Domain Adaptive semantic segmentation using masked autoencoders

KIM, Sukmin
3035445479

Supervisor: Dr. Luo Ping
Table of Contents

1. Project Background .................................................................................................................................................. 3

2. Project Objective ...................................................................................................................................................... 6

3. Project Methodology .................................................................................................................................................. 7
   3.1 Previous works ................................................................................................................................................... 7
   3.2 Pretraining masked autoencoders ...................................................................................................................... 8
   3.3 Finetuning with UDA ......................................................................................................................................... 8
   3.4 Experiments ...................................................................................................................................................... 9

4. Project Schedule and Milestones ............................................................................................................................ 9

5. References ............................................................................................................................................................... 10
1. Project Background

Computer vision has been a widely popular area these days, thus heavily researched and applied to a wide range of tasks. This project is particularly looking for semantic segmentation tasks in autonomous driving. Semantic segmentation is the task of classifying images into predefined categories per pixel. For autonomous driving, popular datasets like Cityscapes classify each pixel in the images into categories such as roads, sidewalks, and trains. The task aims to create a deep learning model that effectively and accurately classifies these images which helps self-driving cars to detect drivable areas like roads.

Autonomous driving is a rising field both in the industry and in research, where recent computer vision breakthroughs are improving rapidly. One of the concerns, however, is that the dataset is usually biased toward certain cities. If the model is trained in a dataset collected from Los Angeles, for example, we cannot ensure the performance will be similar to Hong Kong due to its varying road size, traffic light, vegetation and several different factors. This is where the field of domain adaptation becomes popular, aiming for the model to be robust in different cities, or different domains in the more general case.

Another concern is that the semantic segmentation task suffers from collecting massive amounts of labeled images. This is because per-pixel annotation is time-consuming, taking 1.5 hours per image if labeled by a human [1]. Many people today support data-centric deep learning, which emphasizes the role of quality and quantity of data in overall performance. Therefore, there have been several approaches to using datasets more wisely.
First, unsupervised domain adaptation is an approach that reduces the annotation effort by using synthetic data. Unlike real data, synthetic data is collected from a simulated environment, thus auto-labeled by the software. This allows mass collection of data, which can be directly input into the model. The use of synthetic datasets is not limited to autonomous driving but is a trend for all areas of artificial intelligence. Figure 1 below shows synthetic data will completely overshadow real data in the future of AI. However, the biggest challenge is that the domain gap between synthetic and real data results model trained from the synthetic data shows an enormous performance drop on real data. Previous research investigates methods to adapt the network to real data by reducing domain gaps in diverse ways.
Second, the self-supervised method relieves the burden of data labeling. Specifically, masked autoencoders are popular self-supervised models in that the model is trained with an unlabeled dataset [3]. Unlike supervised learning which inputs RGB image and then outputs the labeled image associated with the task, self-supervised learning like masked autoencoders input masked RGB image and then output reconstructed RGB image without masking, letting the model predict the masked area of images. Then the model is the finetuned respect to the task. It is common now in many deep learning tasks where the self-supervised approach surpasses the results of supervised learning. I believe this superiority can be applied to the autonomous driving task. On top of that, the model is more efficient by using less GPU memory due to masking images. As masking images leave about 25% of images into training, this ensures faster training with a larger batch size or more complex model. Such an advantage matches the current problem of unsupervised domain adaptation models in self-driving cars. Current SOTA models face challenges to GPU memory constraints, thus images are resized or randomly cropped into a smaller resolution. I believe the use of reduced resolution is the main cause why current UDA models still have a huge gap in performance to supervised models in autonomous driving systems.
The project proposes to overcome the dataset constraint presented above by fusing both approaches - unsupervised domain adaptation and self-supervised learning. Previous research work attempts to improve accuracy using the former method but never mixed two approaches. I believe if they are used together appropriately, the performance can surpass previous models.

2. Project Objective

- With associated empirical analysis, propose a methodology and architecture that outperforms the previous unsupervised domain adaptation models on synthetic-to-real adaptation
- Implement a code of the paper, based on mmsegmentation framework, that includes configuration files, training steps, and codes used for experiments
- Present a thorough analysis of the proposed architecture and methodology with different datasets, ablation studies, and comparisons with previous research. Experiments would be done exhaustively by exploring different hyperparameters and their tuned results
- Write a paper as a final deliverable, submitted to the conference if possible. While writing a paper, learn how to write a research paper in a conference style
3. Project Methodology

3.1 Previous works

Previous works use self-training for unsupervised domain adaptation to effectively adapt the model trained from the source(synthetic) image to the target(real) image. Figure 3 presents the self-training algorithms, which consist of two types of images, losses and models. Unsupervised domain adaptation assumes the label from the target domain is not available, thus using the term “unsupervised”. Therefore, self-training relies on pseudolabels created from the teacher model. To reduce the GPU memory burden, the parameter of the teacher net is updated by the exponential moving average(EMA) from the student net, which is trained by the source image-label pair.

