Towards Practical and Efficient Computer Vision Models for Extreme-Weather Scenarios in Urban Mobility

Vishal Vinod, Savita Choudhary
Sir M. Visvesvaraya Institute of Technology
{vishal_1mv16cs125, savitha_cs}@sirmvit.edu

Abstract
Efficiently prototyping computer vision models for urban mobility and research on more representative datasets are essential steps towards building fully autonomous vehicles. There are several challenges pertaining to the extreme weather and challenging conditions scenario that is now driving research in the field. Specific themes related to the challenges in training computer vision models for autonomous vehicles can be formulated as efficiency and performance-based problems. This paper's primary contribution is to identify potential challenges in training vision models for autonomous vehicles, discuss new and efficient solutions from the latest research, and formulating ideas that help improve the datasets. Hence, the generalization capability of the models being trained. Specifically, we describe Domain Adaptation for semantic segmentation, Domain translation using Image-to-Image translation, approaches to create high-quality datasets, and advances in video translation.

Introduction
Research in mobility has lately been inclined towards complex models and overly-engineered features with benchmarks on idealistic datasets that are non-representative of real-world application. Some of the latest ML models typically have a sizeable computational footprint and are quite complicated in terms of large parameter counts. An impactful step forward will be to work towards efficient models trained on representative data focusing on the resource constraints for model deployment. For tasks that can impact low resource deployments such as vision systems in autonomous vehicles, training efficient and interpretable ML models becomes significantly more important, especially with the alarming increase in adversarial attacks that test the robustness of current vision models. In this paper, we discuss some of the latest in efficient model building as a step towards re-designing efficient research directions, works for real-world applications, and paves the way for a new niche in urban mobility.

Challenges for Autonomous Vehicles
There has been a growing trend of models with improved performance for vision systems with resource-constrained deployment scenarios but have been complex, over-engineered, or difficult to train. The current state-of-the-art for vehicular semantic segmentation on the Cityscapes dataset is the hierarchical multi-scale attention model (Tao et al., 2020) that uses various image scales and attention modules to improve performance. The HRNet model architecture, along with object-contextual representation (Yuan et al., 2019) and SegFix, a segmentation refinement module (Yuan et al., 2020), is another example of a complex model that can be prohibitively expensive to train. This can be tackled effectively by developing simple models such as the Detection Transformer (Carion et al., 2020) or advances in unsupervised domain adaptation for related datasets.

The Conditional Domain Normalization (Su et al., 2020) is one such DA technique that adapts to the domain shifts at different levels of representation in the detection module to adapt to large domain gaps effectively. Such advances invariably lead to fine-tuning models across datasets with a few-shot approach.

Clear-weather Dataset Bias
Most computer vision models for self-driving cars are trained on clear-weather daylight images and video. This inherently biases the deployed models towards performing well only for such conditions. These models need to have fast real-time performance and must also be able to perform well in adverse conditions with noisy data such as from rain, snow, or low-light conditions. This topic makes for an exciting investigation because it is a relatively new research field with the potential for work that reduces the
dataset bias and helps build and train efficient models that can generalize well across various weather conditions.

ForkGAN (Zheng et al., 2020) is an interesting model building on the concepts of cross-cycle consistency and from DRIT (Lee, H. et al., 2018) precisely to translate images from night scenes to clear-weather daylight scenes. Such models enable practitioners to utilize state-of-the-art detectors trained on large scale daylight road scenes without using few-shot domain adaptation approaches and instead use an image-to-image translation model. Several such models for diverse image-to-image translation in a multimodal setting may suffer from mode collapse, especially when the domain invariant attributes are similar across various scenes.

Tackling Dataset Bias for Extreme-weather scenes

Image deraining is the task of effectively modeling the heterogeneous noise in rainy images and restoring them to clear weather images. Several techniques exist for image deraining, such as using a physics-based approach to model rain (Yang, W. et al., 2020), using attention models like SPANet (Wang, T. et al., 2019), and modeling raindrops in multiple scales using MSPFN - Multi-Scale Progressive Fusion Network (Jiang, K. et al., 2020).

Creating a dataset for such rainy images, especially for road scenes, is challenging and has led several practitioners to use synthetic noise to train models or use domain adaptation techniques. Interestingly, (Xiaowei et al., 2019) have shown that synthetic rain in images modeled as a 2D noise layer does not capture the inherent depth model of rain. They conclusively proved that training deraining models on synthetic rainy images modeled as a depth completed rain layer performs much better on real-world rain images. Several works on de-raining and domain adaptation focus on generalizing across domains with varying amounts of data augmentation techniques that mimic extreme weather conditions. Their performance on real-world extreme weather scenes lacks especially because augmentation techniques sometimes model the noise layer uniformly based on pseudorandom parameters for raindrop size or fog visibility, whereas real scenes have a very stochastic noise model. Hence, real-world extreme weather datasets are necessary for training vision models in urban mobility.

The contribution of the Foggy-Cityscapes dataset from (Sakaridis, C. et al., 2018) is an essential step towards creating a synthetic dataset that is still a good representation of the real-world depth-completed noise model. The procedure to create such an extreme-weather version of the Cityscapes dataset (R M. Cordts, et al., 2016) is very resource-intensive and requires semantic maps and depth information to apply a smoothened noise model to the clear-weather image. Specifically, the process includes depth completion, depth de-noising, and applying the noise layer, with approximations for the depth and speck size for rain and snow. Newer techniques with segmentation map-guided depth completion also show promising results but require depth information for the domain conversion.

This task can also be formulated as a style transfer problem and can effectively perform the translation even from a zero-shot setting with effective disentangling of the domain style space. Several models such as the peer-regularized feature recombination for zero-shot style transfer (Svoboda et al., 2020), style transfer using feature perturbations (Wang, Z. et al., 2020), and style transfer using graph cuts (Zhang Y. et al., 2019) are efficient and effective solutions for creating rainy or snow-based datasets consistent with real-world images.

Photorealistic Real-time Video Translation

Recent advances in video-to-video translation from photorealistic vid2vid (Wang T. et al., 2018) leading up to the very recent world-consistent vid2vid (Mallya et al., 2020) shows great potential to model challenging scenarios for vehicles consistent with the semantic map of the scene.

Semantic image synthesis (Park et al., 2019) is very useful to generate photorealistic road scenes given just the semantic segmentation map. Video-to-video synthesis uses the semantic map and the depth map to generate structure consistent videos of road scenes. World-consistent vid2vid aims to solve the long-term consistency problem in vid2vid by using guidance images to ensure consistency across views and time. It can render and generate video translations with consistent viewpoints in near real-time speeds and colorize the scene consistent with the 3D point clouds.

Broader Impact

Many of the challenges in autonomous vehicles and mobility can be formulated as research ideas to build and train performant models focusing on the efficiency requirements to deploy in self-driving vehicles. Such advances in solving challenging problems using efficient methods show immense potential to tackling the challenges for level 5 fully-autonomous systems. Models also need to robust to attacks as well be well trained for challenging scenarios. Efficient modeling techniques and the methods for creating datasets without accruing a large budget goes a long way towards attracting scientific minds to work on such niche ideas. Striving for high-quality data, a well-formulated real-world problem, and a focus on real-world applicability can hasten the growth in urban mobility.

References


