Vilnius
University

Idea Paper: Monocular Depth Estimation Pre-training for Imitation-based Autonomous Driving

Authors:
Shubham Juneja (Presenter)
Virginijus Marcinkevičius
Povilas Daniušis

Presenter's Details:
Institute of Data Science & Digital Technologies, Vilnius University
shubham.juneja@mif.stud.vu.lt

Vilnius University

About me

- Originally from Mumbai, India.
- Current positions:
  - PhD Student at Vilnius University; Topic: Deep Imitation Learning for mobile robot navigation.
  - Researcher at Neurotechnology, located in Vilnius, Lithuania.
- Current and past projects:
  - Brain Computer Interface research.
  - Mobile robot navigation research.
- Education Background:
  - Masters in Informatics
  - Bachelors in Computer Engineering
  - Pre-Bachelors Diploma in Computer Engineering

Vilnius University

Vilnius University's Institute of Data Science & Digital Technologies

- The institute is multi-disciplinary.
- Virginijus Marcinkevičius' lab covers various areas of research.
- Some of them are:
  - Robotics
  - Computer vision
  - Artificial Intelligence
- Some other topics researched in our institute:
  - NLP
  - Security

Vilnius University

Urban Driving

- The problem of Urban Driving (or autonomous driving) represents a making a self driving car being able to drive seamlessly in real world urban environments.
- The techniques used in research area also closely involve the field of robot navigation, where robots are made to be able to navigate in environments such as factories, workshops, etc.
- Due to the rise of the need of automation in recent years, this research area has gotten lots of attention.

Vilnius University

Urban Driving

- Driving in urban areas is a complex problem, due to the vast possibilities of situations one can run into.
- Current state-of-the-art is either based on modular approaches or end-to-end learned approaches.
- **Modular approaches** are able to generalise to unseen environments and use multiple sensors to perceive the world, but they are hard to engineer.
- **End-to-end approaches** learn the skill of driving directly from data, hence do not require high engineering effort.
- But end-to-end approaches struggle to generalise well to new environments.
- Our work revolves around extending the research in end-to-end approaches.

Pre-training

- Training a learner (neural network model) from scratch requires excessive amounts of data, resources and time.
- Hence pre-training of models has been widely adopted in natural language processing, object detection, etc.
- Most state-of-the-art autonomous driving methods, based on end-to-end approaches, use ImageNet based classification pre-training for training driving models.
- Only a handful of methods have so far explored pre-training methods specifically for the task of driving.

Problems with current trend of pre-training

- The task of image classification over the ImageNet dataset maybe highly unrelated to the task of driving.
- Classification also tends to narrow down the idea of image understanding to a single concept.
- There seems to be a lack of exploration in this research direction.

Vilnius University

Monocular Depth Estimation

- Monocular depth estimation is a task of estimating depth given a RGB image.
- The idea is to remove the need of expensive depth cameras and be able to infer the depth of an image from a simple RGB camera.
- Recent methods have attempted to approach this task with the use of neural network based learning methods.
- Current state-of-the-art methods show impressive ability to infer depth of RGB images.

Vilnius University

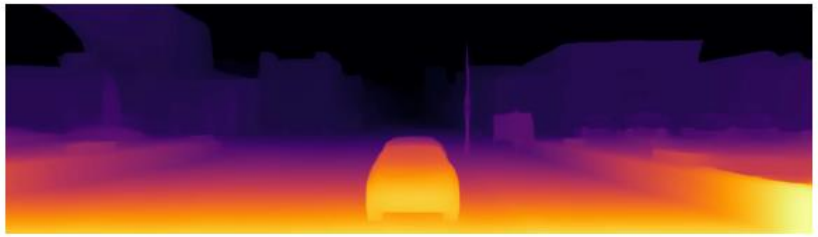
Depth Anything

- Depth Anything is a method which trains on large dataset of 62 million images for the task of monocular depth estimation.
- With such a large dataset, Depth Anything forms a foundational model that can be used for many tasks.
- Depth Anything shows seamless ability to perform zero-shot generalization for estimating depth.
- Trained using the attention-based transformer model.

Vilnius University

Depth Anything

(a)



(b)

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

Vilnius University

Proposed method

- We propose pre-training an encoder over the task of monocular depth estimation using the Depth Anything method.
- Thereon, using the pre-trained encoder to train over the task of autonomous driving.
- We propose using the architecture from recent methods that explore pre-training [1] and compare with a baseline approach.
- The baseline model could be the most common form of pre-training, i.e., classification-based pre-training over the ImageNet dataset, as in [1].

[1] S. Juneja, P. Daniusis, and V. Marcinkevicius, “Visual place recognition pre-training for end-to-end trained autonomous driving agent,” IEEE Access, vol. 11, pp. 128 421–128 428, 2023.

Proposed method

- Additionally, our proposed method also hypothesizes that additional understanding of depth of an image can provide more information than just what has so far been provided by the usual image encoder.
- To test out the proposed method, we revealed the planned implementation details in the paper.

Implementation plan

Vilnius University

- We plan to:
  - Collect data as per the CARLA Leaderboard benchmark standards, with use of the Roach method (cited in the paper).
  - Re-training the whole architecture consisting of the pre-trained encoder with imitation learning.
  - Iteratively collect additional data with the DAgger algorithm, if needed.
  - Evaluate the performances on the Leaderboard benchmark.
  - Compare with a baseline model that uses the most common form of pre-training.

Vilnius
University

Thank you

Contact details:
shubham.juneja@mif.stud.vu.lt