what it would take to acquire each of the skills independently, these approaches do not provide the ability to readily pick up a new skill from a single demonstration.

Our approach heavily relies on an attention model over the demonstration and an attention model over the current observation. We use the soft attention model proposed in [6] for machine translations, and which has also been successful in image captioning [63]. The interaction networks proposed in [7, 12] also leverage locality of physical interaction in learning. Our model is also related to the sequence to sequence model [55, 13], as in both cases we consume a very long demonstration sequence and, effectively, emit a long sequence of actions.

3 One Shot Imitation Learning

3.1 Problem Formalization

We denote a distribution of tasks by \mathbb{T} , an individual task by $t \sim \mathbb{T}$, and a distribution of demonstrations for the task t by $\mathbb{D}(t)$. A policy is symbolized by $\pi_\theta(a|o, d)$, where a is an action, o is an observation, d is a demonstration, and θ are the parameters of the policy. A demonstration $d \sim \mathbb{D}(t)$ is a sequence of observations and actions : $d = [(o_1, a_1), (o_2, a_2), \dots, (o_T, a_T)]$. We assume that the distribution of tasks \mathbb{T} is given, and that we can obtain successful demonstrations for each task. We assume that there is some scalar-valued evaluation function $R_t(d)$ (e.g. a binary value

indicating success) for each task, although this is not required during training. The objective is to maximize the expected performance of the policy, where the expectation is taken over tasks $t \in \mathbb{T}$, and demonstrations $d \in \mathbb{D}(t)$.

3.2 Block Stacking Tasks

To clarify the problem setting, we describe a concrete example of a distribution of block stacking tasks, which we will also later study in the experiments. The compositional structure shared among these tasks allows us to investigate nontrivial generalization to unseen tasks. For each task, the goal is to control a 7-DOF Fetch robotic arm to stack various numbers of cube-shaped blocks into a specific configuration specified by the user. Each configuration consists of a list of blocks arranged into towers of different heights, and can be identified by a string. For example, ab cd ef gh means that we want to stack 4 towers, each with two blocks, and we want block A to be on top of block B, block C on top of block D, block E on top of block F, and block G on top of block H. Each of these configurations correspond to a different task. Furthermore, in each episode the starting positions of the blocks may vary, which requires the learned policy to generalize even within the training tasks. In a typical task, an observation is a list of (x, y, z) object positions relative to the gripper, and information if gripper is opened or closed. The number of objects may vary across different task instances. We define a *stage* as a single operation of stacking one block on top of another. For example, the task ab cd ef gh has 4 stages.

3.3 Algorithm

In order to train the neural network policy, we make use of imitation learning algorithms such as behavioral cloning and DAGGER [44], which only require demonstrations rather than reward functions to be specified. This has the potential to be more scalable, since it is often easier to demonstrate a task than specifying a well-shaped reward function [38].

We start by collecting a set of demonstrations for each task, where we add noise to the actions in order to have wider coverage in the trajectory space. In each training iteration, we sample a list of tasks (with replacement). For each sampled task, we sample a demonstration as well as a small batch of observation-action pairs. The policy is trained to regress against the desired actions when conditioned on the current observation and the demonstration, by minimizing an ℓ_2 or cross-entropy loss based on whether actions are continuous or discrete. A high-level illustration of the training procedure is given in Fig. 1(c). Across all experiments, we use Adamax [25] to perform the optimization with a learning rate of 0.001.

4 Architecture

While, in principle, a generic neural network could learn the mapping from demonstration and current observation to appropriate action, we found it important to use an appropriate architecture. Our architecture for learning block stacking is one of the main contributions of this paper, and we believe it is representative of what architectures for one-shot imitation learning could look like in the future when considering more complex tasks.

Our proposed architecture consists of three modules: the demonstration network, the context network, and the manipulation network. An illustration of the architecture is shown in Fig. 2. We will describe the main operations performed in each module below, and a full specification is available in the Appendix.

4.1 Demonstration Network

The demonstration network receives a demonstration trajectory as input, and produces an embedding of the demonstration to be used by the policy. The size of this embedding grows linearly as a function of the length of the demonstration as well as the number of blocks in the environment.

Temporal Dropout: For block stacking, the demonstrations can span hundreds to thousands of time steps, and training with such long sequences can be demanding in both time and memory usage. Hence, we randomly discard a subset of time steps during training, an operation we call *temporal dropout*, analogous to [53, 27]. We denote p as the proportion of time steps that are thrown away.

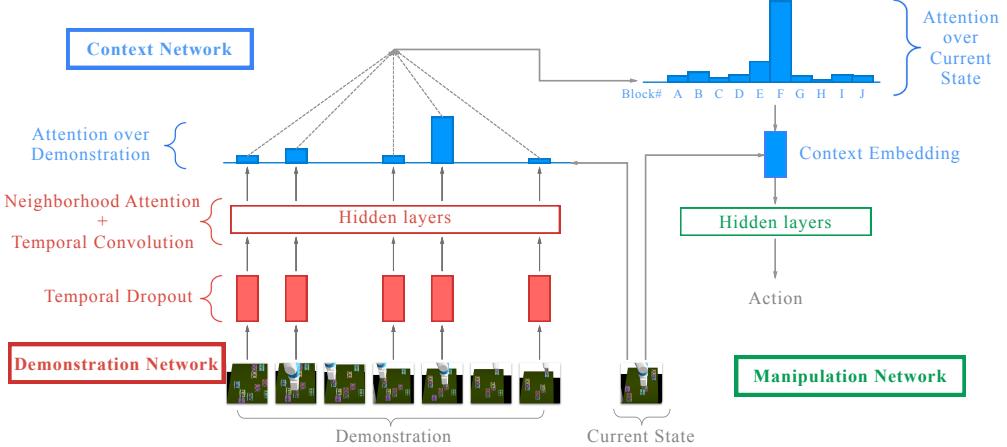


Figure 2: Illustration of the network architecture.

In our experiments, we use $p = 0.95$, which reduces the length of demonstrations by a factor of 20. During test time, we can sample multiple downsampled trajectories, use each of them to compute downstream results, and average these results to produce an ensemble estimate. In our experience, this consistently improves the performance of the policy.

Neighborhood Attention: After downsampling the demonstration, we apply a sequence of operations, composed of dilated temporal convolution [65] and neighborhood attention. We now describe this second operation in more detail.

Since our neural network needs to handle demonstrations with variable numbers of blocks, it must have modules that can process variable-dimensional inputs. Soft attention is a natural operation which maps variable-dimensional inputs to fixed-dimensional outputs. However, by doing so, it may lose information compared to its input. This is undesirable, since the amount of information contained in a demonstration grows as the number of blocks increases. Therefore, we need an operation that can map variable-dimensional inputs to outputs with comparable dimensions. Intuitively, rather than having a single output as a result of attending to all inputs, we have as many outputs as inputs, and have each output attending to all other inputs in relation to its own corresponding input.

We start by describing the soft attention module as specified in [6]. The input to the attention includes a query q , a list of context vectors $\{c_j\}$, and a list of memory vectors $\{m_j\}$. The i th attention weight is given by $w_i \leftarrow v^T \tanh(q + c_i)$, where v is a learned weight vector. The output of attention is a weighted combination of the memory content, where the weights are given by a softmax operation over the attention weights. Formally, we have output $\leftarrow \sum_i m_i \frac{\exp(w_i)}{\sum_j \exp(w_j)}$. Note that the output has the same dimension as a memory vector. The attention operation can be generalized to multiple query heads, in which case there will be as many output vectors as there are queries.

Now we turn to neighborhood attention. We assume there are B blocks in the environment. We denote the robot's state as s_{robot} , and the coordinates of each block as $(x_1, y_1, z_1), \dots, (x_B, y_B, z_B)$. The input to neighborhood attention is a list of embeddings $h_1^{in}, \dots, h_B^{in}$ of the same dimension, which can be the result of a projection operation over a list of block positions, or the output of a previous neighborhood attention operation. Given this list of embeddings, we use two separate linear layers to compute a query vector and a context embedding for each block: $q_i \leftarrow \text{Linear}(h_i^{in})$, and $c_i \leftarrow \text{Linear}(h_i^{in})$. The memory content to be extracted consists of the coordinates of each block, concatenated with the input embedding. The i th query result is given by the following soft attention operation: $\text{result}_i \leftarrow \text{SoftAttn}(\text{query}: q_i, \text{context}: \{c_j\}_{j=1}^B, \text{memory}: \{(x_j, y_j, z_j), h_j^{in}\}_{j=1}^B)$.

