Monocular Depth Estimation Pre-training for Imitation-based Autonomous Driving

Shubham Juneja
Institute of Data Science & Digital Technologies
Vilnius University
Vilnius, Lithuania
Email: shubham.juneja@mif.stud.vu.lt

Virginijus Marcinkevičius
Institute of Data Science & Digital Technologies
Vilnius University
Vilnius, Lithuania
Email: virginijus.marcinkevicius@mif.vu.lt

Povilas Daniušis
Research Institute of Natural and Technological Sciences
Vytautas Magnus University
53361 Kaunas, Lithuania
Email: povilas.daniusis@vdu.lt

***Abstract*—Artificial intelligence based systems have taken industries and research by storm, one of such systems are employed in autonomous driving. Recent empirical findings in imitation learning for autonomous driving indicate that pre-training on various tasks can enhance the effectiveness of the learner method (e.g., neural network). We propose pre-training neural networks over the task of monocular depth estimation could be beneficial in terms of estimating another modality and extending the scene understanding capabilities of the learner method. We also outline a plan for further investigation of this approach, aiming to integrate new experimental results with existing findings in this line of research, i.e., pre-training for autonomous driving.**

***Keywords*—imitation learning; autonomous driving; monocular depth estimation; pre-training**

I. Introduction

Autonomous driving systems are one of many Artificial Intelligence (AI) based systems that have been transformative for multiple industries and research. These systems follow one of two paradigms, namely, the modular paradigm or the end-to-end paradigm. The modular paradigm forms a pipeline of modules where each module takes responsibility of a task, while the end-to-end paradigm often relies on imitation learning based learners that learn to imitate the whole task of driving. The end-to-end systems learn from data consisting of demonstrations from an expert and require very low engineering efforts as opposed to systems following the modular paradigm. Hence, making the use of imitation learning a promising area of research.

Imitation based methods suffer on encountering the problem of co-variate shift, where a trained driving agent faces scenarios during the time of testing, that were not presented during training. This leads to weak generalisation and unexpected driving behaviour. Generalisation ability, being of critical importance, has led research to explore various directions. Recently popular lines of work in the area have explored varying data generation methods [1], architectures [2][3], smarter data aggregation methods [4], incorporating additional modalities [5] and more to alleviate this issue. Our work delves into the line of work in imitation based autonomous driving that explores pre-training of learning methods [6][7].

To train a new learner for a particular system from scratch, can require excessive amounts of data, resources and time. Therefore, performing pre-training has become a standard approach in order to fast forward the process of training, majorly in applications of natural language processing, object detection, object recognition, etc. Autonomous driving systems are increasingly becoming complicated to train with the aim of achieving better ability to generalise, making some kind of pre-training a must. Hence, most imitation based methods have a default reliance on ImageNet pre-trained vision encoders [1][2], rather than training the models all the way from random weight initialisation. Meanwhile, some recent works solely drop this reliance and explore other forms of pre-training [6][7].

ImageNet pre-training tends to narrow down the concept of image understanding to a single concept, i.e., classification. Therefore, works exploring pre-training methods propose training on alternate tasks that bring in an additional perspective, like visual place recognition [7] or contrastive representation learning [6], in order to improve generalisation. One such task that features a high potential of scene understanding while estimating another modality from a RGB image is monocular depth estimation [8]. A very recent method for estimating depth, Depth Anything [9] proposes a foundation model formation by training upon a massive dataset of 62 million images, and shows a strong ability of zero-shot generalisation for estimating depth.

Considering the limitations of current research in pre-training of learning methods in order to improve generalisation on unseen driving scenarios, we make the contribution of proposing another kind of pre-training. We propose pre-training an agent on the task of monocular depth estimation using the depth anything method, followed by training the agent on the task of driving. We hypothesise that pre-training on the task of depth estimation on a large scale dataset may embed the ability to estimate distances between important objects in the visible environment and therefore improve scene understanding. And hence, we also propose evaluation of the proposed method on the offline Leaderboard [1] benchmark standard against a baseline method and recent methods.

The remainder of this paper is organised as follows: Section II reviews the essential literature related to autonomous driving and monocular depth estimation. Section III describes our proposed approach, detailing both the implementation and the evaluation plan. Finally, Section IV concludes the paper.

