

ACT Algorithm Based on Long-Short Distance Attention Mechanism

Haoling Yan

School of Software, Northeastern University, Shenyang, China

**Corresponding Author. Email: 3120067268@qq.com*

Abstract: Robot manipulation technology has always been a core area of research, and imitation learning algorithms are crucial for improving task accuracy and efficiency. The ACT algorithm proposed by Zhao et al. performs excellently in fine-grained tasks; however, when handling complex tasks with long time spans, its memory capability and ability to capture long-range temporal dependencies need improvement. Additionally, the increased task complexity leads to a surge in computational resource demands. This paper proposes an improved algorithm based on the ACT algorithm, incorporating a long-short distance attention mechanism. The aim is to enhance memory efficiency for long-duration tasks and improve the ability to capture long-range temporal dependencies, optimize the understanding and utilization of long-term dependencies, improve the overall performance of complex tasks, and reduce the consumption of computational resources such as CPU and GPU.

Keywords: Long-Short Distance Attention, Action Chunking, Imitation Learning, Transformer

1. Introduction

In robot manipulation, fine-grained tasks have always posed a significant challenge. These tasks involve robots performing complex actions under high precision and detailed requirements, such as multi-object interaction, flexible object grasping, and placement. A common difficulty is the issue of “compound errors” [2]. In traditional imitation learning methods, over time, the prediction errors at each time step gradually accumulate, causing the robot to deviate from its training data distribution. This accumulated error is especially critical in fine-grained tasks. Even small errors can trigger a chain reaction, leading to task failure. Therefore, effectively addressing the compound error issue and time-related confounding factors [3] has become one of the core challenges in current robotics research.

To address this issue, Zhao et al. (2020) proposed an innovative algorithm called ACT (Action Chunking with Transformers) in their work “Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware” [1]. ACT solves the compound error problem and the unpredictability of external states by introducing action chunking [4] and the Transformer architecture [5]. The core idea of the ACT algorithm is to reduce the impact of prediction errors at each step on the final task execution by combining action predictions from multiple time steps. Specifically, the ACT algorithm uses a weighted averaging method to merge action predictions from different time steps, ensuring that even if the action selection at a particular time step is incorrect, the correct prediction from the previous step can still adjust the current action, significantly improving task robustness. Moreover,

the powerful capabilities of the Transformer network structure allow ACT to handle multimodal inputs, including vision, joint states, and even language instructions, thereby enhancing the model's adaptability and scalability [1].

The effectiveness and robustness of the ACT algorithm have been widely verified, attracting more researchers to optimize and improve it. Thanpimon Buamanee et al. combined bilateral control with ACT to enhance the robot's perception of position and force information, enabling the robot to perform more delicate and responsive operations, thus improving its performance in complex manipulation tasks [6]. Kelin Li et al. incorporated fingertip force data into the ACT framework, optimizing the robot's performance in fine-grained tasks, allowing it to handle more compliant and adaptable tasks, particularly in those involving deformable objects [7]. Kevin Rohling introduced text embeddings into ACT, enabling the robot to understand and perform multi-task operations based on natural language instructions, significantly enhancing the robot's language comprehension ability and operational flexibility [8]. These improvements demonstrate ACT's strong adaptability and scalability across different domains, further solidifying its effectiveness and robustness in fine-grained manipulation tasks.

These studies optimize the ACT algorithm by embedding more information; however, the improvement proposed in this paper focuses on enhancing the memory capacity of the ACT algorithm itself and modeling long-range temporal dependencies. Although the ACT algorithm has made significant progress in solving compound errors, there is still room for improvement in memory capability and long-range dependency modeling, especially in tasks with long time spans. Currently, ACT relies on the traditional Transformer structure, which, although capable of handling time-series data, may encounter issues with long-range dependencies [9] when dealing with tasks spanning long durations. Specifically, the self-attention mechanism in Transformers, while effective at capturing short-term relationships, is not as effective for modeling dependencies across long time steps, which may lead to suboptimal performance in long-duration tasks.