Intuitively, this operation allows each block to query other blocks in relation to itself (e.g. find the closest block), and extract the queried information. The gathered results are then combined with each block's own information, to produce the output embedding per block. Concretely, we have

$\text{output}_i \leftarrow \text{Linear}(\text{concat}(h_i^{in}, \text{result}_i, (x_i, y_i, z_i), s_{\text{robot}}))$. In practice, we use multiple query heads per block, so that the size of each result_i will be proportional to the number of query heads.

4.2 Context network

The context network is the crux of our model. It processes both the current state and the embedding produced by the demonstration network, and outputs a context embedding, whose dimension does not depend on the length of the demonstration, or the number of blocks in the environment. Hence, it is forced to capture only the relevant information, which will be used by the manipulation network.

Attention over demonstration: The context network starts by computing a query vector as a function of the current state, which is then used to attend over the different time steps in the demonstration embedding. The attention weights over different blocks within the same time step are summed together, to produce a single weight per time step. The result of this temporal attention is a vector whose size is proportional to the number of blocks in the environment. We then apply neighborhood attention to propagate the information across the embeddings of each block. This process is repeated multiple times, where the state is advanced using an LSTM cell with untied weights.

Attention over current state: The previous operations produce an embedding whose size is independent of the length of the demonstration, but still dependent on the number of blocks. We then apply standard soft attention over the current state to produce fixed-dimensional vectors, where the memory content only consists of positions of each block, which, together with the robot’s state, forms the *context embedding*, which is then passed to the manipulation network.

Intuitively, although the number of objects in the environment may vary, at each stage of the manipulation operation, the number of relevant objects is small and usually fixed. For the block stacking environment specifically, the robot should only need to pay attention to the position of the block it is trying to pick up (the *source* block), as well as the position of the block it is trying to place on top of (the *target* block). Therefore, a properly trained network can learn to match the current state with the corresponding stage in the demonstration, and infer the identities of the source and target blocks expressed as soft attention weights over different blocks, which are then used to extract the corresponding positions to be passed to the manipulation network. Although we do not enforce this interpretation in training, our experiment analysis supports this interpretation of how the learned policy works internally.

4.3 Manipulation network

The manipulation network is the simplest component. After extracting the information of the source and target blocks, it computes the action needed to complete the current stage of stacking one block on top of another one, using a simple MLP network.¹ This division of labor opens up the possibility of modular training: the manipulation network may be trained to complete this simple procedure, without knowing about demonstrations or more than two blocks present in the environment. We leave this possibility for future work.

5 Experiments

We conduct experiments with the block stacking tasks described in Section 3.2.² These experiments are designed to answer the following questions:

- How does training with behavioral cloning compare with DAGGER?
- How does conditioning on the entire demonstration compare to conditioning on the final state, even when it already has enough information to fully specify the task?
- How does conditioning on the entire demonstration compare to conditioning on a “snapshot” of the trajectory, which is a small subset of frames that are most informative?

¹In principle, one can replace this module with an RNN module. But we did not find this necessary for the tasks we consider.

²Additional experiment results are available in the Appendix, including a simple illustrative example of particle reaching tasks and further analysis of block stacking

- Can our framework generalize to tasks that it has never seen during training?

To answer these questions, we compare the performance of the following architectures:

- **BC:** We use the same architecture as previous, but and the policy using behavioral cloning.
- **DAGGER:** We use the architecture described in the previous section, and train the policy using DAGGER.
- **Final state:** This architecture conditions on the final state rather than on the entire demonstration trajectory. For the block stacking task family, the final state uniquely identifies the task, and there is no need for additional information. However, a full trajectory, one which contains information about intermediate stages of the task’s solution, can make it easier to train the optimal policy, because it could learn to rely on the demonstration directly, without needing to memorize the intermediate steps into its parameters. This is related to the way in which reward shaping can significantly affect performance in reinforcement learning [38]. A comparison between the two conditioning strategies will tell us whether this hypothesis is valid. We train this policy using DAGGER.
- **Snapshot:** This architecture conditions on a “snapshot” of the trajectory, which includes the last frame of each stage along the demonstration trajectory. This assumes that a segmentation of the demonstration into multiple stages is available at test time, which gives it an unfair advantage compared to the other conditioning strategies. Hence, it may perform better than conditioning on the full trajectory, and serves as a reference, to inform us whether the policy conditioned on the entire trajectory can perform as well as if the demonstration is clearly segmented. Again, we train this policy using DAGGER.

We evaluate the policy on tasks seen during training, as well as tasks unseen during training. Note that generalization is evaluated at multiple levels: the learned policy not only needs to generalize to new configurations and new demonstrations of tasks seen already, but also needs to generalize to new tasks.

Concretely, we collect 140 training tasks, and 43 test tasks, each with a different desired layout of the blocks. The number of blocks in each task can vary between 2 and 10. We collect 1000 trajectories per task for training, and maintain a separate set of trajectories and initial configurations to be used for evaluation. The trajectories are collected using a hard-coded policy.

5.1 Performance Evaluation

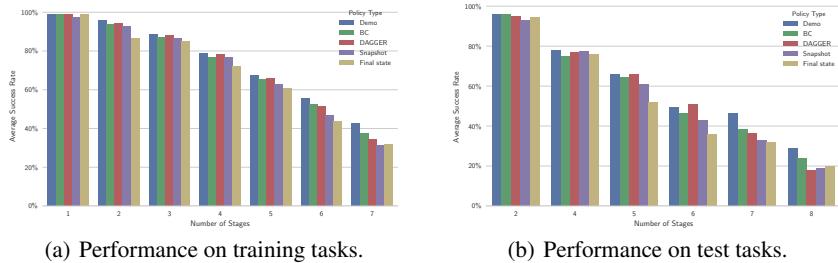


Figure 3: Comparison of different conditioning strategies. The darkest bar shows the performance of the hard-coded policy, which unsurprisingly performs the best most of the time. For architectures that use temporal dropout, we use an ensemble of 10 different downsampled demonstrations and average the action distributions. Then for all architectures we use the greedy action for evaluation.

Fig. 3 shows the performance of various architectures. Results for training and test tasks are presented separately, where we group tasks by the number of stages required to complete them. This is because tasks that require more stages to complete are typically more challenging. In fact, even our scripted policy frequently fails on the hardest tasks. We measure success rate per task by executing the greedy policy (taking the most confident action at every time step) in 100 different configurations, each conditioned on a different demonstration unseen during training. We report the average success rate over all tasks within the same group.

From the figure, we can observe that for the easier tasks with fewer stages, all of the different conditioning strategies perform equally well and almost perfectly. As the difficulty (number of stages) increases, however, conditioning on the entire demonstration starts to outperform conditioning on the final state. One possible explanation is that when conditioned only on the final state, the policy may struggle about which block it should stack first, a piece of information that is readily accessible from demonstration, which not only communicates the task, but also provides valuable information to help accomplish it.

More surprisingly, conditioning on the entire demonstration also seems to outperform conditioning on the snapshot, which we originally expected to perform the best. We suspect that this is due to the regularization effect introduced by temporal dropout, which effectively augments the set of demonstrations seen by the policy during training.

Another interesting finding was that training with behavioral cloning has the same level of performance as training with DAGGER, which suggests that the entire training procedure could work without requiring interactive supervision. In our preliminary experiments, we found that injecting noise into the trajectory collection process was important for behavioral cloning to work well, hence in all experiments reported here we use noise injection. In practice, such noise can come from natural human-induced noise through tele-operation, or by artificially injecting additional noise before applying it on the physical robot.

5.2 Visualization

We visualize the attention mechanisms underlying the main policy architecture to have a better understanding about how it operates. There are two kinds of attention we are mainly interested in, one where the policy attends to different time steps in the demonstration, and the other where the policy attends to different blocks in the current state. Fig. 4 shows some of the attention heatmaps.

