Thesis Storyline

Giulio Starace

2023-08-03 16:35:31

Out-of-distribution (OOD) generalization, the ability to perform well on data not distributed identically to the training data, is a key challenge in the field of Machine Learning (ML) [1]. Recent advances in the field have led to increasingly urgent discussions on the associated safety implications [2–5]. In this context, Goal Misgeneralization (GMG) [6, 7] is a form of OOD generalization failure often introduced when discussing potential safety-critical failure modes of ML models.

GMG occurs when a model trained to pursue a given goal during training, misgeneralizes during testing and capably pursues a different goal instead. GMG is of particular concern due to the retained capability of the model when generalizing, allowing for the possibility of visiting arbitrarily undesirable states.

We frame GMG as a consequence of causal confusion [8], where the underlying causal model for achieving our goal is spuriously correlated with some other variable during training time, which the model confusingly learns instead. We identify three possible solutions to the problem:

1. Performing interventions on the data, to better discover the underlying causal model.
2. Increasing the variety of the training data, to reduce the chance of spurious correlations.
3. Improve task specification, to reduce ambiguity and similarly reduce the chance of spurious correlations.

While we expect a combination of these solutions to be most effective, we choose to focus on this solution: improving task specification. We focus on this direction because we intuit that most examples of goal misgeneralization are due to the limited nature by which tasks are specified. It is perhaps unfair to blame the model for its causal confusion when the bandwidth for specifying the task is limited to e.g. a one-dimensional signal such as reward in Reinforcement Learning [9].

We focus on the problem area of of sequential decision making (SDM) and choose to leverage recent advancements in natural language processing, leading us to language-informed sequential decision making (LISDM). We choose this direction because we view task specification as a *communication* between task requester

and task executor. For communicative intents, natural language provides the most expressive potential.

Here, we require language representations to be "understood" by our policy, what is also referred to as *grounding*. We make use of a vision-language multimodal foundation model, CLIP, as our starting point and study a potential modification route for it to be used as a hypothetical foundation model for SDM.

Ultimately, we present the following contributions:

1. We connect GMG to the phenomenon of Causal Confusion
2. We provide a high level description of the possible solutions to GMG
3. We show how an existing vision-language foundation model can be modified to simulate a hypothetical foundation model for SDM.
4. We show how such a hypothetical foundation model could be leveraged for improved task specification for the case of addressing GMG.

References

[1] Arjovsky, M. (2020). *Out of Distribution Generalization in Machine Learning*. PhD thesis, New York University, USA.

[2] Hendrycks, D., Carlini, N., Schulman, J. and Steinhardt, J. (2022). Unsolved Problems in ML Safety.

[3] Ngo, R., Chan, L. and Mindermann, S. (2022). The alignment problem from a deep learning perspective.

[4] Hendrycks, D., Mazeika, M. and Woodside, T. (2023). An Overview of Catastrophic AI Risks.

[5] Critch, A. and Russell, S. (2023). TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI.

[6] Langosco, L. L. D., Koch, J., Sharkey, L. D., Pfau, J. and Krueger, D. (2022). Goal Misgeneralization in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning* pp 12004–19. PMLR.

[7] Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J. and Kenton, Z. (2022). Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals.

[8] de Haan, P., Jayaraman, D. and Levine, S. (2019). Causal Confusion in Imitation Learning. In *Advances in Neural Information Processing Systems* vol 32. Curran Associates, Inc.

[9] Vamplew, P., Smith, B. J., Källström, J., Ramos, G., Rădulescu, R., Roijers, D. M., Hayes, C. F., Heintz, F., Mannion, P., Libin, P. J. K., Dazeley, R. and Foale, C. (2022). Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems* **36** 41.