Therefore, this paper proposes an improved scheme for the ACT algorithm based on a long-short distance attention mechanism. By introducing this mechanism, the model's memory capacity for long-duration tasks can be enhanced, and the relationships between remote time steps [1] can be better captured. Specifically, a new self-attention strategy is incorporated within the ACT framework, enabling the model to maintain sensitivity to recent time steps while effectively capturing correlations between distant time steps, thereby improving the stability and accuracy of the model in multi-step tasks.

The main contribution of this study is the proposal and verification of the application of the long-short distance attention mechanism in the ACT algorithm. It is demonstrated that this mechanism can significantly improve the model's performance in complex task operations while reducing computational costs. Experiments in a simulated environment show that the improved ACT algorithm outperforms the traditional ACT in multi-task operations, particularly when the robot performs long-duration tasks. The enhanced algorithm exhibits notable improvements in accuracy, stability, and robustness.

2. Related work

2.1. Action chunking using transformer

With the increasing demand for efficient imitation learning methods in the robotics field, the Action Chunking Transformer (ACT) algorithm represents a significant advancement, particularly in addressing compounding errors and out-of-distribution state issues in traditional behavior cloning algorithms. ACT integrates a Conditional Variational Autoencoder (CVAE) [10] with the Transformer architecture, significantly improving task execution efficiency in robotics by predicting

multi-step action sequences (i.e., action chunks) from multimodal inputs. The core of this approach lies in modeling scene variability using CVAE, enabling robots to learn from limited human demonstrations. By employing a weighted averaging mechanism for temporal aggregation of predicted actions, the algorithm effectively mitigates failure risks caused by error accumulation. Even if the model makes erroneous action selections at certain time steps, correct predictions from previous steps exert a regulatory effect on subsequent actions, thereby alleviating the compounding error problem. Moreover, the Transformer's self-attention mechanism [5] effectively captures global dependencies within input sequences, helping robots comprehend long-term dependencies and dynamic changes in tasks, thus enhancing their ability to execute complex tasks with precision.

Despite the superior performance of the self-attention mechanism in capturing global dependencies, ACT encounters limitations when handling long-horizon tasks. As the input sequence length increases, computational complexity and memory consumption rise significantly, and dependencies between distant time steps may be overshadowed by local dependencies, reducing the model's capability to capture long-term dependencies. To address this issue, this paper proposes an improved approach based on long-short distance attention mechanisms. By introducing a multi-level self-attention strategy, the method simultaneously focuses on short-term details while effectively capturing long-term dependencies, overcoming the limitations of conventional self-attention mechanisms. This approach enhances computational efficiency while improving the stability and robustness of the ACT algorithm in long-sequence tasks, particularly in executing multi-step tasks in complex environments, thereby significantly increasing task success rates and accuracy.

2.2. Transformer based on long-short distance attention mechanism

The traditional Transformer model has been widely adopted due to its outstanding performance in modeling long-range dependencies through global attention mechanisms. However, studies have shown that the feedforward network (FFN) layers in standard Transformers consume a disproportionately high share of computational resources. This imbalance arises from the classical bottleneck structure, which compresses feature dimensions in the attention layers to reduce computational costs but inadvertently weakens critical contextual modeling capabilities [11]. Against this backdrop, the Transformer architecture has made a breakthrough by introducing the Long-Short Range Attention (LSRA) framework. LSRA employs a dual-path feature encoding mechanism: the global path retains the standard multi-head attention structure but reduces computational density through a dynamic channel-wise partitioning strategy; the local path integrates depthwise separable convolution networks [12], leveraging a sliding window mechanism [13] to capture local syntactic structures within sequences. By implementing hard feature channel partitioning and dynamically merging information in the subsequent FFN layers, LSRA optimizes the balance between global and local modeling capabilities under computational constraints, effectively redefining the Transformer's computational paradigm through structural innovation.

Beyond reducing computational costs, LSRA enhances memory retention in long-horizon tasks by optimizing computational allocation in the attention layers. Compared to the traditional Transformer, which struggles with long-sequence dependency modeling, LSRA strikes a well-calibrated balance between capturing global and local contextual information. This enables the model to effectively track dependencies across distant time steps, thereby improving stability and accuracy in multi-step tasks. This improvement is particularly beneficial for tasks requiring long-term memory, ensuring high accuracy and robustness throughout extended execution periods.