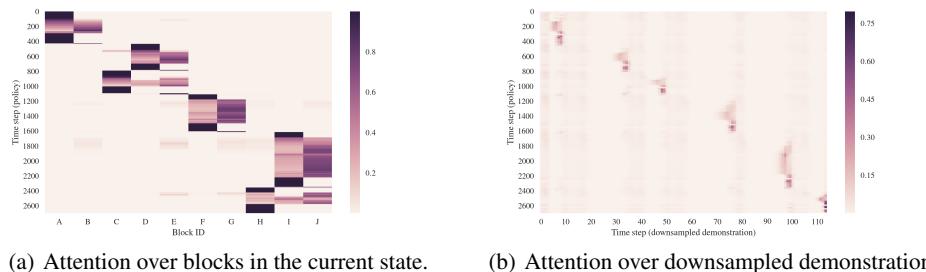


Figure 4: Visualizing attentions performed by the policy during an entire execution. The task being performed is ab cde fg hij. Note that the policy has multiple query heads for each type of attention, and only one query head per type is visualized. (a) We can observe that the policy almost always focuses on a small subset of the block positions in the current state, which allows the manipulation network to generalize to operations over different blocks. (b) We can observe a sparse pattern of time steps that have high attention weights. This suggests that the policy has essentially learned to segment the demonstrations, and only attend to important key frames. Note that there are roughly 6 regions of high attention weights, which nicely corresponds to the 6 stages required to complete the task.

6 Conclusions

In this work, we presented a simple model that maps a single successful demonstration of a task to an effective policy that solves said task in a new situation. We demonstrated effectiveness of this approach on a family of block stacking tasks. There are a lot of exciting directions for future work. We plan to extend the framework to demonstrations in the form of image data, which will allow more end-to-end learning without requiring a separate perception module. We are also interested in enabling the policy to condition on multiple demonstrations, in case where one demonstration does not fully resolve ambiguity in the objective. Furthermore and most importantly, we hope to scale up

our method on a much larger and broader distribution of tasks, and explore its potential towards a general robotics imitation learning system that would be able to achieve an overwhelming variety of tasks.

7 Acknowledgement

We would like to thank our colleagues at UC Berkeley and OpenAI for insightful discussions. This research was funded in part by ONR through a PECASE award. Yan Duan was also supported by a Huawei Fellowship. Jonathan Ho was also supported by an NSF Fellowship.

References

- [1] Pieter Abbeel and Andrew Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.
- [2] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Neural Information Processing Systems (NIPS)*, 2016.
- [3] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.
- [4] Yusuf Aytar and Andrew Zisserman. Tabula rasa: Model transfer for object category detection. In *2011 International Conference on Computer Vision*, pages 2252–2259. IEEE, 2011.
- [5] Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Neural Information Processing Systems (NIPS)*, 2016.
- [6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [7] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, pages 4502–4510, 2016.
- [8] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule. In *Optimality in Artificial and Biological Neural Networks*, pages 6–8, 1992.
- [9] Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. *Learning a synaptic learning rule*. Université de Montréal, Département d’informatique et de recherche opérationnelle, 1990.
- [10] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 560–564. IEEE, 1995.
- [11] Sylvain Calinon. *Robot programming by demonstration*. EPFL Press, 2009.
- [12] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. In *Int. Conf. on Learning Representations (ICLR)*, 2017.
- [13] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014.
- [15] Lixin Duan, Dong Xu, and Ivor Tsang. Learning with augmented features for heterogeneous domain adaptation. *arXiv preprint arXiv:1206.4660*, 2012.

- [16] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL²: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.
- [17] Harrison Edwards and Amos Storkey. Towards a neural statistician. *International Conference on Learning Representations (ICLR)*, 2017.
- [18] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, 2016.
- [19] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.
- [20] Abhishek Gupta, Coline Devin, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Learning invariant feature spaces to transfer skills with reinforcement learning. In *Int. Conf. on Learning Representations (ICLR)*, 2017.
- [21] Nicolas Heess, Gregory Wayne, David Silver, Tim Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. In *Advances in Neural Information Processing Systems*, pages 2944–2952, 2015.
- [22] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.
- [23] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*. Springer, 2001.
- [24] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*, 2013.
- [25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.
- [26] Gregory Koch. Siamese neural networks for one-shot image recognition. *ICML Deep Learning Workshop*, 2015.
- [27] David Krueger, Tegan Maharaj, János Kramár, Mohammad Pezeshki, Nicolas Ballas, Nan Rosemary Ke, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, Aaron Courville, et al. Zoneout: Regularizing rnns by randomly preserving hidden activations. *arXiv preprint arXiv:1606.01305*, 2016.
- [28] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792. IEEE, 2011.
- [29] S. Levine, Z. Popovic, and V. Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.
- [30] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.
- [31] Ke Li and Jitendra Malik. Learning to optimize. *arXiv preprint arXiv:1606.01885*, 2016.
- [32] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- [33] Mingsheng Long and Jianmin Wang. Learning transferable features with deep adaptation networks. *CoRR, abs/1502.02791*, 1:2, 2015.
- [34] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

- [35] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- [36] Devang K Naik and RJ Mammone. Meta-neural networks that learn by learning. In *International Joint Conference on Neural Networks (IJCNN)*, 1992.
- [37] Andrew Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.
- [38] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.
- [39] Andrew Y Ng, H Jin Kim, Michael I Jordan, Shankar Sastry, and Shiv Ballianda. Autonomous helicopter flight via reinforcement learning. In *NIPS*, volume 16, 2003.
- [40] Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697, 2008.
- [41] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, pages 305–313, 1989.
- [42] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Under Review, ICLR*, 2017.
- [43] Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. *International Conference on Machine Learning (ICML)*, 2016.
- [44] Stéphane Ross, Geoffrey J Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, volume 1, page 6, 2011.
- [45] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- [46] Fereshteh Sadeghi and Sergey Levine. (cad)² rl: Real single-image flight without a single real image. 2016.
- [47] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning (ICML)*, 2016.
- [48] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242, 1999.
- [49] Jurgen Schmidhuber. Evolutionary principles in self-referential learning. *On learning how to learn: The meta-meta... hook.) Diploma thesis, Institut f. Informatik, Tech. Univ. Munich*, 1987.
- [50] Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 1992.
- [51] John Schulman, Sergey Levine, Pieter Abbeel, Michael I Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, pages 1889–1897, 2015.
- [52] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [53] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

- [54] Bradlie Stadie, Pieter Abbeel, and Ilya Sutskever. Third person imitation learning. In *Int. Conf. on Learning Representations (ICLR)*, 2017.
- [55] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [56] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
- [57] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.
- [58] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 1998.
- [59] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.
- [60] Eric Tzeng, Coline Devin, Judy Hoffman, Chelsea Finn, Xingchao Peng, Pieter Abbeel, Sergey Levine, Kate Saenko, and Trevor Darrell. Towards adapting deep visuomotor representations from simulated to real environments. *arXiv preprint arXiv:1511.07111*, 2015.
- [61] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Neural Information Processing Systems (NIPS)*, 2016.
- [62] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.
- [63] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015.
- [64] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 188–197. ACM, 2007.
- [65] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2016.
- [66] B. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

A Illustrative Example: Particle Reaching

The particle reaching problem is a very simple family of tasks. In each task, we control a point robot to reach a specific landmark, and different tasks are identified by different landmarks. As illustrated in Fig. 1, one task could be to reach the orange square, and another task could be to reach the green triangle. The agent receives its own 2D location, as well as the 2D locations of each of the landmarks. Within each task, the initial position of the agent, as well as the positions of all the landmarks, can vary across different instances of the task.

Without a demonstration, the robot does not know which landmark it should reach, and will not be able to accomplish the task. Hence, this setting already gets at the essence of one-shot imitation, namely to communicate the task via a demonstration. After learning, the agent should be able to identify the target landmark from the demonstration, and reach the same landmark in a new instance of the task.

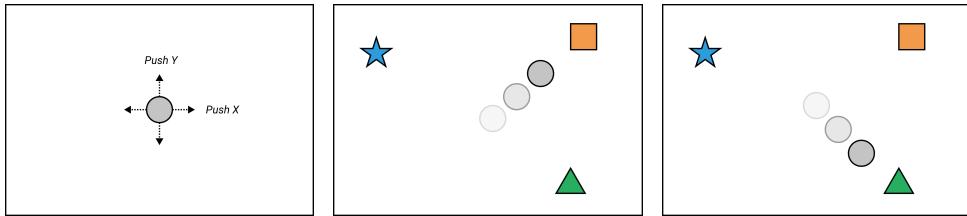


Figure 1: The robot is a point mass controlled with 2-dimensional force. The family of tasks is to reach a target landmark. The identity of the landmark differs from task to task, and the model has to figure out which target to pursue based on the demonstration. (left) illustration of the robot; (middle) the task is to reach the orange box, (right) the task is to reach the green triangle.

