A3C-S: Automated Agent Accelerator Co-Search towards Efficient Deep Reinforcement Learning

Abstract—Driven by the explosive interest in applying deep reinforcement learning (DRL) agents for numerous real-time control and decision-making applications, there has been a growing demand to deploy DRL agents to empower daily-life intelligent devices, while the prohibitive complexity of DRL stands at the odds with the limited on-device resources. In this work, we propose an Automated Agent Accelerator Co-Search (A3C-S) framework, which to our best knowledge is the first to automatically co-search the optimally matched DRL agents and accelerators that maximize both the test scores and hardware efficiency. Extensive experiments consistently validate the superiority of our A3C-S over state-of-the-art techniques.

Index Terms—Network Accelerator Co-design, Deep Reinforcement Learning, AutoML

I. INTRODUCTION

Recent successes of deep reinforcement learning (DRL) [1], which integrates reinforcement learning (RL) and deep neural networks (DNNs), has triggered tremendous enthusiasm to develop and deploy DRL-powered intelligence into numerous inference and control applications, including robotics and autonomous vehicles. Many of them, such as autonomous vehicles, require real-time control and decision-making policies for which the DRL agents have to derive real-time policies using real-time data for dynamic systems. However, real-time control and decision-making for DRL can be prohibitively challenging in many real-world applications due to DRL’s integrated complex DNNs and edge devices’ constrained resources, calling for DRL-powered intelligent solutions that favor both the test score and processing efficiency.

To address the large gap between the growing need for on-device DRL and DRL’s prohibitive complexity, emerging network and accelerator co-exploration (NACoS) methods [2], [3] are promising as they can boost DNN acceleration efficiency. However, directly applying NACoS methods to design DRL agents can easily fail due to the commonly observed vulnerability and instability of DRL training, which occurs with a high variance [4]. Furthermore, such instability will be further exacerbated when considering differentiable NAS (DNAS) based NACoS methods, which requires competitively low search cost and thus can enable efficient navigation over the large DRL agent and accelerator joint space, because the success of NAS requires unbiased gradient estimation with low variance. To this end, we aim to develop a novel NACoS framework dedicated to DRL agent acceleration to promote fast development and highly efficient DRL-powered solutions. Specifically, we make the following contributions:

• We propose an Automated Agent Accelerator Co-Search (A3C-S) framework, which to our best knowledge is the first to automatically co-search the optimally matched DRL agents and accelerators that maximize both the test scores and hardware efficiency.
• A3C-S integrates and demonstrates the first DNAS search dedicated to DRL that features a novel distillation mechanism to effectively stabilize the agent search, despite the instability of DRL training.
• A3C-S incorporates a parameterized micro-architecture with over $10^{27}$ searchable choices of accelerators and dataflows to enable differentiable search for DRL agent accelerators. A3C-S is generic and can be applied on top of different accelerator templates.
• Through FPGA measurements, extensive experiments and ablation studies validate A3C-S’s effectiveness in generating efficient DRL agents/accelerators that consistently outperform state-of-the-art (SOTA) agents/accelerators.

II. RELATED WORKS

Deep reinforcement learning. DRL integrates traditional RL algorithms with DNNs to handle higher-dimensional and more complex problems, e.g., DQN [1] introduces DNNs to Q-Learning and [5] utilizes DNNs to model both the actor and critic in AC-based DRL. More DRL works can be found in [6]. Despite DRL’s promising success, automating the design of DRL agents has not yet been explored, while there is a growing need for fast development of DRL-powered solutions.

Hardware-aware neural architecture search. NAS has been widely adopted to automate the design of efficient DNNs. To tackle the prohibitive search cost of previous RL-based NAS methods, DNAS [7], [8] has gained more attention thanks to its excellent search efficiency. However, the possibility of applying NAS or DNAS in DRL hasn’t been explored.

DNN accelerator design. SOTA DNN accelerators [9] tackle the prohibitive complexity of DNNs through novel micro-architectures/dataflows to maximize data reuses, and thus improve the acceleration efficiency. Early works require experts’ manual design, and thus is time-consuming. Later, various design flow and automation tools [10], [11] have been developed. However, they mostly explore DNN acceleration, leaving automated DRL accelerator design unexplored.