Building on LSRA's design principles, this study integrates the mechanism into the ACT algorithm to enhance its performance in precision operation tasks. By incorporating long-short distance attention into the ACT framework, the model can better capture dependencies across different time steps in long-sequence tasks while maintaining high performance under limited computational

resources. This innovation strengthens ACT's memory retention and task execution stability in long-horizon scenarios. In subsequent experiments, it will be demonstrated how this enhancement offers significant advantages over the conventional ACT algorithm in multi-task operations and prolonged task execution.

2.3. Simulation experiments

This study constructs a dual-arm collaborative operation simulation platform based on the MuJoCo physics engine and the DeepMind Control Suite [14], using the ViperX robotic system with two six-degree-of-freedom robotic arms for precision operation validation. The experimental design includes two contact-constrained tasks: Cube Transfer Task: The right arm's end effector must precisely grasp a red cube from the tabletop and transfer it to the constrained space within the left arm's end effector. Plug-in Assembly Task: The two robotic arms must respectively grasp a blue socket and a red plug and complete an aerial alignment and insertion task without fixed support. Throughout the simulation, the computational resource usage of the ACT algorithm enhanced with the long-short distance attention mechanism is monitored, and the final experimental results are recorded.

3. Experimental methods

3.1. Basic framework of action chunking transformer (ACT)

Action Chunking Transformer (ACT) is a Transformer-based model specifically designed for learning fine dual-arm manipulation skills in robotic systems. This model fully leverages the advantages of the Transformer architecture in capturing long-range dependencies and efficiently processing sequential data, enabling the generation of coherent action sequences to accomplish complex manipulation tasks.

The ACT model consists of several key components: a vision encoder based on a pre-trained ResNet model [15], a Transformer-based action encoder and decoder, and a Variational Autoencoder (VAE). During training, the VAE compresses the real action sequences and joint positions into a style variable z , which is discarded during inference. The Transformer encoder processes visual features, current joint positions, and the style variable z to generate a latent representation, which is then passed to the decoder to produce the corresponding action sequence.

3.2. Transformer enhanced with long-short distance attention mechanism

The traditional Transformer architecture, with its powerful self-attention mechanism, has achieved outstanding performance in various tasks. However, when handling long-duration tasks, the self-attention mechanism faces certain limitations. Specifically, processing long sequences requires capturing dependencies between distant time steps, which can lead to a sharp increase in computational complexity and, in some cases, an inability to fully capture long-range temporal dependencies. To overcome these challenges, this study proposes an improved approach based on the Long-Short Distance Attention (LSRA) mechanism. By dynamically adjusting the weighting between global and local attention, LSRA can simultaneously focus on long-term and short-term dependencies, addressing the limitations of conventional Transformers.

Design of the Long-Short Distance Attention Mechanism

LSRA separates global and local attention to resolve the performance bottlenecks of traditional Transformers in long-duration tasks. Specifically, LSRA employs two dedicated attention branches: one for capturing global dependencies and the other for processing local dependencies. Global dependencies are modeled using the standard self-attention module, while local dependencies are

captured through lightweight convolutional layers. This design enables the model to effectively represent both global and local contexts, improving its performance in handling long-duration tasks.

In LSRA, input features are split along the channel dimension into two branches, allowing the model to process global and local information separately. These features are then dynamically fused through a subsequent feedforward network (FFN) [11]. This approach not only optimizes computational resource allocation but also enhances efficiency and accuracy in processing long-sequence data. By specializing in capturing both global and local contexts, LSRA significantly enhances the model's memory capacity for long-duration tasks, improving stability and robustness in handling long-term dependencies.

3.3. Improved transformer architecture

Based on the LSRA mechanism, this study introduces several improvements to the traditional Transformer: **Global Attention Branch:** Uses the standard multi-head attention (MHA) mechanism [16] to capture global dependencies. This branch models long-range dependencies within the sequence, ensuring that the model perceives global contextual information. **Local Attention Branch:** Incorporates lightweight convolutional layers or windowed self-attention mechanisms to focus on capturing local dependencies within the sequence. By restricting the attention scope to a local window, this branch reduces computational complexity while enhancing the modeling of local details. **Feature Fusion Module:** Combines the output features of the global and local attention branches. Specifically, the outputs of both branches are merged using concatenation or weighted averaging and further processed through a feedforward network (FFN) for feature integration. This fusion mechanism allows the model to leverage both global and local information, generating more expressive feature representations. **Dynamic Weight Allocation:** Introduces learnable weight parameters in the feature fusion process to dynamically adjust the contribution of global and local features. This mechanism enables the model to adaptively balance the use of global and local information based on task requirements.