We consider three architectures for this problem:

- **Plain LSTM:** The first architecture is a simple LSTM with 512 hidden units. It reads the demonstration trajectory, the output of which is then concatenated with the current state, and fed to a multi-layer perceptron (MLP) to produce the action.
- **LSTM with attention:** In this architecture, the LSTM outputs a weighting over the different landmarks from the demonstration sequence. Then, it applies this weighting in the test scene, and produces a weighted combination over landmark positions given the current state. This 2D output is then concatenated with the current agent position, and fed to an MLP to produce the action.
- **Final state with attention:** Rather than looking at the entire demonstration trajectory, this architecture only looks at the final state in the demonstration (which is already sufficient to communicate the task), and produce a weighting over landmarks. It then proceeds like the previous architecture.

Notice that these three architectures are increasingly more specialized to the specific particle reaching setting, which suggests a potential trade-off between expressiveness and generalizability.

The experiment results are shown in Fig. 2. We observe that as the architecture becomes more specialized, we achieve much better generalization performance. For this simple task, it appears that conditioning on the entire demonstration hurts generalization performance, and conditioning on just the final state performs the best even without explicit regularization. This makes intuitive sense, since the final state already sufficiently characterizes the task at hand.

However, the same conclusion does not appear to hold as the task becomes more complicated, as shown by the block stacking tasks in the main text.

Fig. 3 shows the learning curves for the three architectures designed for the particle reaching tasks, as the number of landmarks is varied, by running the policies over 100 different configurations, and computing success rates over both training and test data. We can clearly observe that both LSTM-based architectures exhibit overfitting as the number of landmarks increases. On the other hand, using attention clearly improves generalization performance, and when conditioning on only the final state, it achieves perfect generalization in all scenarios. It is also interesting to observe that

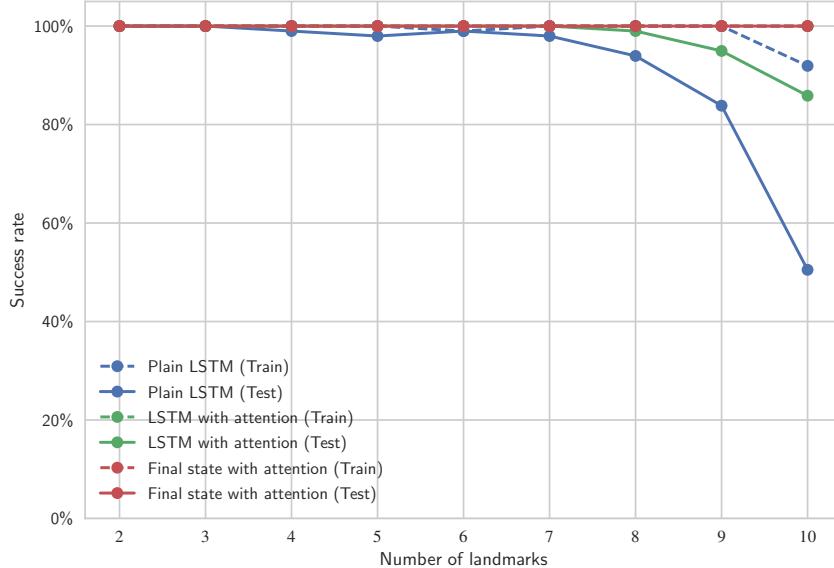


Figure 2: Success rates of different architectures for particle reaching. The “Train” curves show the success rates when conditioned on demonstrations seen during training, and running the policy on initial conditions seen during training, while the “Test” curves show the success rates when conditioned on new trajectories and operating in new situations. Both attention-based architectures achieve perfect training success rates, and the curves are overlapped.

learning undergoes a phase transition. Intuitively, this may be when the network is learning to infer the task from the demonstration. Once this is finished, the learning of control policy is almost trivial.

Table 1 and Table 2 show the exact performance numbers for reference.

#Landmarks	Plain LSTM	LSTM with attention	Final state with attention
2	100.0%	100.0%	100.0%
3	100.0%	100.0%	100.0%
4	100.0%	100.0%	100.0%
5	100.0%	100.0%	100.0%
6	99.0%	100.0%	100.0%
7	100.0%	100.0%	100.0%
8	100.0%	100.0%	100.0%
9	100.0%	100.0%	100.0%
10	91.9%	100.0%	100.0%

Table 1: Success rates of particle reaching conditioned on seen demonstrations, and running on seen initial configurations.

B Further Details on Block Stacking

B.1 Full Description of Architecture

We now specify the architecture in pseudocode. We omit implementation details which involve handling a minibatch of demonstrations and observation-action pairs, as well as necessary padding and masking to handle data of different dimensions. We use weight normalization with data-dependent initialization Salimans and Kingma [2016] for all dense and convolution operations.

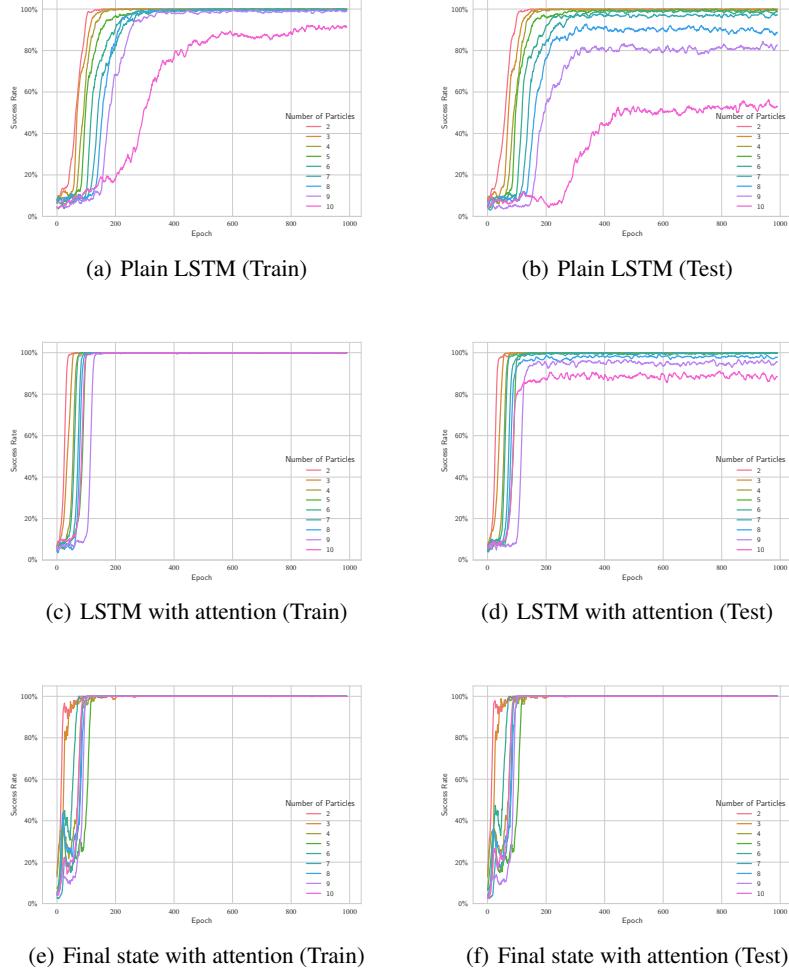


Figure 3: Learning curves for particle reaching tasks. Shown success rates are moving averages of past 10 epochs for smoother curves. Each policy is trained for up to 1000 epochs, which takes up to an hour using a Titan X Pascal GPU (as can be seen from the plot, most experiments can be finished sooner).

B.1.1 Demonstration Network

Assume that the demonstration has T time steps and we have B blocks. Our architecture only make use of the observations in the input demonstration but not the actions. Each observation is a $(3B + 2)$ -dimensional vector, containing the (x, y, z) coordinates of each block relative to the current position of the gripper, as well as a 2-dimensional gripper state indicating whether it is open or closed.