Network and accelerator co-search. Jointly exploring DNNs and their accelerators is very promising towards efficient DNN solutions, as shown in pioneering works [12], [13], where the former suffers from large search time and the latter considers a limited search space. These works motivate us to explore jointly searching for DRL agents and their accelerators to maximize the test scores and hardware efficiency.
III. PRELIMINARIES OF DRL

Here we describe the preliminaries of DRL.

RL can be viewed a Markov Decision Process (MDP) determined by a tuple \((S,A,T,R,\gamma)\), where \(S\) is the state space, \(A\) is the action space, \(T(s'|s,a)\) is the transition probability of ending up in state \(s'\) when executing the action \(a\) in the state \(s\), \(R\) is the reward function, and \(\gamma\) is a discount factor. A policy \(\pi(s_t,a_t) = p(a_t|s_t)\) defines the probability for an agent in the MDP to execute the action \(a\) in the state \(s\). In particular, the agent performs the actions \(a_t \in A\) sampled from the policy \(\pi(s_t,a_t)\) at each time step \(t\) and the state \(s_t \in S\), leading to the next state \(s_{t+1} \in S\) and acquiring a reward \(r_t\).

In DRL, a policy is parameterized by \(\theta\), i.e., the weights of a DNN, and the agent’s goal is learning an optimal policy to maximize the expected cumulative reward:

\[
\theta^*_\pi = \arg \max_{\theta} J(\pi(\cdot|\theta_\pi)) = \arg \max_{\theta} \mathbb{E}_\pi \left[ \sum_{t=0}^H \gamma^t r_t \right]
\]  

where \(H\) is the time horizon, and \(\gamma\) is a discount factor.

To solve the optimization problem in Eq. 1, stochastic policy gradient methods [14] is widely adopted where \(\nabla_{\theta_\pi} J(\pi(\cdot|\theta_\pi))\):

\[
\nabla_{\theta_\pi} J(\pi(\cdot|\theta_\pi)) = \mathbb{E}_\pi \left[ \sum_{t=0}^H \delta_t \nabla_{\theta_\pi} \log \pi(a_t|s_t|\theta_\pi) \right]
\]  

In this work, we adopt the temporal difference error (td-error) \(\delta_t = r_t + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)\) to reduce the variance in policy gradients, where \(V_\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^H \gamma^t r_t | s_0 = s \right]\) is the value function which estimates the expected cumulative reward of the policy \(\pi\) starting from the state \(s\). Since the value function \(V_\pi\) is hard to estimate, AC-based DRL methods [15] parameterize the value function (i.e., the critic) with the learnable parameter \(\theta_v\), i.e. the weights of a DNN in DRL. The objective of \(\theta_v\) is to minimize the td-error of the estimated value between consecutive states:

\[
\theta_v^* = \arg \min_{\theta_v} \mathbb{E}_\pi \left[ \sum_{t=0}^H \frac{1}{2} (r_t + \gamma V_\pi(s_{t+1}|\theta_v) - V_\pi(s|\theta_v))^2 \right]
\]  

Therefore, in the AC framework, the actor and critic parameterized by \(\theta_\pi\) and \(\theta_v\), respectively can be iteratively updated to lead the agent towards an optimal policy.

IV. THE PROPOSED A3C-S FRAMEWORK

There exist three main challenges for designing A3C-S: (1) the huge joint search space, (2) the non-differentiable accelerator parameters and the gap between DNA's required layer-wise hardware-cost penalty and the optimal accelerators' dependency on all the layers, and (3) the training instability with the high variance of DRL which may lead to the failure of applying NAS to search DRL agents. In this section, we first introduce the component techniques of A3C-S in Sec. IV-A that tackles the first two challenges with a novel differentiable search strategy and a differentiable accelerator search engine, and then A3C-S’s AC-distillation mechanism in Sec. IV-B to tackle the third challenge.

A. A3C-S: the co-search pipeline

**A3C-S formulation.** We formulate A3C-S as below:

\[
\min_{\theta_\pi,\theta_v,\alpha} L_{task}(\theta_\pi, \theta_v, net(\alpha)) + \lambda L_{cost}(hw(\phi^*), net(\alpha))
\]  

s.t. \(\phi^* = \arg \min_{\phi} L_{cost}(hw(\phi), net(\alpha))\)

where \(\alpha\) and \(\phi\) are the variables maintaining the probability of choosing different (1) network operators and (2) accelerator parameters, \(\theta_\pi\) and \(\theta_v\) are the supernet weights of the actor and critic in an AC-based DRL, respectively. \(L_{task}\) and \(L_{cost}\) are the task loss (see Sec. IV-B for details) and the hardware-cost loss, respectively, and \(net(\alpha)\) and \(hw(\phi)\) denote the network and accelerator parameterized by \(\alpha\) and \(\phi\), respectively.