Compared to traditional Transformer architectures, the LSRA-enhanced Transformer model demonstrates multiple advantages. By explicitly separating global and local attention mechanisms, this design significantly improves the efficiency of capturing long-range dependencies while reducing computational overhead, making it particularly effective for complex tasks involving extended time spans. Additionally, the built-in local attention branch enhances the precision of modeling short-term dependencies within sequences, which is crucial for executing fine-grained manipulation tasks. Notably, the introduction of dynamic weight allocation enables the model to flexibly adjust the fusion ratio of global and local information based on task requirements, granting the model greater adaptability while significantly improving its generalization and robustness.

4. Experimental design

4.1. Simulation environment and tasks

To evaluate the performance of the improved ACT algorithm in fine manipulation tasks, two challenging tasks were designed in the MuJoCo simulation environment: the block transfer task and the dual-arm insertion task. These tasks simulate real-world scenarios where robotic arms perform high-precision operations in confined spaces, with the objective of validating the superiority of the improved algorithm in terms of precise control, dynamic adaptability, and computational efficiency.

In the block transfer task, the objective is to transfer a randomly generated block within a $20\text{ cm} \times 20\text{ cm}$ range from the right robotic arm to the left robotic arm. The task consists of three key steps: the right robotic arm must first make precise contact with the block, then lift it off the ground, and finally transfer it to the gripper of the left robotic arm. The challenge lies in the fact that the gap

between the block and the left-arm gripper is only about 1 cm, meaning that even the slightest positional or postural deviation could result in a collision and cause task failure. Moreover, the task requires the robotic arms to achieve high-precision position and force control in a dynamic environment while quickly adapting to the randomness of the initial position, imposing strict demands on the algorithm's accuracy and robustness.

In the dual-arm insertion task, the goal is to insert a plug into a socket. The initial positions of the plug and socket are randomly generated within a $10\text{ cm} \times 20\text{ cm}$ rectangular area near the left and right robotic arms, respectively. The task consists of three key steps: the left and right robotic arms must first grasp the plug and socket separately, then move them into contact, and finally insert the plug precisely into the socket. This task is particularly challenging because the gap between the plug and socket is only about 5 mm, requiring millimeter-level precision. Additionally, the task demands coordinated motion between both arms and highly accurate position and posture control, where even the slightest error could result in insertion failure. The stringent precision requirements not only test the algorithm's control capability but also validate its performance in dual-arm coordination tasks.

Through the design of these two tasks, the improved ACT algorithm demonstrates its advantages in high-precision, high-dynamic, and high-complexity operational scenarios. The block transfer task verifies the algorithm's ability to perform precise control and maintain robustness under centimeter-level constraints, while the dual-arm insertion task further tests its millimeter-level operational precision and dual-arm coordination capabilities. Moreover, the experiment continuously monitors computational resource usage in real time, ensuring that the algorithm maintains high performance while achieving computational efficiency. The design of these tasks not only provides a comprehensive validation of the improved ACT algorithm but also lays a solid technical foundation for practical robotic arm control applications.