II. Related Work

The classical idea of imitation learning-based autonomous driving consists of data collection from demonstrations fol-

lowed by training a neural network to predict the actions given vision inputs from the demonstrations [10]. Further development attempts revolve around improving architectures [2], improving data quality using a reinforcement learning agent [1], increasing the perception ability [3], and so on. Some of the few works that investigate the advantages of pre-training the vision encoders of a driving agent, use an alternate task in the pre-training phase. Action conditioned pre-training method [6] trains over contrastive representation learning on understanding how the visual representations differ as per actions used. Another recent method pre-trains over the task of visual place recognition in order to increase scene understanding in context of changing lighting and weather conditions [7]. Despite its popularity, pre-training still tends to be under explored in the area of autonomous driving.

Depth is often used as an additional or sometimes as the only modality for the task of driving, often requiring expensive depth sensors. Monocular depth estimation is a task which aims to predict the depth modality given an RGB image. Recent method depth anything [9] shows ability to estimate the depth in images with zero-shot learning. It accomplishes this by training over a massive dataset that combines labelled and unlabelled images under a feature alignment loss, and hence, shows possession of rich feature understanding. Although several methods exist for depth estimation, the depth anything method stands out since it forms a foundational model for its task of interest, demonstrating a wider generalisation ability. This shows promising capability to explore its use to transfer learning. In our proposed idea, we consider building on top of such capabilities by leveraging in the form of pre-training of learning methods.

III. Method and Experiments

A. Approach

Our approach is to pre-train a visual encoder over the task of monocular depth estimation following the depth anything method [9]. Further on, to embed the vision encoder into the architecture based on conditional imitation learning [2]. Then we plan to train the whole architecture over the task of autonomous driving, using the data collected with the reinforcement learning agent commonly known as Roach [1]. We plan to base the framework on our previous work [7] for better comparability, as this work follows a similar line of research.

(a)

(b)

Fig. 1. (a) An RGB image from CARLA simulator. (b) Same RGB image's estimated depth with depth anything [9].

B. Environment and Benchmarks

For the training and testing environment, we select the CARLA simulator [11] as it enables testing under varying weather conditions and towns settings. We then plan to assess the performance based on the offline Leaderboard [1] benchmark, which establishes standardised train and test route settings. The choice of using a simulated environment may present challenges in transferring the trained method to real-world environment application, to address them domain-transfer techniques can be further investigated. For the purposes of this study, we plan our steps around the use of the simulated environments, as it is sufficient for validating the proposed concept and aligns with the scope of our research. To concretely measure the performance, we plan to work with the metrics of route completion and distance completion percentages. Additionally, calculated depths can be compared to the ground truth depths, before and after training of the encoder on the task of driving. This may require additional exploration in modification of decoders.

C. Implementation plan

For the initial pre-training we plan to utilise the pre-trained encoder from the transformer network provided by the depth anything work. We show the capability this pre-trained encoder in estimating depth of a RGB image from simulation environment in Figure 1. Later to train on the task of driving, we collect images from the front camera of the car together with the command from the expert agent and a higher level command as done in other works [1][2][7]. To improve comparability, we plan to inherit the implementation specifics from our recent work [7] that aligns with this line of work.

IV. Conclusion

Our work proposes further exploration of pre-training in the area of imitation-based autonomous driving. We hypothesise that the idea of pre-training on the task of monocular depth estimation followed by training on how to drive holds potential in bringing in better scene understanding and additionally in estimating the depth modality while driving, therefore, potentially resulting in better decision making in unseen scenarios and improving overall generalisation. In our further work, we plan to follow the implementation plan and generate empirical results.

References

[1] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 222–15 232.

[2] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9329–9338.

[3] Y. Xiao, F. Codevilla, D. Porres, and A. M. López, "Scaling vision-based end-to-end autonomous driving with multi-view attention learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 1586–1593.

[4] A. Prakash, A. Behl, E. Ohn-Bar, K. Chitta, and A. Geiger, "Exploring data aggregation in policy learning for vision-based urban autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 763–11 773.

[5] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 878–12 895, 2022.

[6] Q. Zhang, Z. Peng, and B. Zhou, "Learning to drive by watching youtube videos: Action-conditioned contrastive policy pretraining," in *European Conference on Computer Vision*. Springer, 2022, pp. 111–128.

[7] S. Juneja, P. Daniušis, and V. Marcinkevičius, "Visual place recognition pre-training for end-to-end trained autonomous driving agent," *IEEE Access*, vol. 11, pp. 128 421–128 428, 2023.

[8] R. Birkl, D. Wofk, and M. Müller, "Midas v3.1 – a model zoo for robust monocular relative depth estimation," *arXiv preprint arXiv:2307.14460*, 2023.

[9] L. Yang *et al.*, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.

[10] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Proceedings of the 1st International Conference on Neural Information Processing Systems*, ser. NIPS'88. Cambridge, MA, USA: MIT Press, 1988, p. 305–313.

[11] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.