Doublet Contrastive End-to-End Semantic Segmentation for Autonomous Driving under Adverse Weather

Jongoh Jeong and Jong-Hwan Kim
Korea Advanced Institute of Science and Technology (KAIST)

INTRODUCTION

Motivation
- Intelligent driving systems require safe and accurate perception of the surroundings in dynamically changing environments, most heavyweight models that focus primarily on the performance are not suitable for practical real-time deployment.
- Most semantic segmentation applications under adverse driving conditions, or "unusual road or traffic conditions that were not known" as defined by the US Federal Motor Carrier Safety Administration [1] (e.g. fog, nighttime, rain, snow), has yet to be explored further [2].
- 1.3 million death toll of road traffic crashes every year, and the risk of accidents in rainy weather, for example, is 70% higher than in normal conditions [3, 4].

Contributions
- We propose an end-to-end doubly (image- and pixel-levels) contrastive learning strategy for a lightweight semantic segmentation model to eliminate the pre-training stage in the conventional contrastive learning approach without requiring a large training batch size or a memory bank.
- Our training method achieves 1.34%p increase in mIoU measure from the baseline for loss-only objective with the SwiftNet architecture (ResNet-18 backbone), running inference at up to 66.7 FPS in 2048x1024 resolution on a single Nvidia RTX 3090 Mobile GPU.
- We verify that replacing image-level supervision with self-supervision in our supervised contrastive objective achieves comparable performance when pre-trained with clear weather images.

PRELIMINARIES

Self-supervised Contrast
- For a set of $N$ randomly sampled image-label pairs ($X_i$, $y_i$), i = 1, ..., $N$, and a corresponding multi-view set of augmented samples from the same sources ($X_i'$, $y_i'$), $i$ = 1, ..., $N$, compute a similarity score:

$$s_{xy} = \sum_{i=1}^{N} \left( X_i, X_i' \right)$$

where

$$\left( X_i, X_i' \right) = \sum_{i} \left( x_i, x_i' \right)$$

Supervised Contrast
- Image-level
  - More than one sample belonging to each image class label

$$s_{xy} = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{N} \exp \left( s_{xy} \right)$$

- Pixel-level
  - Similar to image-level, but applied to pixel-wise features

$$s_{xy} = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{N} \exp \left( s_{xy} \right)$$

where $N$ is the number of pixels.

METHODOLOGY

Doubly Contrastive Supervised Semantic Segmentation
- Base model: multi-scale pyramidal encoder followed by up-sampling blocks (SwiftNet).
- Training objective: Segmentation loss + Contrastive Losses

$$L = \lambda_c \cdot (L_{\text{Seg}} + L_{\text{pix}}) + \lambda_s \cdot L_{\text{sup}}$$

where

$$L_{\text{sup}}(\theta(p), \tilde{y}_k) = -\sum_{y_k} \log p(y_k | \theta(p))$$

EXPERIMENTAL RESULTS

Datasets
- Cityscapes (pre-training)
- Adverse Conditions Dataset with Correspondences (ACDC) for training & evaluation

PRELIMINARIES

Self-supervised Contrast
- For a set of $N$ randomly sampled image-label pairs ($X_i$, $y_i$), i = 1, ..., $N$, and a corresponding multi-view set of augmented samples from the same sources ($X_i'$, $y_i'$), $i$ = 1, ..., $N$, compute a similarity score:

$$s_{xy} = \sum_{i=1}^{N} \left( X_i, X_i' \right)$$

where

$$\left( X_i, X_i' \right) = \sum_{i} \left( x_i, x_i' \right)$$

Supervised Contrast
- Image-level
  - More than one sample belonging to each image class label

$$s_{xy} = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{N} \exp \left( s_{xy} \right)$$

- Pixel-level
  - Similar to image-level, but applied to pixel-wise features

$$s_{xy} = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{N} \exp \left( s_{xy} \right)$$

where $N$ is the number of pixels.

RESULTS

Our approach achieves 1.34%p and 1.33%p increases in mIoU using the ResNet-18 and 34, respectively, from the baseline. In addition, we found that our method is more advantageous in more adverse conditions like nighttime, rain and snow than relatively easier condition like fog. We assume this is due to the difficulties where the visibility is limited by the low illumination, rain streaks, rain/snow covered roads, and highly dense fog. Our approach corrects false positive predictions observed in the baseline results, and results in more consistent predictions of objects of relatively large sizes.

CONCLUSIONS

We proposed an end-to-end doubly contrastive learning approach to semantic segmentation for self-driving under adverse weather, exploiting image-level labels to semantically correlate RGB images taken under various weather conditions and pixel-level labels to obtain more semantically meaningful representations. In our method, the two supervised contrasts complement each other to effectively improve the performance of a lightweight model, without a need for pre-training or a memory bank to associate images across various adverse conditions for global consistency.

REFERENCES

1. Federal Motor Carrier Safety Administration. code of federal regulations. Accessed May 23, 2022 (ONLINE).
2. Christos Sakardis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. ICCV, October 2021.
3. Jean Andrey and Sam Yagar. A temporal analysis of rain-related crash risk. Accident Analysis & Prevention, 25(4):465-472, 1993. ISBN 0001-4575. doi: https://doi.org/10.1016/0001-4575(93)90076-9.
4. World Health Organization. Road traffic injuries, 2021.

ACKNOWLEDGMENTS

This work was supported by the Korea Institute for Information & Communications Technology (IITP) under Grant 2021-004575 through the Korean Government (Ministry of Science and ICT (MSIT)); Development of Artificial Intelligence Technology that Continuously Improves itself as the Situation Changes in the Real World.)