4.2. Experimental results

This study conducts a comparative analysis of the improved ACT, the ACT proposed by Tony Z. Zhao, and four other imitation learning algorithms. Among them, BC-ConvMLP [17] is a fundamental and widely used model that leverages convolutional neural networks to process real-time image data, integrating extracted features with joint position information to generate action predictions. BeT [18], while also based on the Transformer framework, differs significantly in two key aspects: first, it does not require chunking actions but directly outputs a single action based on the historical observation sequence; second, its image data is processed by an independent pre-trained visual encoder, which remains frozen during training, meaning that the perception module and decision-making module are not optimized in an end-to-end manner. RT-1 [19] is another Transformer-based approach that predicts a single action based on a fixed-length historical observation sequence. Notably, both BeT and RT-1 adopt an action-space discretization strategy, meaning their output is a probability distribution over discrete intervals. However, BeT introduces an additional continuous offset within the interval center to improve precision. Additionally, VINN [20], a non-parametric method, relies on demonstration data during the testing phase. Specifically, when receiving new observational data, VINN retrieves the k most similar samples from a database based on visual features and generates actions using a weighted k -nearest neighbors algorithm. The method's visual feature extractor is based on a pre-trained ResNet architecture, further fine-tuned on demonstration data through unsupervised learning. Tony Z. Zhao's ACT method directly outputs continuous actions, a design aimed at meeting the high-precision control requirements of fine manipulation tasks. However, while the self-attention mechanism in Transformer can effectively capture short-term dependencies, it is less effective in modeling long-term dependencies across extended time steps. The ACT model in this study, incorporating Long-Short Distance Attention

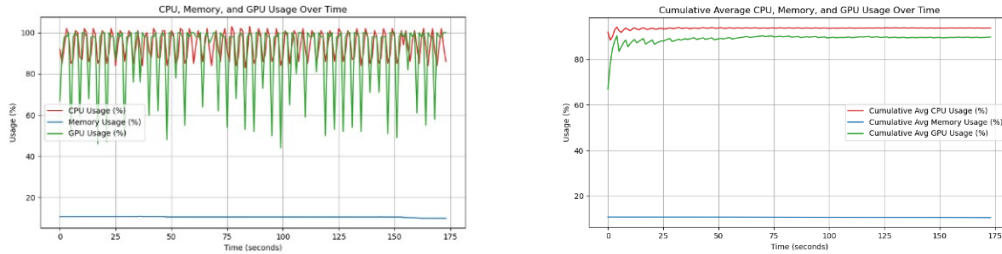
(LSRA), enhances its ability to capture both global and local context, significantly improving its memory capacity for long-horizon tasks while slightly reducing computational costs.

Table 1: Comparison of movement across different algorithms

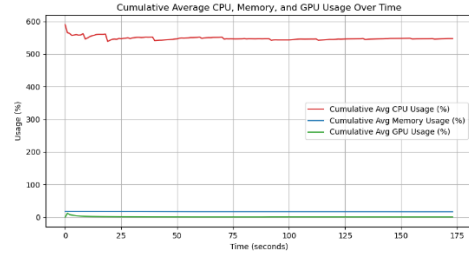
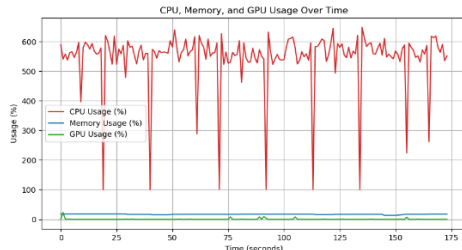
Imitation Learning Algorithm	Cube Transfer
BC-ConvMLP	1%
BeT	27%
RT-1	2%
VINN	3%
ACT	86%
my-ACT	94%

Imitation Learning Algorithm	Cube Transfer
BC-ConvMLP	1%
BeT	3%
RT-1	1%
VINN	1%
ACT	32%
my-ACT	56%

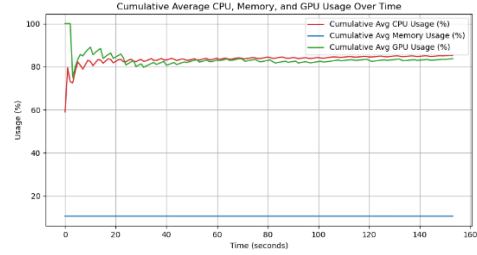
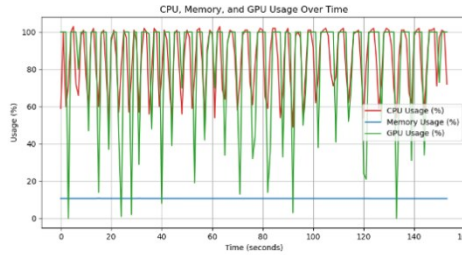
According to the experimental results presented in Table 1, the proposed ACT algorithm, improved with a long- and short-range attention mechanism, demonstrates significant advantages in both the block transfer and dual-hand insertion tasks. In the block transfer task, traditional imitation learning algorithms generally exhibit low success rates, whereas the ACT algorithm proposed by Zhao et al. achieves a notable 86% success rate. However, the improved ACT algorithm in this study further enhances performance by an additional 8 percentage points, reaching a 94% success rate. In the dual-hand insertion task, the success rates of traditional algorithms do not exceed 32%, with Zhao’s ACT algorithm performing the best among them. In contrast, the improved ACT algorithm achieves a 56% success rate, surpassing the best baseline by 24 percentage points. The experimental results indicate that the improved ACT algorithm effectively mitigates the compounding error problem, enhances the modeling capability for non-Markovian behaviors, and improves success rates in complex tasks.