The full sequence of operations is given in Module 1. We first apply temporal dropout as described in the main text. Then we split the observation into information about the block and information about the robot, where the first dimension is time and the second dimension is the block ID. The robot state is broadcasted across different blocks. Hence the shape of outputs should be $\bar{T} \times B \times 3$ and $\bar{T} \times B \times 2$, respectively.

Then, we perform a 1×1 convolution over the block states to project them to the same dimension as the per-block embedding. Then we perform a sequence of neighborhood attention operations and 1×1 convolutions, where the input to the convolution is the concatenation of the attention result,

Module 1 Demonstration Network

Input: Demonstration $d \in \mathbb{R}^{T \times (3B+2)}$

Hyperparameters: $p = 0.95, D = 64$

Output: Demonstration embedding $\in \mathbb{R}^{\tilde{T} \times B \times D}$, where $\tilde{T} = \lceil T(1-p) \rceil$ is the length of the down-sampled trajectory.

```
d' ← TemporalDropout(d, probability=p)
block_state, robot_state ← Split(d')
h ← Conv1D(block_state, kernel_size=1, channels=D)
for a ∈ {1, 2, 4, 8} do
    // Residual connections
    h' ← ReLU(h)
    attn_result ← NeighborhoodAttention(h')
    h' ← Concat({h', block_state, robot_state}, axis=-1)
    h' ← Conv1D(h', kernel_size=2, channels=D, dilation=a)
    h' ← ReLU(h')
    h ← h + h'
end for
demo_embedding ← h
```

the current block position, and the robot’s state. This allows each block to query the state of other blocks, and reason about the query result in comparison with its own state and the robot’s state. We use residual connections during this procedure.

B.1.2 Context Network

The pseudocode is shown in Module 2. We perform a series of attention operations over the demonstration, followed by attention over the current state, and we apply them repeatedly through an LSTM with different weights per time step (we found this to be slightly easier to optimize). Then, in the end we apply a final attention operation which produces a fixed-dimensional embedding independent of the length of the demonstration or the number of blocks in the environment.

B.1.3 Manipulation Network

Given the context embedding, this module is simply a multilayer perceptron. Pseudocode is given in Module 3.

B.2 Evaluating Permutation Invariance

During training and in the previous evaluations, we only select one task per equivalence class, where two tasks are considered equivalent if they are the same up to permuting different blocks. This is based on the assumption that our architecture is invariant to permutations among different blocks. For example, if the policy is only trained on the task abcd, it should perform well on task dcba, given a single demonstration of the task dcba. We now experimentally verify this property by fixing a training task, and evaluating the policy’s performance under all equivalent permutations of it. As Fig. 4 shows, although the policy has only seen the task abcd, it achieves the same level of performance on all other equivalent tasks.

B.3 Effect of Ensembling

We now evaluate the importance of sampling multiple downsampled demonstrations during evaluation. Fig. 5 shows the performance across all training and test tasks, as the number of ensembles varies from 1 to 20. We observe that more ensembles helps the most for tasks with fewer stages. On the other hand, it consistently improves performance for the harder tasks, although the gap is smaller. We suspect that this is because the policy has learned to attend to frames in the demonstration trajectory where the blocks are already stacked together. In tasks with only 1 stage, for example, it is very easy for these frames to be dropped in a single downsampled demonstration. On the other hand, in tasks with more stages, it becomes more resilient to missing frames. Using more than 10

Module 2 Context Network

Input: Demonstration embedding $h_{in} \in \mathbb{R}^{\tilde{T} \times B \times D}$, current state $s \in \mathbb{R}^{3B+2}$

Hyperparameters: $D = 64, t_{lstm} = 4, H = 2$

Output: Context embedding $\in \mathbb{R}^{2+6H}$

```
// Split the current state into block state  $\in \mathbb{R}^{B \times 3}$  and robot state broadcasted to all blocks  $\in \mathbb{R}^{B \times 2}$ 
block_state, robot_state  $\leftarrow$  SplitSingle(s)
// Initialize LSTM output  $\in \mathbb{R}^{B \times D}$  and state (including hidden and cell state)  $\in \mathbb{R}^{B \times 2D}$ 
output, state  $\leftarrow$  InitLSTMState(size=B, hidden_dim=D)
for t = 1 to  $t_{lstm}$  do
    // Temporal attention: every block attend to the same time step
    x  $\leftarrow$  output
    if t > 1 then
        x  $\leftarrow$  ReLU(x)
    end if
    // Computing query for attention over demonstration  $\in \mathbb{R}^{B \times D}$ 
    q  $\leftarrow$  Dense(x, output_dim=D)
    // Compute result from attention  $\in \mathbb{R}^{H \times B \times D}$ 
    temp  $\leftarrow$  SoftAttention(query=q, context=h_in, memory=h_in, num_heads=H)
    // Reorganize result into shape  $B \times (HD)$ 
    temp  $\leftarrow$  Reshape(Transpose(temp, (1, 0, 2)), (B, H*D))

    // Spatial attention: each block attend to a different block separately
    x  $\leftarrow$  output
    if t > 1 then
        x  $\leftarrow$  ReLU(x)
    end if
    x  $\leftarrow$  Concat({x, temp}, axis=-1)
    // Computing context for attention over current state  $\in \mathbb{R}^{B \times D}$ 
    ctx  $\leftarrow$  Dense(x, output_dim=D)
    // Computing query for attention over current state  $\in \mathbb{R}^{B \times D}$ 
    q  $\leftarrow$  Dense(x, output_dim=D)
    // Computing memory for attention over current state  $\in \mathbb{R}^{B \times (HD+3)}$ 
    mem  $\leftarrow$  Concat({block_state, temp}, axis=-1)
    // Compute result from attention  $\in \mathbb{R}^{B \times H \times (HD+3)}$ 
    spatial  $\leftarrow$  SoftAttention(query=q, context=ctx, memory=mem, num_heads=H)
    // Reorganize result into shape  $B \times H(HD + 3)$ 
    spatial  $\leftarrow$  Reshape(spatial, (B, H * (H*D+3)))
    // Form input to the LSTM cell  $\in \mathbb{R}^{B \times (H(HD+3)+HD+8)}$ 
    input  $\leftarrow$  Concat({robot_state, block_state, spatial, temp}, axis=-1)
    // Run one step of an LSTM with untied weights (meaning that we use different weights per
    // time step
    output, state  $\leftarrow$  LSTMOneStep(input=input, state=state)
end for
// Final attention over the current state, compressing an  $O(B)$  representation down to  $O(1)$ 
// Compute the query vector. We use a fixed, trainable query vector independent of the input data,
with size  $\in \mathbb{R}^{2 \times D}$  (we use two queries, originally intended to have one for the source block and
one for the target block)
q  $\leftarrow$  GetFixedQuery()
// Get attention result, which should be of shape  $2 \times H \times 3$ 
r  $\leftarrow$  SoftAttention(query=q, context=output, memory=block_state, num_heads=H)
// Form the final context embedding (we pick the first robot state since no need to broadcast here)
context_embedding  $\leftarrow$  Concat({robot_state[0], Reshape(r, 2*H*3)})
```

Module 3 Manipulation Network

Input: Context embedding $h_{in} \in \mathbb{R}^{2+6H}$

Hyperparameters: $H = 2$

Output: Predicted action distribution $\in \mathbb{R}^{|A|}$

```

h ← ReLU(Dense(h_in, output_dim=256))
h ← ReLU(Dense(h, output_dim=256))
action_dist ← Dense(h, output_dim=|A|)

```

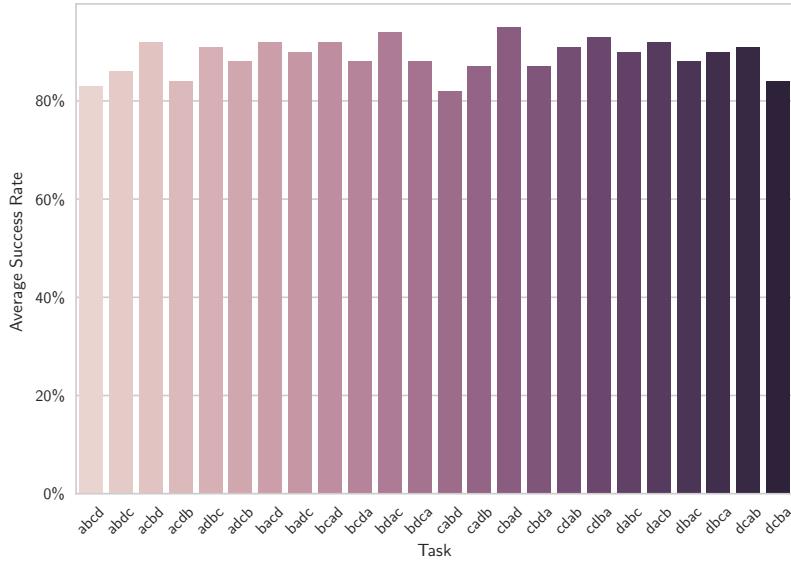


Figure 4: Performance of policy on a set of tasks equivalent up to permutations.

ensembles appears to provide no significant improvements, and hence we used 10 ensembles in our main evaluation.

