Poly-Autoregressive Prediction for Modeling Interactions

Large language models (LLMs) have been very successful with natural language processing (NLP) tasks, which require accurate reasoning over relationships words have within a body of text. A key component of LLMs is autoregressive (AR) modeling, where each word token is predicted based on a sequence of preceding word tokens. Building on the success of AR modeling in the NLP community, this work focuses on modeling the dynamic relationships between agents and entities in everyday interactions that occur in the physical world, such as social interactions, driving, and hand-object interactions.

Unlike language which is structured through grammar and semantics, interactions in the physical world are dictated by both the laws of physics (e.g. how a hand grasps an object) and the internal state of each agent (e.g. the trajectory an agent chooses to move a grasped object in) , a latent variable that we know nothing about. Furthermore, as opposed to text where each word follows another unidimensionally, the states of multiple agents are changing simultaneously. For example, in social situations, the history of a single person's past states does not alone determine the dynamics of their future states; we also need to consider the states of other agents. We argue, therefore, that AR modeling alone is insufficient.

In this paper, we introduce poly-autoregressive (PAR) modeling, a simple unifying approach to model the influence of other agents and entities on one's behavior. We model behavior as a temporal sequence of states and predict an ego agent's future behavior conditioned on the history of behavior of the ego agent as well as the rest of the agents. By considering other agents' behaviors, we demonstrate that our approach significantly improves upon the ill-posed problem of single-agent prediction in interactive settings.

The PAR framework uses a transformer-based model for next-token prediction. Transformers have shown great success in language modeling and naturally lend themselves to predicting behavior over time. In an interaction scenario of N agents, our model predicts the future behavior of an ego (Nth) agent conditioned on its past behavior and the behavior of the other $N - 1$ non-ego agents. Each prediction task may model a different behavior modality of interest, e.g. actions for social action prediction, or 6DoF pose for object forecasting in hand-object interactions. We apply PAR to three case studies of common real-world interactions:

1. *Social action prediction.* We test our method on the AVA benchmark [1] for action forecasting. By incorporating both the ego and another agent using PAR, we get a overall +1.9 absolute mAP gain over AR, which only models the ego agent to predict its future behavior, and an absolute +3.5 mAP gain on 2-person interaction classes.
2. *Trajectory prediction for autonomous vehicles.* When forecasting future xy locations of an ego vehicle, incorporating the locations of neighboring vehicles with PAR outperforms AR, which only uses the ego vehicle's preceding trajectory as input. Specifically, on the nuScenes dataset [2], PAR outperforms AR with a relative improvement of 6.3% ADE and 6.4% FDE.
3. *6DoF object pose forecasting during hand-object interaction.* We use the DexYCB dataset [3], where we treat the object as the ego agent and the hand as the interacting agent. While PAR integrates that hand's 3D location and object pose history, AR only uses the object's pose history to predict the object's 6DoF pose. PAR outperforms AR with relative improvements of 8.9% and 41.0% for the rotation and translation predictions, respectively.

In all of these different settings, we find that incorporating the behavior of other agents in the scene improves predictions of the ego agent's behavior. All of these problems are modeled via the same simple PAR framework and implemented using the same proof-of-concept 4 million parameter transformer without any modifications to the base framework or architecture, only to data pre-processing and choice of tokenization. We also provide an example of a simple way to build on our architecture through a location positional encoding.

The primary contribution of this work is a versatile framework that can be applied to a diverse range of settings, without modifications aside from domain-specific data processing. Our results suggest that PAR provides a simple formulation that, with a more complex transformer backbone and larger datasets, could enhance prediction of diverse multiagent interactions across various problem domains. To facilitate further exploration and development, we have released our code, which contains the building blocks to use PAR for modeling other types of multi-agent interactions.

Citations

[1] Chunhui Gu et al., AVA: A video dataset of spatio-temporally localized atomic visual actions, CVPR 2018.
[2] Holger Caesar et al., nuScenes: A multimodal dataset for autonomous driving, CVPR 2020.
[3] Yu-Wei Chao, et al. Dexycb: A benchmark for capturing hand grasping of objects. CVPR, 2021.