SKILL COMPOSITION

Vaishnava Hari

Dream A marketplace of skills, from which a robot installs skills. Skill is the ability to perform part of a task. This task could be new for the robot and set in an unseen environment. And a framework to decompose a task into subtasks which then is solved using a subset of skills.

1 Overview of approach

1.1 Decomposition of task

First challenge is to decompose a task into smaller sub-tasks, which then can be executed by a subset of skills. Or in other words, given a set of skills, how to select a subset of skills to execute a task during inference. This is a well studied problem in the field of Hierarchical Reinforcement Learning (HRL) [1] .

1.2 Subtask discovery

Next challenge is to identify which skills to learn during training. This is usually done by

- **Bottom-up approach:** Handpicking a set of primitive skills, intuitively, and training them individually. Later, they are frozen and put together for training the high-level policy. This method is widely used. [2].

However, this approach can not be generalized. And also handcrafting the set of skills is not optimal [3].

- **Top-down approach:** Train both high-level and low-level policies together. And number of low-level policies or skills is not fixed. This introduces a new hurdle: non stationary transistion function for the high-level policy.

2 Previous Works

2.1 Hirarchical Reinforcement Learning

Hirarchical Reinforcement Learning (HRL) is a framework for breaking down complex tasks into simpler subtasks. It is accomplished by defining multiple layers of policies, where higher-level policies select subgoals or subtasks for lower-level policies to execute.

Bottom-up approach: Started with a set of low level skills, trained individually and freeze. Introduce a high level policy to select a subset of skills to execute. This method is widely used.

Top-down approach: Train both high level and low level policies together. The main hurdles are:

1. number of low level policies: decomposing a task into a set of skills is still an open problem.
2. credit assignment problem (cap): how to assign credit to the right low level policy.

2.2 Vision-Language-Action

Vision-Language-Action (VLA) models are a class of models based on the transformer architecture and are designed to process and understand both visual and textual information. They are made up of two parts:

1. VLM (Vision-Language Model): This part of the model is responsible for understanding and processing visual and textual information.

2. Action Head: Made of diffusion transformers and are responsible for generating actions using the embeddings from the VLM. They can also take in additional inputs such as proprioceptive and previleged environment information.

Some of the popular VLA models include: - SoFar [4], state of the art in robot manipulation tasks in the Google SimplerEnv dataset [5].

- OpenVLA [6], a 7B parameter model. It uses Lama 2 as the backbone and is trained on 970k robot demos in Open X-Embodiment dataset [7].

- SAM2ACT [8] is a new model focused on manipulation tasks. It is the leader in the RLbench dataset [9].

References

[1] M. Hutsebaut-Buysse, K. Mets, and S. Latré, "Hierarchical Reinforcement Learning: A Survey and Open Research Challenges," *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 172–221, Mar. 2022.

[2] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek, "Hierarchical Reinforcement Learning: A Comprehensive Survey," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 109:1–109:35, Jun. 2021.

[3] D. Silver and R. S. Sutton, "Welcome to the Era of Experience."

[4] Z. Qi, W. Zhang, Y. Ding, R. Dong, X. Yu, J. Li, L. Xu, B. Li, X. He, G. Fan, J. Zhang, J. He, J. Gu, X. Jin, K. Ma, Z. Zhang, H. Wang, and L. Yi, "SoFar: Language-Grounded Orientation Bridges Spatial Reasoning and Object Manipulation," Feb. 2025.

[5] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "Evaluating Real-World Robot Manipulation Policies in Simulation," May 2024.

[6] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An Open-Source Vision-Language-Action Model," Sep. 2024.

[7] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Mart'in-Mart'in, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H.

Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," Jun. 2024.

[8] H. Fang, M. Grotz, W. Pumacay, Y. R. Wang, D. Fox, R. Krishna, and J. Duan, "SAM2Act: Integrating Visual Foundation Model with A Memory Architecture for Robotic Manipulation," Feb. 2025.

[9] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The Robot Learning Benchmark & Learning Environment," Sep. 2019.