**A3C-S’s co-search pipeline.** During co-search, A3C-S starts by updating the accelerator parameters \(\phi\) given the current network structure \(net(\alpha)\), and then updates \(\theta_\pi, \theta_v, \alpha\) in the same iteration based on the accelerator \(hw(\phi^*)\) resulting from the previous step. Our A3C-S adopts one-level optimization [16] instead of bi-level optimization [7], considering that the one-step SGD approximation of bi-level optimization will lead to biased gradient estimation [17] which can largely suffer from the high variance of DRL training as evaluated in Sec. V-D. The updates of \(\alpha\) and \(\phi\) follow:

**Forward**: \(A_{l+1} = \sum_{i=1}^N GS_{hard}(\alpha_i^l) O_i(A^l) = O_{fw}^l(A^l)\)

**Backward**: \(\frac{\partial L_{task}}{\partial \alpha_i^l} = \sum_{k=1}^K \frac{\partial L_{task}}{\partial GS(\alpha_k^l)} \frac{\partial GS(\alpha_k^l)}{\partial \alpha_i^l}\)

\(\frac{\partial L_{cost}}{\partial \alpha_i^l} = 1 (GS_{hard}(\alpha_i^l) = 1) L_{cost}(hw(\phi^*), net(\alpha_i^l))\)

where \(A^l\) and \(A^{l+1}\) are the feature maps of the \(l\)-th and \((l+1)\)-th layer, respectively, \(GS_{hard}\) is the hard Gumbel Softmax operator generating a one-hot output, i.e., only one operator \(O_{fw}^l\), will be activated during forward, \(N\) is the total number of operator choices, and \(O_i^l\) is the \(i\)-th operator in the \(l\)-th layer parameterized by \(\alpha_i^l\). Meanwhile, \(GS\) is a Gumbel Softmax function and \(K\) is the number of activated paths with the top \(K\) probability, and similar to [18], \(K \in (1,N)\) in A3C-S to control the computational cost. In Eq. 8, \(\mathbb{1}\) is an indicator denoting whether \(\alpha_i^l\) (i.e., the \(i\)-th operator in the \(l\)-th layer) is activated during forward.

**A3C-S’s co-search strategy.** A3C-S integrates a novel search strategy to solve Eq. 4 for effective yet efficient search to avoid memory explosion due to the large joint search spaces.

Single-path forward: (see Eq. 6) A3C-S adopts hard Gumbel Softmax sampling [19], i.e., only the choice with the highest probability will be activated to narrow the gap between the supernet and the finally derived network thanks to the single-path property of hard Gumbel Softmax sampling.
Multi-path backward: (see Eq. 7) A3C-S activates multiple paths to approximate the gradients of \( \alpha \) via Gumbel Softmax relaxation to balance the search efficiency (prefer fewer activated paths) and stability (prefer more activated paths), inspired by [18] which targets DNAs for DNNs.

Hardware-cost penalty: The network search in Eq. 4 requires layer-wise hardware-cost penalties assuming the layer-wise operators running on the final optimal accelerator \( hw(\phi^*) \), which is not yet available at each co-search epoch as the optimal network is still unknown, i.e., the chicken-and-egg problem. To handle this, we approximate the layer-wise hardware-cost by assuming that the single-path network derived during each forward is close to the final derived network, since the network operators that have higher probabilities are also more likely to appear in the final optimal network.

A3C-S’s Differentiable accelerator search (DAS). EDD [13] made a pioneering effort to differentially co-search the network and accelerator, yet their accelerator search space is limited to the parallel factor of their template, which can be analytically fused into their computational cost, whereas this is not always applicable to other naturally non-differentiable accelerator design knobs such as PE numbers and buffer allocation strategies. A more general search engine is desirable towards generic DAS.