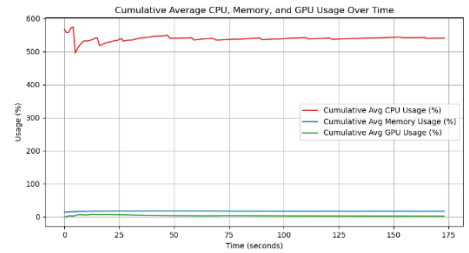
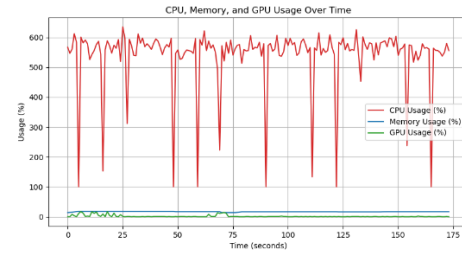
Pre-Improvement ACT – Cube (Train)



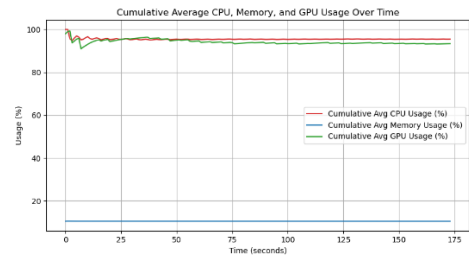
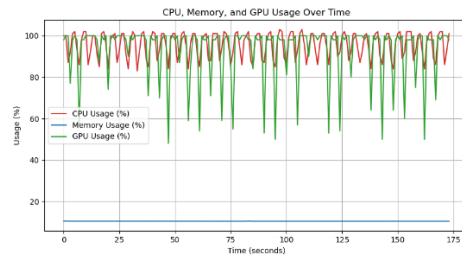
Pre-Improvement ACT – Cube (Test)



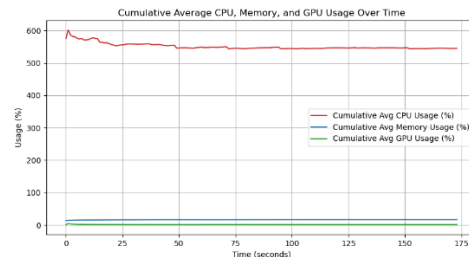
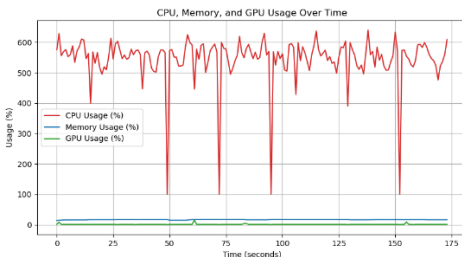
Post-Improvement ACT – Cube (Train)



Post-Improvement ACT – Cube (Test)



Pre-Improvement ACT – Insert (Train)



Pre-Improvement ACT – Insert (Test)