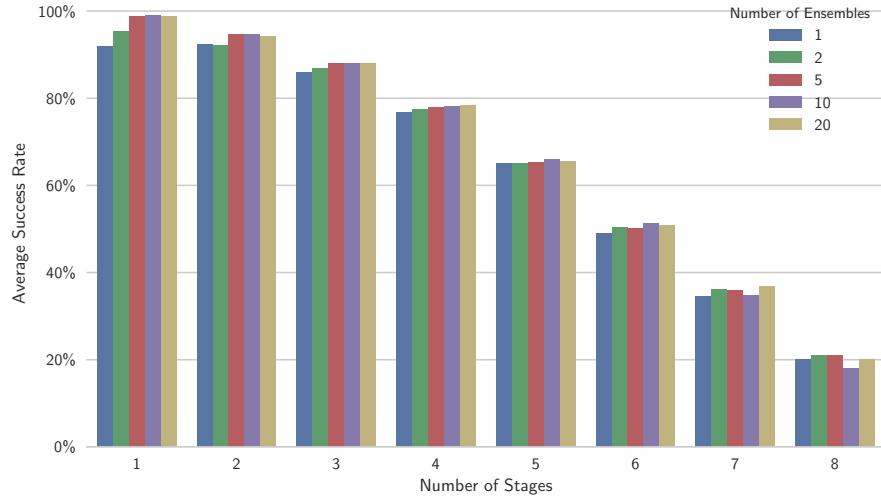


Figure 5: Performance of various number of ensembles.

B.4 Breakdown of Failure Cases

To understand the limitations of the current approach, we perform a breakdown analysis of the failure cases. We consider three failure scenarios: “Wrong move” means that the policy has arranged a layout incompatible with the desired layout. This could be because the policy has misinterpreted the demonstration, or due to an accidental bad move that happens to scramble the blocks into the wrong layout. “Manipulation failure” means that the policy has made an irrecoverable failure, for example if the block is shaken off the table, which the current hard-coded policy does not know how to handle. “Recoverable failure” means that the policy runs out of time before finishing the task, which may be due to an accidental failure during the operation that would have been recoverable given more time. As shown in Fig. 6, conditioning on only the final state makes more wrong moves compared to other architectures. Apart from that, most of the failure cases are actually due to manipulation failures that are mostly irrecoverable.¹ This suggests that better manipulation skills need to be acquired to make the learned one-shot policy more reliable.

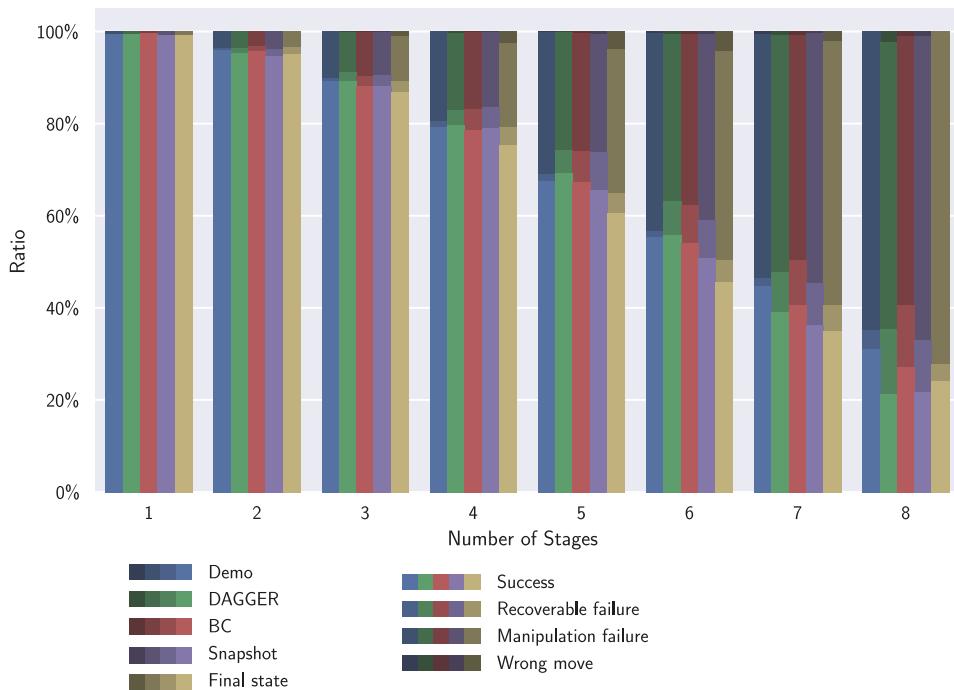


Figure 6: Breakdown of the success and failure scenarios. The area that each color occupies represent the ratio of the corresponding scenario.

B.5 Learning Curves

Fig. 7 shows the learning curves for different architectures designed for the block stacking tasks. These learning curves do not reflect final performance: for each evaluation point, we sample tasks and demonstrations from training data, reset the environment to the starting point of some particular stage (so that some blocks are already stacked), and only run the policy for up to one stage. If the training algorithm is DAGGER, these sampled trajectories are annotated and added to the training set. Hence this evaluation does not evaluate generalization. We did not perform full evaluation as training proceeds, because it is very time consuming: each evaluation requires tens of thousands of

¹Note that the actual ratio of misinterpreted demonstrations may be different, since the runs that have caused a manipulation failure could later lead to a wrong move, were it successfully executed. On the other hand, by visually inspecting the videos, we observed that most of the trajectories categorized as “Wrong Move” are actually due to manipulation failures (except for policy conditioning on the final state, which does seem to occasionally execute an actual wrong move).

trajectories across over > 100 tasks. However, these figures are still useful to reflect some relative trend.

From these figures, we can observe that while conditioning on full trajectories gives the best performance which was shown in the main text, it requires much longer training time, simply because conditioning on the entire demonstration requires more computation. In addition, this may also be due to the high variance of the training process due to downsampling demonstrations, as well as the fact that the network needs to learn to properly segment the demonstration. It is also interesting that conditioning on snapshots seems to learn faster than conditioning on just the final state, which again suggests that conditioning on intermediate information is helpful, not only for the final policy, but also to facilitate training. We also observe that learning happens most rapidly for the initial stages, and much slower for the later stages, since manipulation becomes more challenging in the later stages. In addition, there are fewer tasks with more stages, and hence the later stages are not sampled as frequently as the earlier stages during evaluation.

B.6 Exact Performance Numbers

Exact performance numbers are presented for reference:

- Table 3 and Table 4 show the success rates of different architectures on training and test tasks, respectively;
- Table 5 shows the success rates across all tasks as the number of ensembles is varied;
- Table 6 shows the success rates of tasks that are equivalent to $abcd$ up to permutations;
- Table 7, Table 8, Table 9, Table 10, and Table 11 show the breakdown of different success and failure scenarios for all considered architectures.

B.7 More Visualizations

Fig. 8 and Fig. 9 show the full set of heatmaps of attention weights. Interestingly, in Fig. 8, we observe that rather than attending to two blocks at a time, as we originally expected, the policy has learned to mostly attend to only one block at a time. This makes sense because during each of the grasping and the placing phase of a single stacking operation, the policy needs to only pay attention to the single block that the gripper should aim towards. For context, Fig. 10 and Fig. 11 show key frames of the neural network policy executing the task.

References

- Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–901, 2016.

#Landmarks	Plain LSTM	LSTM with attention	Final state with attention
2	100.0%	100.0%	100.0%
3	100.0%	100.0%	100.0%
4	99.0%	100.0%	100.0%
5	98.0%	100.0%	100.0%
6	99.0%	100.0%	100.0%
7	98.0%	100.0%	100.0%
8	93.9%	99.0%	100.0%
9	83.8%	94.9%	100.0%
10	50.5%	85.9%	100.0%

Table 2: Success rates of particle reaching conditioned on unseen demonstrations, and running on unseen initial configurations.