A3C-S’s accelerator search algorithm: We propose a general DAS engine to efficiently search for the optimal accelerator, including the micro-architectures and dataflows, given a DNN based on the single-path sampling in Eq. 6, i.e.:

\[
\phi^* = \arg \min_{\phi} \sum_{m=1}^{M} GS_{hard}(\phi^m) \hat{L}
\]  

(9)

where \( \hat{L} = L_{cost}(hw(\{GS_{hard}(\phi^m)\}), \text{net}(\alpha)) \)

where \( M \) is the number of accelerator parameters. Given the network \( \text{net}(\alpha) \) which is the most likely network sampled during the single-path forward, the search engine utilizes hard Gumbel Softmax \( GS_{hard}(\phi^m) \) sampling on each design parameter \( \phi^m \) to build an accelerator \( hw(\{GS_{hard}(\phi^m)\}) \) and penalize each sampled accelerator parameter with the overall hardware-cost \( L_{cost} \) through relaxation in a differentiable manner.

A3C-S’s accelerator template: We adopt a parameterized accelerator template built upon a SOTA chunk-based pipeline micro-architecture [20]. The accelerator template comprises multiple sub-accelerators (i.e., chunks) and executes DNNs in a pipeline fashion. In particular, each chunk is assigned with multiple but not necessarily consecutive layers which are executed sequentially within the chunk. Similar to Eyeriss, each chunk consists of levels of buffers/memories (e.g., on-chip buffer and local register files) and processing elements (PEs) to facilitate data reuses and parallelism with searchable accelerator parameters, including PE interconnections (i.e., Network-on-chip), buffer sizes, and MAC operations’ scheduling and tiling (i.e., dataflows) (see more details in Sec. V-A).

B. A3C-S: the AC-distillation mechanism

Motivation. Policy distillation [21] shows that the distillation from a teacher agent can effectively reduce the variance of gradient estimates and stabilize the training process of the student agent, motivating us to introduce the distillation mechanism to stabilize the DNAs process for DRL. However, vanilla policy distillation merely distills the policy without considering the value function which can play a critical role in both assisting the policy updates and reducing the variance of vanilla policy gradients. We conjecture that further distilling the value function from the teacher agent can better improve the training stability and the convergence.

A3C-S’s AC-distillation. In A3C-S, we propose an AC-distillation mechanism to distill knowledge from both the actor and critic of a pretrained teacher agent to the student agent, where the two distillation losses for the actor and critic are:

\[
L_{actor}^{\text{distill}} = \mathbb{E}_{\pi} \left[ \sum_{t=0}^{H} \pi(a_t, s_t|\theta^{tea}_\pi) \log \frac{\pi(a_t, s_t|\theta^{tea}_\pi)}{\pi(a_t, s_t|\theta^{stu}_\pi)} \right] 
\]

(10)

\[
L_{critic}^{\text{distill}} = \mathbb{E}_{\pi} \left[ \sum_{t=0}^{H} \frac{1}{2} (V_\pi(s_t|\theta^{stu}_\pi) - V_\pi(s_t|\theta^{tea}_\pi))^2 \right] 
\]

(11)

where \( \pi(a_t, s_t|\theta^{tea}_\pi) \) and \( \pi(a_t, s_t|\theta^{stu}_\pi) \) are the actor and student actor, respectively, and \( V_\pi(s_t|\theta^{tea}_\pi) \) and \( V_\pi(s_t|\theta^{stu}_\pi) \) are the teacher and student critic, respectively. We adopt KL divergence to distill the knowledge from the teacher actor following [21] and the MSE loss as a soft constraint to enforce the student critic to mimic the estimated value of the teacher critic. The final objective during both search and training is:

\[
L_{task} = L_{policy} + L_{value} + \beta_1 L_{entropy} + \beta_2 L_{actor}^{\text{distill}} + \beta_3 L_{critic}^{\text{distill}} 
\]

(12)

where \( \beta_1, \beta_2, \) and \( \beta_3 \) are the weighted coefficients. Here \( L_{policy} \) is the policy gradient loss as in [14], \( L_{value} \) is the value loss based on the td-error, and \( L_{entropy} \) is the entropy loss on top of the policy to encourage exploration, i.e.:

\[
L_{policy} = \mathbb{E}_\pi \left[ - \sum_{t=0}^{H} \delta_t \log(\pi(a_t, s_t|\theta^{stu}_\pi)) \right] 
\]

