APPENDIX

We evaluate CLEAS as well as other alternative methods on numerous sequential classification tasks. The results lend great credence to the fact that CLEAS is able to achieve higher classification accuracy while using simpler neural architectures. Compared to the state-of-the-art method RCL (Xu & Zhu, 2018), we improve the model accuracy relatively by 0.21%, 0.21% and 6.70% on the three benchmark datasets and reduce network complexity by 29.9%, 19.0% and 51.0%, respectively.

Each task in MNIST Permutations or MNIST Rotations contains 55,000 training samples, 5,000 validation samples. and 10,000 test samples. Each task in CIFAR-100 contains 5,000 samples for training and 1,000 for testing. We randomly select 1,000 samples from each task training samples as the validation samples and assure each class in a task has at least 100 validation samples. The model observes the tasks one by one and does not see any data from previous tasks.

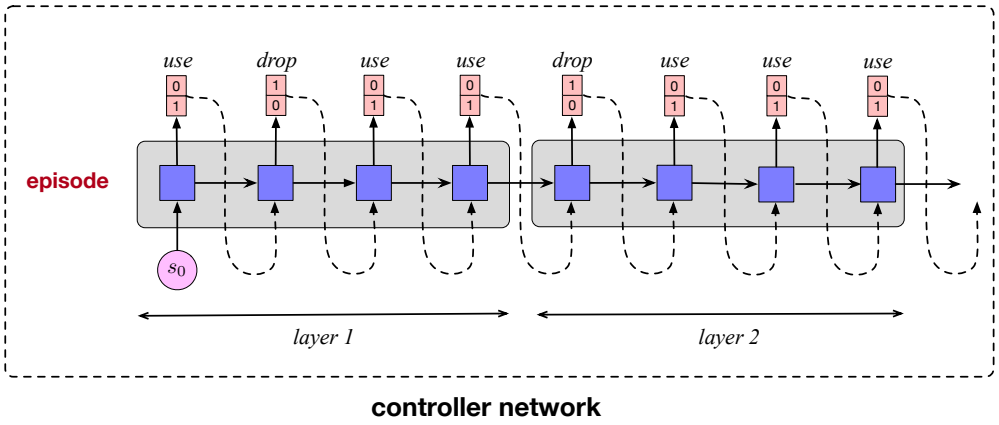
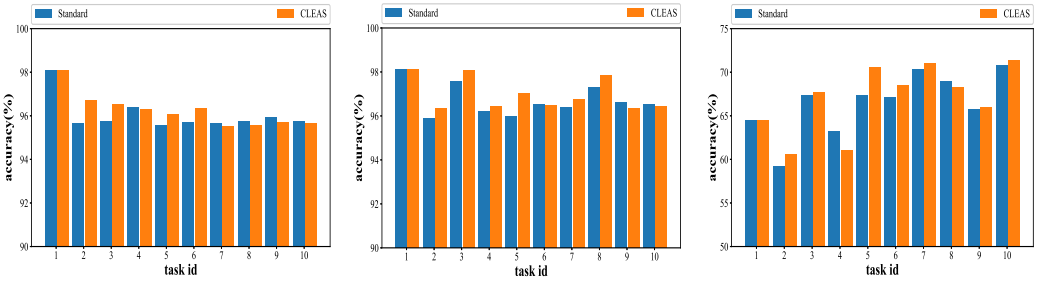


Figure 8: The standard implementation of NAS controller.



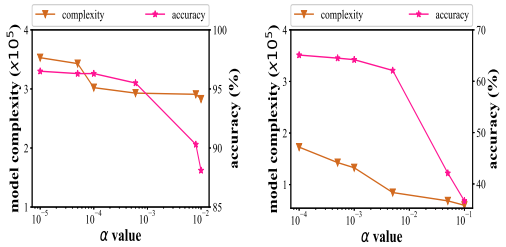
(a) MNIST permutations. (b) Rotated MNIST. (c) CIFAR-100.

Figure 9: Task accuracy of standard NAS controller vs. CLEAS NAS controller.

CLEAS vs. Standard NAS Controller Here we experimentally compare CLEAS to the standard implementation of a NAS controller that considers each output of an RNN-based network as a individual action and it starts with only one initialized state s_0, as shown in Fig 8. Therefore, such a controller considers $(s_0, a_0, 0, s_1, a_1, 0, \ldots, s_{n-1}, a_{n-1}, R)$ as one episode where $s_j = a_{j-1}$, and the real reward R is given only after all states and actions are played. However, the controller of CLEAS considers a sequence of candidate task networks as one episode, and each candidate receives a reward immediately. That is, CLEAS considers $(s^1_{1:n}, \bar{a}^1_{1:n}, R^1, s^2_{1:n}, \bar{a}^2_{1:n}, R^2, \ldots, s^{\mathcal{U}}_{1:n}, \bar{a}^{\mathcal{U}}_{1:n}, R^{\mathcal{U}}, s^{\mathcal{U}+1}_{1:n})$ as one episode (recall this from Section 3.1).

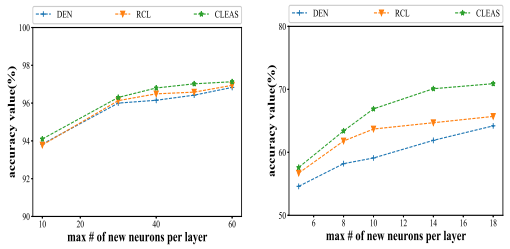
We evaluate these two versions on the same three datasets that were used in Section 4. Fig 9 shows each task accuracy of the three datasets. We find that the controller implemented in the standard way achieves inferior model performances, which are 96.0%, 96.7%, 66.4% in average accuracy on the three datasets respectively. By comparison, CLEAS achieves 96.3%, 97.0%, 66.9%, thus yielding **0.31%**, **0.29%**, and **0.75%** relative improvement. Besides, we also implement a random search version for the task network search by using the same number of models as our CLEAS. Such

simple version archives 95.1%, 96.3% and 62.7% on three datasets, which is actually worse than our CLEAS and Standard NAS Controller. Since the simple version still has a probability to extend the task network by adding new neurons, whereby it can keep active in training for new coming task. However, the random search from old neurons might destroy the inherent information for old tasks.



(a) MNIST permutations. (b) CIFAR-100.

Figure 10: Hyperparameter sensitivity.



(a) MNIST permulation. (b) CIFAR-100

Figure 11: Maximum number of new neurons.

Hyperparameter Sensitivity Lastly, Fig. 10 shows the sensitivity of hyperparameter α in (4). We can see the clear trade-off between model performance and complexity. The best choice of α for MNIST is between $[10^{-4}, 10^{-3}]$ where the network is simpler but preserves good performance as well. For CIFAR-100 α should be between $[10^{-3}, 10^{-2}]$. In Fig. 11 we vary another hyperparameter that is the maximum number of new neurons that can be allocated per layer to a new task. As expected, as the maximum number increases the overall model performances raises as well. But we see that CLEAS always achieves the highest accuracy under different settings.