Draft Proposal

FAST Tokenization for Time-Discrete Control: Learning with Compressed Action Representations

Ruiming Wu
Technical University of Munich
ruiming.wu@tum.de

Baiheng Li
Technical University of Munich
go76ren@mytum.de

I. Objective

In this project, we want to explore how different action tokenization methods affect the learning of control policies. We plan to train a small Transformer model in simple simulation environments like CartPole and Pendulum, using two different tokenization approaches: the standard Naïve Binning method and the newer FAST Tokenization technique.

Our goal is to compare them in terms of training speed, control performance, and prediction accuracy, and to better understand whether compression-based tokenization brings clear advantages in these settings.

II. Related Work

Many current Vision-Language-Action (VLA) models like π_0 [1] and OpenVLA [2] use simple binning methods to turn continuous robot actions into discrete tokens. This works well for low-frequency tasks, but once the control becomes fast or more fine-grained, these models tend to struggle—mainly because the token sequences get too long and redundant.

To deal with this, FAST (Frequency-space Action Sequence Tokenization) was recently proposed. It compresses action sequences using Discrete Cosine Transform (DCT) and Byte-Pair Encoding (BPE), which helps reduce sequence length and improve training speed. Research showed that combining FAST with models like π_0 or OpenVLA makes them train faster and perform better on more challenging, high-frequency tasks [3]. That's why we're interested in testing this idea in our own setup and comparing it to the more traditional binning approach.

III. Technical Pipeline (estimated)

We will primarily use Python to work on the project, and OpenAI Gym and Pytorch will be the main repositories:

- Model a simple physics-based simulation environment (e.g. Pendulum or CartPole) with continuous action space
- Collect action trajectories (e.g. 1-second sequences) as training data using a PID controller
- Apply two different tokenization methods:
  a) Naïve binning (per-dimension, per-timestep)
  b) FAST tokenization (DCT → quantization → BPE)
- Train two small Transformer-based autoregressive models, each using one of the tokenization methods to predict action sequences from observations
- Compare both models in terms of:
  a) Control performance (accuracy, stability)
  b) Training speed and convergence behavior
  c) Token sequence length and compression efficiency
  d) Generalization under slight variations in dynamics (e.g., mass, damping)

IV. References

[1] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi 0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

[2] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

[3] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. arXiv preprint arXiv:2501.09747, 2025.