(13)

\[
L_{value} = \mathbb{E}_\pi \left[ \sum_{t=0}^{H} \frac{1}{2} (r_t + \gamma V_\pi(s_{t+1}|\theta^{stu}_\pi) - V_\pi(s_t|\theta^{stu}_\pi))^2 \right] 
\]

(14)

\[
L_{entropy} = \mathbb{E}_\pi \left[ \sum_{t=0}^{H} \pi(a_t, s_t|\theta^{stu}_\pi) \log(\pi(a_t, s_t|\theta^{stu}_\pi)) \right] 
\]

(15)

The search algorithm of A3C-S is summarized in Alg. 1.

V. EXPERIMENT RESULTS

In this section, we first introduce our experiment setting, and then present ablation studies to evaluate A3C-S’s component techniques and evaluation of our A3C-S framework.
A. Experiment setup

Models and tasks. We evaluate the performance of the AC-based DRL when its feature extractor backbone adopts the searched networks from A3C-S or five SOTA networks with different model sizes, including the small network in DQN [1] (termed as Vanilla) and ResNet-14/-20/-38/-74, on Atari 2600 games based on the Arcade Learning Environment. For all the ResNets, we modify the stride of the first convolution to be 2 and the output dimension of the final FC layer to be 256 to adapt them to the 84x84 resolution of Atari games.

Training settings. We use the same training and test hyper-parameters settings for all the models on all the tasks in this paper. Specifically, we train a DRL agent on each task for 3e7 steps with a discount factor ($\gamma$ in Eq. 1) of 0.99 and a rollout length of 5; we use the RMSProp optimizer [1] with an initial learning rate of 1e-3 which keeps constant in the first 1e7 steps and then linearly decays to 1e-4; and the reported test score is averaged on 30 episodes with null-op starts following [1].

Distillation Settings. For the proposed AC-distillation, we train a ResNet-20 model as the teacher agent and $\beta_1$, $\beta_2$, and $\beta_3$ in Eq. 12 are set to be 1e-2, 1e-1, and 1e-3, respectively.

Algorithm 1 Automated Agent Accelerator Co-Search (A3C-S)

Initialize the step counter $t \leftarrow 1$
repeat
$t_{\text{start}} = t$
Get state $s_t$
repeat
Perform $a_t \sim \pi(s_t, s_t|\theta^{s_{\text{stu}}})$ based on Eq. 6
Receive reward $r_t$ and new state $s_{t+1}$
$t \leftarrow t + 1$
until terminal $s_t$ or $t = t_{\text{start}}$ $\Rightarrow$ rollout length $L$
Update $\phi$ in Eq. 9 to acquire $\phi^*$
for $i \in \{t_{\text{start}}, \ldots, t - 1\}$ do
$\delta_t = r_t + \gamma V_s(s_{t+1}|\phi^{s_{\text{stu}}}) - V_s(s_t|\phi^{s_{\text{stu}}})$
Calculate $L_{\text{task}}$ in Eq. 12 based on $\delta_t$, $\pi(\phi(s_t|\theta^{\text{stu}}))$, and $V_s(s_t|\theta^{\text{stu}})$
Calculate $L_{\text{cost}}$ in Eq. 4 based on $\phi^*$
Update $\theta^s_{\text{stu}}$; $\theta^r_{\text{stu}} \leftarrow \theta^r_{\text{stu}} - \eta_t \nabla_{\theta^r_{\text{stu}}} L_{\text{task}}$
Update $\theta^r_{\text{stu}}$; $\theta^r_{\text{stu}} \leftarrow \theta^r_{\text{stu}} - \eta_t \nabla_{\theta^r_{\text{stu}}} L_{\text{task}}$
Update $\alpha$; $\alpha \leftarrow \alpha - \eta_t \nabla_{\alpha} (L_{\text{task}} + \lambda L_{\text{cost}})$
end for
until $t > T_{\text{max}}$
Derive the final agent and accelerator with the highest $\alpha$ and $\phi$ respectively
return the final agent and accelerator

B. Ablation study: DRL with different model sizes

Observations and Analysis. Fig. 1 visualizes the test core evolution during the training process on various Atari games, when adopting different networks for the DRL agents, where the highest achieved test scores is listed in Tab. I. Two observations can be made. First, networks with a larger size in general favors the achieved test scores especially in all the experiments.