#Stages	Demo	DAGGER	BC	Snapshot	Final state
1	99.1%	99.1%	99.1%	97.2%	98.8%
2	95.6%	94.3%	93.7%	92.6%	86.7%
3	88.5%	88.0%	86.9%	86.7%	84.8%
4	78.6%	78.2%	76.7%	76.4%	71.9%
5	67.3%	65.9%	65.4%	62.5%	60.6%
6	55.7%	51.5%	52.4%	47.0%	43.6%
7	42.8%	34.3%	37.5%	31.4%	31.5%

Table 3: Success rates of different architectures on training tasks of block stacking.

#Stages	Demo	DAGGER	BC	Snapshot	Final state
2	95.8%	94.9%	95.9%	92.8%	94.1%
4	77.6%	77.0%	74.8%	77.2%	75.8%
5	65.9%	65.9%	64.3%	61.1%	51.9%
6	49.4%	50.6%	46.5%	42.6%	35.9%
7	46.5%	36.5%	38.5%	32.8%	32.0%
8	29.0%	18.0%	24.0%	19.0%	20.0%

Table 4: Success rates of different architectures on test tasks of block stacking.

#Stages	1 Ens.	2 Ens.	5 Ens.	10 Ens.	20 Ens.
1	91.9%	95.4%	98.8%	99.1%	98.7%
2	92.3%	92.2%	94.5%	94.6%	94.1%
3	86.0%	86.8%	87.9%	88.0%	87.9%
4	76.6%	77.4%	77.9%	78.0%	78.3%
5	65.1%	65.0%	65.3%	65.9%	65.5%
6	49.0%	50.4%	50.1%	51.3%	50.8%
7	34.4%	36.1%	36.0%	34.9%	36.8%
8	20.0%	21.0%	21.0%	18.0%	20.0%

Table 5: Success rates of varying number of ensembles using the DAGGER policy conditioned on full trajectories, across both training and test tasks.

Task ID	Success Rate
abcd	83.0%
abdc	86.0%
acbd	92.0%
acdb	84.0%
adbc	91.0%
adcb	88.0%
bacd	92.0%
badc	90.0%
bcad	92.0%
bcda	88.0%
bdac	94.0%
bdca	88.0%
cabd	82.0%
cadb	87.0%
cbad	95.0%
cbda	87.0%
cdab	91.0%
cdba	93.0%
dabc	90.0%
dacb	92.0%
dbac	88.0%
dbca	90.0%
dcab	91.0%
dcba	84.0%

Table 6: Success rates of a set of tasks that are equivalent up to permutations, using the DAGGER policy conditioned on full trajectories.

#Stages	Success	Recoverable failure	Manipulation failure	Wrong move
1	99.3%	0.0%	0.7%	0.0%
2	95.9%	0.4%	3.7%	0.0%
3	89.1%	0.7%	10.1%	0.1%
4	79.2%	1.2%	19.4%	0.1%
5	67.5%	1.4%	30.9%	0.2%
6	55.2%	1.4%	43.1%	0.3%
7	44.6%	1.7%	53.2%	0.6%
8	30.9%	4.3%	64.9%	0.0%

Table 7: Breakdown of success and failure scenarios for Demo policy.

#Stages	Success	Recoverable failure	Manipulation failure	Wrong move
1	99.4%	0.0%	0.6%	0.0%
2	95.3%	0.9%	3.8%	0.0%
3	89.1%	1.9%	8.8%	0.1%
4	79.5%	3.5%	16.7%	0.3%
5	69.1%	5.0%	25.6%	0.3%
6	55.8%	7.3%	36.4%	0.5%
7	39.0%	8.6%	51.5%	0.8%
8	21.2%	14.1%	62.4%	2.4%

Table 8: Breakdown of success and failure scenarios for DAGGER policy.

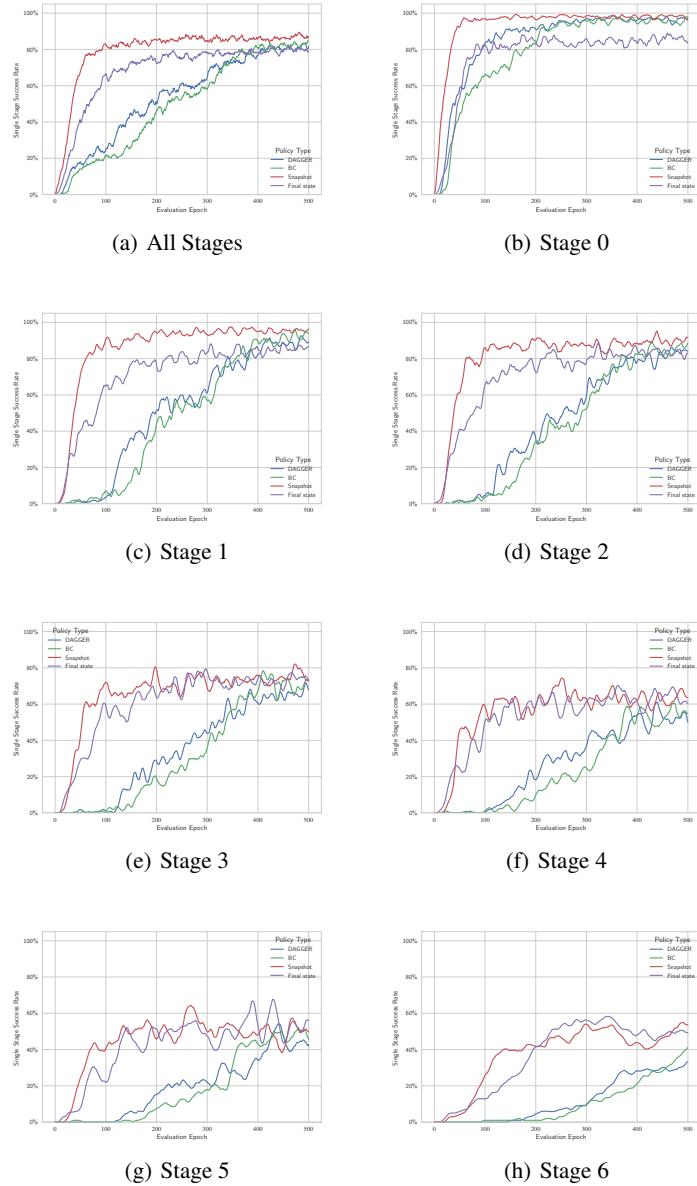


Figure 7: Learning curves of block stacking task. The first plot shows the average success rates over initial configurations of all stages. The subsequent figures shows the breakdown of each stage. For instance, “Stage 3” means that the first 3 stacking operations are already completed, and the policy is evaluated on its ability to perform the 4th stacking operation.

#Stages	Success	Recoverable failure	Manipulation failure	Wrong move
1	99.6%	0.0%	0.4%	0.0%
2	95.6%	1.1%	3.2%	0.1%
3	88.1%	2.2%	9.5%	0.2%
4	78.5%	4.5%	16.8%	0.2%
5	67.2%	6.6%	25.7%	0.4%
6	53.9%	8.3%	37.1%	0.6%
7	40.6%	9.8%	48.7%	0.9%
8	27.0%	13.5%	58.4%	1.1%

Table 9: Breakdown of success and failure scenarios for BC policy.

#Stages	Success	Recoverable failure	Manipulation failure	Wrong move
1	99.1%	0.0%	0.9%	0.0%
2	94.5%	1.6%	3.8%	0.1%
3	88.0%	2.5%	9.3%	0.2%
4	78.9%	4.6%	16.2%	0.3%
5	65.6%	8.0%	25.8%	0.6%
6	50.8%	8.3%	40.2%	0.7%
7	36.1%	9.2%	54.2%	0.4%
8	21.6%	11.4%	65.9%	1.1%

Table 10: Breakdown of success and failure scenarios for Snapshot policy.