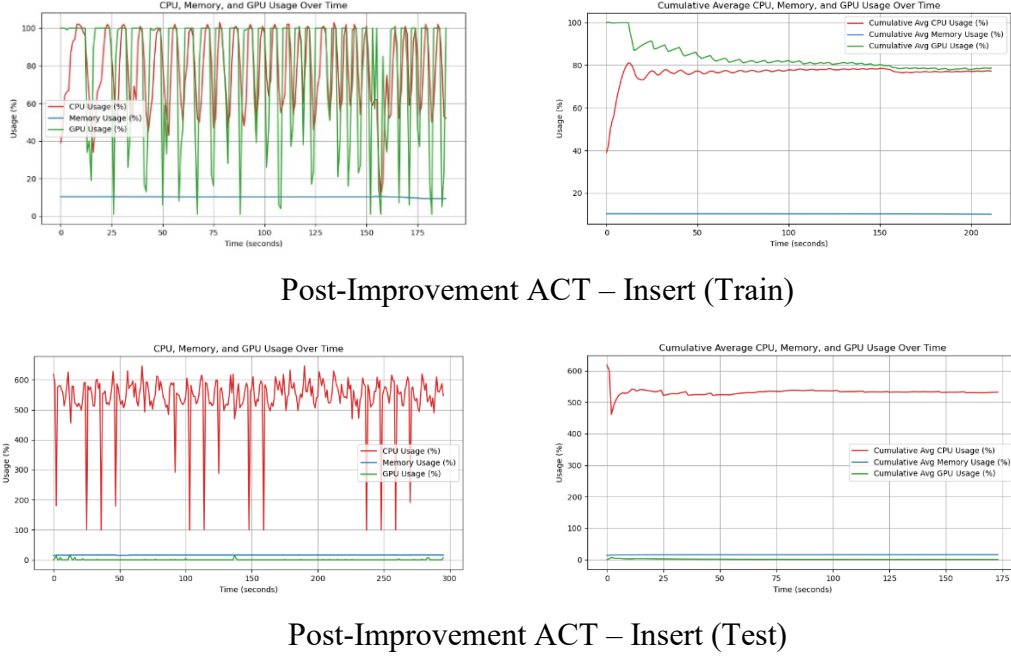


Figure 1: Comparison of computational resource consumption across different simulated tasks

As shown in Figure 1, a detailed comparison of computational resource consumption during the training and testing phases for two simulated tasks (block transfer and dual-hand insertion) is presented. Specifically, in the training phase of the block transfer task, the ACT algorithm proposed by Zhao et al. exhibits an average cumulative CPU and GPU utilization of approximately 95% and 90%, respectively. In contrast, the improved ACT architecture, based on a long- and short-range attention mechanism, maintains significantly lower levels of around 80% for both resources. Similarly, during the training phase of the dual-hand insertion task, Zhao’s ACT algorithm demonstrates a high resource consumption, with CPU and GPU utilization rates both nearing 95%, whereas the improved architecture significantly reduces resource usage, stabilizing at approximately 80% for both. Notably, during the validation phase of both simulated experiments, CPU and GPU usage remain roughly equivalent, showing no significant differences. In summary, these findings clearly indicate that the optimized ACT algorithm effectively reduces computational resource consumption during training, demonstrating higher resource utilization efficiency.

5. Conclusion and future prospects

Building upon the solid foundation of the ACT algorithm proposed by Zhao et al., this study introduces an improved approach incorporating a long- and short-range attention mechanism. This enhancement is specifically designed to strengthen the model’s memory efficiency in handling long-span complex tasks and to improve its ability to capture long-range temporal dependencies accurately. Not only does this provide a novel theoretical perspective on algorithm design, but it also exhibits significant potential for enhancing task execution efficiency and accuracy in practical applications. Experimental validation demonstrates that the improved ACT algorithm achieves breakthroughs in simulated tasks such as block transfer and dual-hand insertion. The notable increase in success rates directly reflects the enhancement in algorithm performance.

A particularly noteworthy finding is that the improved algorithm optimizes computational resource utilization while enhancing task execution efficiency. During the training phase, CPU and GPU usage significantly decrease compared to the original ACT algorithm. This discovery provides strong

support for the efficient operation and resource conservation of the algorithm in real-world deployment. It not only implies that the improved algorithm is better suited for resource-constrained environments but also lays a solid foundation for its broader adoption in resource-sensitive applications.

However, this study represents only an initial exploration in the field of algorithm optimization. With the rapid advancement of robotics, real-world applications impose increasingly higher demands on algorithm performance, adaptability, and robustness. Future research will focus on conducting further in-depth studies on the theoretical foundations and practical applications of the long- and short-range attention mechanism. This includes, but is not limited to, fine-tuning algorithm structures, optimizing hyperparameters, and exploring various improvements to Transformer-based architectures to enhance performance in complex dynamic environments.