Search settings. The supernet structure follows the network design (i.e., #groups and stride) of the ResNet series with 12 sequential searchable cells. The candidate operators are standard convolutions with a kernel size 3/5, inverted residual blocks with a kernel size 3/5, a channel expansion of 1/3/5, and skip connections, leading to a search space of $9^{12}$ choices. We update the architecture parameters using an Adam optimizer with a momentum of 0.9 and a fixed learning rate of 1e-3. The initial temperature [8] in Gumbel Softmax is set to 5 and decayed by 0.98 every 1e5 steps.

Accelerator settings. To evaluate A3C-S’s generated accelerators, we adopt a standard FPGA design flow, i.e., the Vivado HLS flow [23], and performance metric, i.e., frame per second (FPS). When searching for the accelerator parameter, A3C-S makes use of a SOTA accelerator performance predictor [24] to obtain fast and reliable estimation during search. The accelerator parameters we optimize upon include 1) parallel processing elements (PE) settings: the number and inter-connections of PEs, 2) buffer management: allocation of lower level memories between inputs, weights and outputs, 3) tiling and scheduling of MAC (Multiply Accumulate) computations, and 4) layer allocation: ways to assign each layer to the corresponding pipeline stage (sub-accelerator).

Fig. 1. Test scores averaged over 30 episodes during the training of five models on four Atari games.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakout</td>
<td>523.7</td>
<td>776.5</td>
<td>811.0</td>
<td>818.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Alien</td>
<td>1724.1</td>
<td>9007.2</td>
<td>9323.2</td>
<td>8829.0</td>
<td>4456.3</td>
</tr>
<tr>
<td>Asterix</td>
<td>4850.0</td>
<td>70850.0</td>
<td>856800.0</td>
<td>756120.0</td>
<td>539060.0</td>
</tr>
<tr>
<td>Atlantis</td>
<td>306320.0</td>
<td>3127390.0</td>
<td>3156130.0</td>
<td>3181090.0</td>
<td>3046490.0</td>
</tr>
<tr>
<td>TimePilot</td>
<td>4780.0</td>
<td>9070.0</td>
<td>9680.0</td>
<td>9500.0</td>
<td>9040.0</td>
</tr>
<tr>
<td>SpaceInvaders</td>
<td>1171.1</td>
<td>9848.0</td>
<td>46870.0</td>
<td>17962.0</td>
<td>15111.0</td>
</tr>
<tr>
<td>WizardOfWor</td>
<td>1320.0</td>
<td>2690.0</td>
<td>3580.0</td>
<td>3160.0</td>
<td>1850.0</td>
</tr>
<tr>
<td>Tennis</td>
<td>-23.1</td>
<td>13.8</td>
<td>11.5</td>
<td>19.6</td>
<td>19.3</td>
</tr>
<tr>
<td>Asteroids</td>
<td>2095.0</td>
<td>5690.0</td>
<td>5744.0</td>
<td>1947.0</td>
<td>4792.0</td>
</tr>
<tr>
<td>Assault</td>
<td>10164.0</td>
<td>14470.0</td>
<td>17314.0</td>
<td>12406.5</td>
<td>9649.0</td>
</tr>
<tr>
<td>BattleZone</td>
<td>7600.0</td>
<td>5800.0</td>
<td>13100.0</td>
<td>13300.0</td>
<td>4100.0</td>
</tr>
<tr>
<td>BeamRider</td>
<td>5530.0</td>
<td>23984.0</td>
<td>25961.0</td>
<td>29498.0</td>
<td>30048.0</td>
</tr>
<tr>
<td>Bowling</td>
<td>28.1</td>
<td>53.0</td>
<td>59.2</td>
<td>33.2</td>
<td>50.8</td>
</tr>
<tr>
<td>Boxing</td>
<td>4.2</td>
<td>100.0</td>
<td>100.0</td>
<td>99.3</td>
<td>87.1</td>
</tr>
<tr>
<td>Centipede</td>
<td>5025.0</td>
<td>6690.0</td>
<td>6410.0</td>
<td>6384.6</td>
<td>6899.0</td>
</tr>
<tr>
<td>ChopperCommand</td>
<td>1320.0</td>
<td>11170.0</td>
<td>14910.0</td>
<td>4370.0</td>
<td>8240.0</td>
</tr>
</tbody>
</table>
on more difficult tasks (e.g., BeamRider of Tab. I), since larger networks can provide higher test scores with the same training time steps in most of the games over the smaller vanilla-network and ResNet-14; Second, there always exists a task-specific optimal network size (i.e., further increase won’t improve or even degrade the test score), which is likely due to the increased difficulty of training larger DRL agents. For instance, the vanilla network performs well in the Atlantis game (see Fig. 1), whereas ResNet-38 merely offers a marginally improved score at a cost of 13.7× higher FLOPs (floating-point operations); and ResNet-74 is inferior than ResNet-20/38 in most of these experiments, because it is more difficult to be trained within the limited training steps. Note that even though more training steps or better-tuned hyper-parameters may improve the convergence of DRL with ResNet-74, the associated inefficiency compared to more efficient agents makes it impractical to be widely adopted.