#Stages	Success	Recoverable failure	Manipulation failure	Wrong move
1	99.2%	0.0%	0.8%	0.0%
2	95.1%	1.3%	3.6%	0.0%
3	86.7%	2.5%	9.7%	1.1%
4	75.2%	4.0%	18.3%	2.5%
5	60.5%	4.3%	31.2%	4.0%
6	45.5%	4.7%	45.5%	4.3%
7	34.9%	5.6%	57.3%	2.2%
8	24.1%	3.6%	72.3%	0.0%

Table 11: Breakdown of success and failure scenarios for Final state policy.

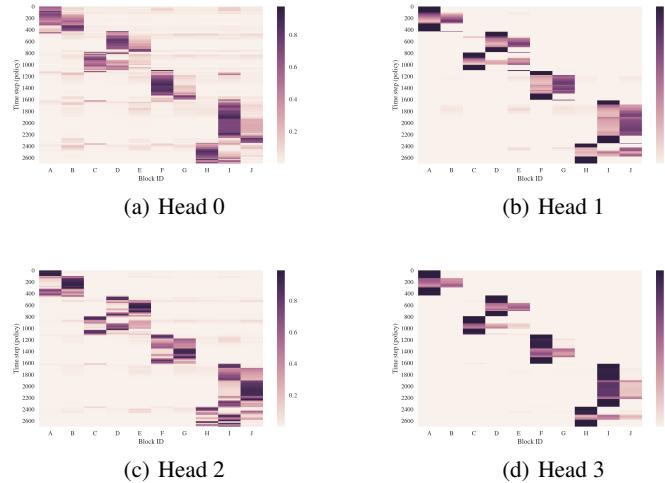


Figure 8: Heatmap of attention weights over different blocks of all 4 query heads.

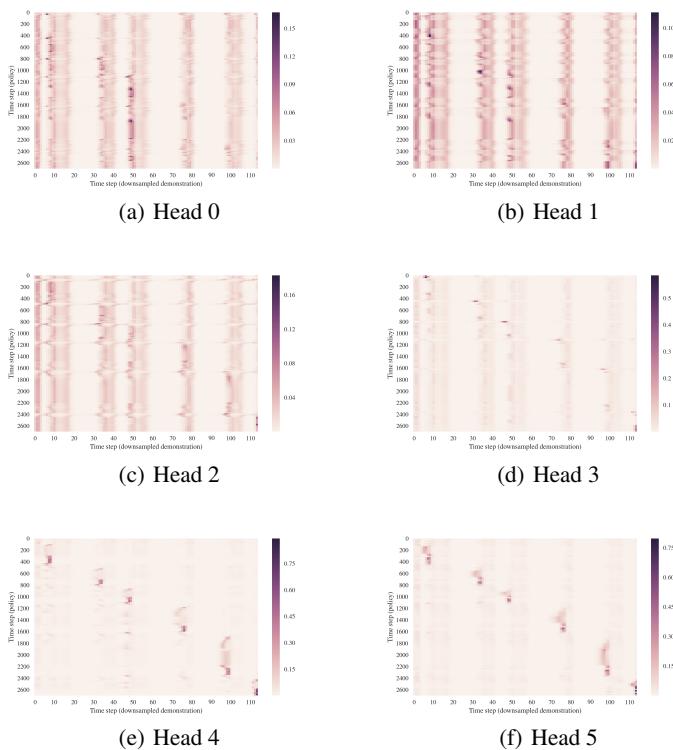


Figure 9: Heatmap of attention weights over downsampled demonstration trajectory of all 6 query heads. There are 2 query heads per step of LSTM, and 3 steps of LSTM are performed.

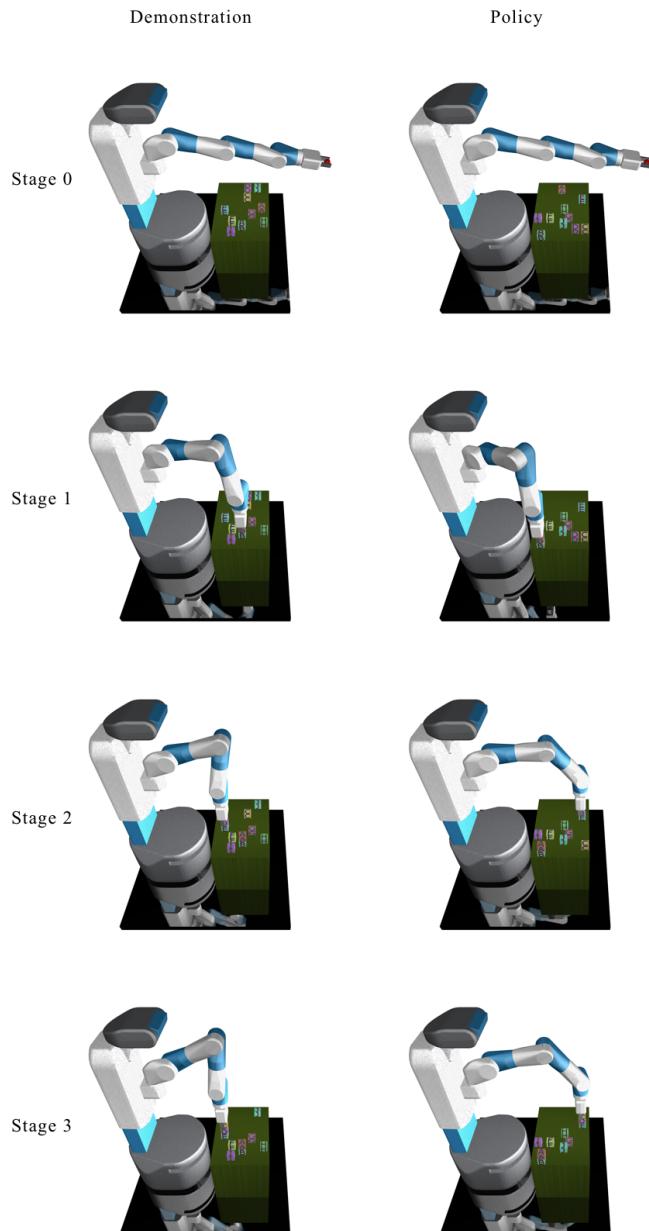


Figure 10: Illustration of the task used for the visualization of attention heatmaps (first half). The task is ab cde fg hij. The left side shows the key frames in the demonstration. The right side shows how, after seeing the entire demonstration, the policy reproduces the same layout in a new initialization of the same task.

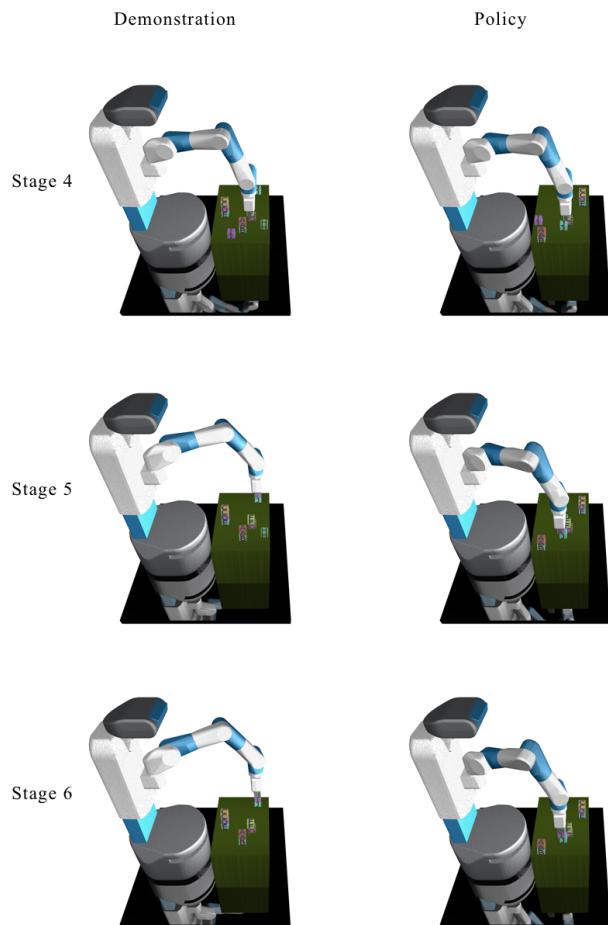


Figure 11: Illustration of the task used for the visualization of attention heatmaps (second half). The task is ab cde fg hij. The left side shows the key frames in the demonstration. The right side shows how, after seeing the entire demonstration, the policy reproduces the same layout in a new initialization of the same task.