Experiments for unsupervised domain adaptation are done in various ways. For autonomous driving, synthetic-to-real and clear-to-adverse weather adaptation are two popular ways to evaluate the model’s performance. However, many models suffer from GPU memory constraints due to datasets with high resolution and complexity of models to be trained. For instance, the GTA dataset has a 2048 by 1024 resolution, but the UDA model is trained with 1024 by 512 to fit into the memory. Also, models may learn unnecessary texture information from the source domain, which harms performance when evaluated in the target domain due to the discrepancy of fine details in the images between the two domains.

Figure 3: self-training module for unsupervised domain adaptation [4]
3.2 Pretraining masked autoencoders

For pretraining, the project will use masked autoencoders, which relieves the GPU memory constraints because it uses only unmasked image area and light decoder to train. Also, the approach preserves and lets the model learn general information such as shape and color, but not the fine texture specific to the domain, which is more appropriate to adapt the model to different domains. Moreover, training is done much faster compared to previous supervised models, implying debugging and experimentation can be done more thoroughly with limited time.

There are several approaches to pretrain the network by varying datasets and model architecture. Currently, a pretrained model with a vision transformer is available in public, but the current SOTA model for unsupervised domain adaptation uses SegFormer, a transformer-based semantic segmentation network, which proved to be more robust compared to previous vision transformers [5]. Therefore, one approach is to pretrain SegFormer-based masked autoencoder and compare it with the previous vision transformer models. Furthermore, there is a wide range of datasets available to be pretrained. All the previous models are pretrained using ImageNet-1K, with about 1.2 million images for training. Although it is possible to pretrain with autonomous driving datasets including GTA, Cityscapes, and Synthia, their dataset size is too insignificant compared to ImageNet-1K, thus highly prone to overfitting. It is possible to implement different augmentation methods to increase dataset size, but it is still insufficient. Therefore, my first attempt for pretraining will be on SegFormer model pretrained with the Imagenet-1K dataset, with visualizing the reconstructed image and logging reconstruction MSE loss.

3.3 Finetuning with UDA

Several UDA models in autonomous driving are based on the self-training module, and the project will also use the module for finetuning. Previous models improved performance significantly by sampling rare classes more frequently. Most recent works use multi-resolution training by models to learn both high-resolution detail and low-resolution context information of the image, then fuse prediction results using
attention mechanisms. However, previous works never attempted to apply the pretrain-finetune paradigm for the task. After pretraining the masked autoencoders, I will evaluate the result first by directly applying it to HRDA, the current SOTA model, to compare the result [6]. In finetuning phase, the decoder of the model will be substituted into the segmentation head, while reusing the pretrained encoder. Further improvement could be done by proposing new architecture more appropriate for fine-tuning.

3.4 Experiments

Experiments will be done in a wide range of aspects. The IoU of each class and mIoU scores, the metric used for evaluating semantic segmentation performance of the model, will be presented to the synthetic-to-real unsupervised domain adaptation task, in comparison with the previous model. Besides the quantitative result, a qualitative comparison will be conducted, which emphasizes how the model is classified into small-sized categories, which previous models suffer to predict well. Furthermore, an experiment by tuning possible hyperparameters – crop size, learning rate, scheduler, EMA rate – will be presented, justifying why certain hyperparameters are selected in the project’s final model. Finally, ablation studies will be conducted to analyze component-wise improvement in performance.

The code will be written based on the mmsegmentation framework, which is a popular framework for semantic segmentation tasks, and will be released to GitHub.

4. Project Schedule and Milestones

<table>
<thead>
<tr>
<th>Period</th>
<th>Milestones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep – Oct 2022</td>
<td>• Continue work from URFP</td>
</tr>
<tr>
<td></td>
<td>• Research recent development on unsupervised domain adaptation and masked autoencoders</td>
</tr>
<tr>
<td></td>
<td>• Research recent development in autonomous driving</td>
</tr>
<tr>
<td></td>
<td>• Prepare for phase 1 deliverables</td>
</tr>
<tr>
<td>Oct – Dec 2022</td>
<td>• Investigate different ways to pretrain masked autoencoders by various encoder, decoder, dataset, image size, masking ratio, and others.</td>
</tr>
<tr>
<td></td>
<td>• Implement codes and visualize results of reconstructed images</td>
</tr>
</tbody>
</table>
Dec 2022 – Jan 2023
• Propose method and architecture to finetune pretrained masked autoencoders
• Prepare for phase 2 deliverables

Jan – Apr 2023
• Experiment with the proposed method and architecture and compare it with previous SOTA results
• Take ablation studies, hyperparameters tuning, and performance comparison with previous models
• Write a paper, following the style of the targeted conference
• Implement the code of experiments and release it to GitHub
• Prepare for phase 3 deliverables

Apr – May 2023
• Prepare for the project exhibition

5. References