**Extracted Insights.** The observation and analysis above imply that (1) DNNs’ architecture plays a critical role in DRL, which is yet under-explored in existing works, and (2) designing task-specific agents is highly desired in optimally balancing the test score and processing efficiency for different tasks. Given that manual design of dedicated agents for different tasks is not practical to handle the growing need for the fast development of DRL-powered solutions for numerous applications, motivating the use of NAS to design DRL agents.

**Observations and Analysis.** We compare our AC-distillation mechanism with the three baselines by applying them to the vanilla network and ResNet-14 evaluated on Atarix games (see Tab. II). Three observations can be made.

First, compared to the designs without distillation, distillation strategies in general favor the test scores, which is consistent with [21]; Second, our AC-distillation mechanism consistently performs the best in achieving the highest test scores on most of tasks, among the three distillation strategies.

**D. Ablation study: one-level vs. bi-level optimization**

In this set of experiments, we visualize the test score evolution in Fig. 2 during the search process of (1) directly applying NAS without distillation, (2) using AC-distillation with bi-level optimization, and (3) using AC-distillation with one-level optimization. We can see that the test scores remain low when searching with the bi-level optimization, validating our mentioned hypothesis that the supernet cannot serve as an accurate proxy to indicate the performance of the sampled subnetworks. In contrast, searching with the one-level optimization leads to a consistent improvement in the test scores during search, demonstrating the first framework that successfully makes NAS possible in DRL.

**E. Evaluating the proposed A3C-S framework**

**A3C-S vs. SOTA networks on searched accelerators.** Here we evaluate A3C-S’s searched agents by comparing their hardware performance (i.e., FPS) with the most competitive SOTA DRL agents, i.e., ResNet-14 with the best trade-off between test scores and hardware efficiency among all our experiments (see Tab. II). For a fair comparison, we use our proposed DAS engine to search for the optimal accelerators under the same search settings and to train both networks with our AC-distillation mechanism. From Fig. 3, we can see that the resulting hardware efficiency of A3C-S’s searched agents is consistently higher than that of the most competitive SOTA

---

**TABLE II**

DRL with the vanilla network and ResNet-14 under: (1) no distillation, (2) only policy distillation, and (3) our AC-distillation.