To further improve model performance, future work will consider multimodal information fusion. Specifically, beyond the current use of positional data, force feedback information will be incorporated to improve a robot's ability to manipulate irregular, fragile, and deformable objects, thereby increasing precision and safety in complex operational tasks. Additionally, integrating natural language understanding modules will be explored, embedding textual information to enhance the robot's semantic comprehension of operational commands, allowing for more flexible task execution. Moreover, to address sample efficiency issues, reinforcement learning strategies based on single-shot demonstrations will be introduced, enabling rapid learning and adaptation to new tasks with minimal demonstration data, thereby improving the algorithm's practicality and scalability.

In conclusion, this study not only successfully improves Zhao's ACT algorithm but also opens up valuable opportunities for subsequent research and applications. The continued development of this algorithm is expected to contribute to advancements in various fields, further driving progress in robotics technology.

References

- [1] Zhao, T. Z., et al. (2023). *Learning fine-grained bimanual manipulation with low-cost hardware*. *arXiv Preprint arXiv:2304.13705*.
- [2] Ke, L., et al. (2021). *Grasping with chopsticks: Combating covariate shift in model-free imitation learning for fine manipulation*. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6421–6427). IEEE.
- [3] Swamy, G., et al. (2022). *Causal imitation learning under temporally correlated noise*. In *International Conference on Machine Learning (Vol. 162, pp. 9909–9919)*. PMLR.
- [4] Lai, L., Huang, A. Z., & Gershman, S. J. (2022). *Action chunking as policy compression*.
- [5] Vaswani, A. (2017). *Attention is all you need*. In *Advances in Neural Information Processing Systems*.
- [6] Buamane, T., et al. (2024). *Bi-ACT: Bilateral control-based imitation learning via action chunking with transformer*. *arXiv Preprint arXiv:2401.17698*.
- [7] Li, K., et al. (2024). *Haptic-ACT: Bridging human intuition with compliant robotic manipulation via immersive VR*. *arXiv Preprint arXiv:2409.11925*.
- [8] Rohling, K. *Integrating natural language instructions into the action chunking transformer for multi-task robotic manipulation*.
- [9] Bengio, Y., Simard, P., & Frasconi, P. (1994). *Learning long-term dependencies with gradient descent is difficult*. *IEEE Transactions on Neural Networks*, 5(2), 157–166.
- [10] Sohn, K., Lee, H., & Yan, X. (2015). *Learning structured output representation using deep conditional generative models*. In *Advances in Neural Information Processing Systems*.
- [11] Hornik, K., Stinchcombe, M., & White, H. (1989). *Multilayer feedforward networks are universal approximators*. *Neural Networks*, 2(5), 359–366.
- [12] Kaiser, L., Gomez, A. N., & Chollet, F. (2017). *Depthwise separable convolutions for neural machine translation*. *arXiv Preprint arXiv:1706.03059*.
- [13] Liu, Z., et al. (2021). *Swin transformer: Hierarchical vision transformer using shifted windows*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE.
- [14] Todorov, E., Erez, T., & Tassa, Y. (2012). *Mujoco: A physics engine for model-based control*. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE.

- [15] He, K., et al. (2016). *Deep residual learning for image recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.
- [16] Voita, E., et al. (2019). *Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned*. *arXiv Preprint arXiv:1905.09418*.
- [17] Jang, E., et al. (2022). *BC-Z: Zero-shot task generalization with robotic imitation learning*. In *Conference on Robot Learning*. PMLR.
- [18] Shafiullah, N. M., et al. (2022). *Behavior transformers: Cloning π modes with one stone*. *Advances in Neural Information Processing Systems*, 35, 22955–22968.
- [19] Brohan, A., et al. (2022). *RT-1: Robotics transformer for real-world control at scale*. *arXiv Preprint arXiv:2212.06817*.
- [20] Pari, J., et al. (2021). *The surprising effectiveness of representation learning for visual imitation*. *arXiv Preprint arXiv:2112.01511*.