<table>
<thead>
<tr>
<th>Atari Games</th>
<th>Vanilla No distillation</th>
<th>Policy distillation only</th>
<th>AC-distillation</th>
<th>ResNet-14 No distillation</th>
<th>Policy distillation only</th>
<th>AC-distillation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alien</td>
<td>1724</td>
<td>3096</td>
<td>3419</td>
<td>9007</td>
<td>14682</td>
<td>15723</td>
</tr>
<tr>
<td>SpaceInvaders</td>
<td>1171</td>
<td>26821</td>
<td>30124</td>
<td>9848</td>
<td>76246</td>
<td>111189</td>
</tr>
<tr>
<td>Asterix</td>
<td>4850</td>
<td>59020</td>
<td>64510</td>
<td>708500</td>
<td>794870</td>
<td>849400</td>
</tr>
<tr>
<td>Asteroids</td>
<td>2095</td>
<td>4131</td>
<td>6467</td>
<td>5690</td>
<td>15371</td>
<td>15947</td>
</tr>
<tr>
<td>Assault</td>
<td>10164</td>
<td>8036.8</td>
<td>9628.5</td>
<td>14470</td>
<td>11697</td>
<td>14052</td>
</tr>
<tr>
<td>BattleZone</td>
<td>7600</td>
<td>14200</td>
<td>14400</td>
<td>5800</td>
<td>16300</td>
<td>17500</td>
</tr>
<tr>
<td>BeamRider</td>
<td>5530</td>
<td>14417</td>
<td>21519</td>
<td>23984</td>
<td>38311</td>
<td>39604</td>
</tr>
<tr>
<td>Boxing</td>
<td>4.2</td>
<td>2.8</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Centipede</td>
<td>5025</td>
<td>5800</td>
<td>6575.5</td>
<td>6690</td>
<td>7744.3</td>
<td>8056.9</td>
</tr>
<tr>
<td>ChopperCommand</td>
<td>1320</td>
<td>15900</td>
<td>19120</td>
<td>11170</td>
<td>26320</td>
<td>31190</td>
</tr>
<tr>
<td>CrazyClimber</td>
<td>118300</td>
<td>138610</td>
<td>145700</td>
<td>128710</td>
<td>135290</td>
<td>138470</td>
</tr>
<tr>
<td>DemonAttack</td>
<td>318349</td>
<td>463823</td>
<td>483490</td>
<td>481818</td>
<td>517801</td>
<td>521051</td>
</tr>
</tbody>
</table>
DRL agents, under a comparable or better test score. This set of experiments again motivate the necessity of applying NAS to search for task-specific optimal agents that balance both the test scores and model complexity.

**A3C-S searched agents on A3C-S searched accelerators vs. SOTA accelerators.** Here we evaluate A3C-S’s DAS engine by comparing the hardware efficiency (i.e., FPS) of A3C-S’s searched agents, when being accelerated by both our DAS’s generated accelerators and a SOTA DNN accelerator, DNNBuilder [25]. For a fair comparison, both accelerators adopt the optimal DRL agents searched by our A3C-S. As shown in Fig. 3, the FPS achieved by DAS’s generated accelerators consistently outperforms that of the SOTA DNN accelerator, under the same DSP limit (900; which is the largest resource in our ZC706 [26]. This set of experiments validate the need for dedicated accelerators for DRL agents, for which our A3C-S is the first demonstration.

**A3C-S vs. SOTA DRL solutions.** To evaluate the combo benefits of our A3C-S, we evaluate A3C-S’s resulting accelerators (i.e., using A3C-S’s DAS to generate optimal accelerators for A3C-S’s searched agents) over a SOTA DRL system FA3C [27], where the performance for the latter is directly obtained from the reported data of the baseline paper. From Tab. III, we can see that our A3C-S’s resulting DRL accelerators achieve $2.1 \times \sim 6.1 \times$ better FPS, while offering consistently higher test scores, as compared to the SOTA DRL system [27]. As expected, jointly searching for both the DRL agents and their accelerators lead to optimal DRL solutions, achieving both higher test scores and better hardware efficiency. Furthermore, our A3C-S’s differentiable search strategy makes it more accessible to researchers without paramount computing resources and facilitates fast development of DRL-powered solutions.

<table>
<thead>
<tr>
<th>Atari Games</th>
<th>FA3C</th>
<th>A3C-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beamrider</td>
<td>3100 / 260</td>
<td>36745 / 617.7</td>
</tr>
<tr>
<td>Breakout</td>
<td>340 / 260</td>
<td>670 / 1596.3</td>
</tr>
<tr>
<td>Pong</td>
<td>690 / 260</td>
<td>209 / 787.4</td>
</tr>
<tr>
<td>Qbert</td>
<td>6100 / 260</td>
<td>15194 / 1222.9</td>
</tr>
<tr>
<td>Seaquest</td>
<td>170 / 260</td>
<td>478940 / 778.1</td>
</tr>
<tr>
<td>SpaceInvaders</td>
<td>830 / 260</td>
<td>109417 / 535.6</td>
</tr>
</tbody>
</table>

**VI. CONCLUSION**

We propose, design, and validate a DRL agent accelerator co-search framework dubbed A3C-S, which to our best knowledge is the first to (1) automatically co-search for the optimally matched DRL agents and accelerators that maximize both the test scores and hardware efficiency, and (2) demonstrate the first successful DNAS for DRL, for which a vanilla DNAS fails due to DRL’s training instability.

**REFERENCES**