JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 16, No. 3, September 2022, pp. 178-184

Switching DNN for Autonomous Driving System

Yu-Seung Ma
Electronics and Telecommunications Research Institute, Daejeon, Korea
ysma@etri.re.kr

Hojae Han and Seung-won Hwang
Dept. of Computer Science and Engineering, Seoul National University, Seoul, Korea
stovecat@snu.ac.kr, seungwonh@snu.ac.kr

Abstract
In autonomous driving system, building a rigorous object detection model unaffected by conditions, such as weather or time-of-day, is essential for safety. However, as deep learning models are often limited in generalizability, training over the entire data collection can be suboptimal, e.g., daytime training instances hinder the training for nighttime prediction. We call this curse of multitasking (CoM), which was first observed in multilingual training, where training a multilingual model can be suboptimal, compared to multiple monolingual models. Our contribution is observing CoM in autonomous driving, overcoming the problem by building multiple mono-task models, or specialized experts for each task, then switching models according to the input condition, enhancing the overall effectiveness of the detection model. We show the effectiveness of using the proposed strategy in both YOLOv3 and RetinaNet models on BDD dataset.

Category: Smart and Intelligent Computing

Keywords: Deep neural network; Object-detection; Autonomous driving; Software engineering

I. INTRODUCTION

Deep learning models have been successful in various tasks for autonomous driving systems, as surveyed in [1]. Our target task is object detection, detecting the presence of relevant objects in images and classifying them into relevant classes; many deep neural network (DNN) models such as Faster R-CNN [2], RetinaNet [3], YOLO [4], and SSD [5] have achieved good performance results. Large-scale and high-quality datasets for training was critical for their success, such that follow-up research focused on increasing both the amount and the diversity of training data (as summarized in Section II).

Conversely, this study explores the dark side of making datasets bigger, hypothesizing that a larger dataset may capture scenarios that are too diverse. This hypothesis is inspired by the curse of multilinguality [6, 7]: Table 1 from [7] shows a multilingual model trained on the entire dataset is less effective than a monolingual model, KO and RU trained on a smaller subset, containing only instances in Korean and Russian, in diverse language tasks, such as Named Entity Recognition (NER), Sentiment Analysis (SA), Question-Answering (QA), Universal Dependency Parsing (UDP), and part-of-speech (POS) tagging.

In this study, we explore whether the diversity in a large-scale autonomous driving dataset, such as BDD100k (Berkeley Deep Drive) [8], has similar negative effects.

Open Access http://dx.doi.org/10.5626/JCSE.2022.16.3.178 http://jcse.kiise.org

Received 22 June 2022; **Accepted** 14 September 2022
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

Table 1. An empirical validation for the curse of multilinguality [7]

Lang	Model	NER	SA	QA	POS
		*F*1	Acc	EM/*F*1	Acc
KO	Mono	**88.8**	**89.7**	**74.2/91.1**	**97.0**
	Multi	86.6	86.7	69.7/89.5	96.0
RU	Mono	**91.0**	**95.2**	**64.3/83.7**	**98.4**
	Multi	90.0	95.0	63.3/82.6	98.2

Monolingual BERT (mono) generally performs better than multilingual BERT (multi) in each target language domain.

In the BDD100k dataset, various driving road images exist representing different weather conditions such as rain and snow, and the time of day and night. This study focuses on images of two distinct domains, daytime and nighttime. It is suitable as a division criterion because it is automatically distinguishable whether it is night or day through time information and illuminance sensor value. We consider that images taken during the daytime and nighttime may have the same effects as documents written in Korean and Russian. We use the curse of multi-tasking (CoM) as an umbrella term to represent both challenges.

In other words, our first research question is validating CoM, by observing whether a multi-task model trained on all images M_{all} is less effective under specific conditions, such as day or night, than a mono-task model, M_{day} or M_{night}, trained only on images taken during the daytime or nighttime. Our next step is to propose an approach that outperforms the current state-of-the-art by switching between mono- task models for autonomous object detection systems.

The following items are the main contributions of our study.

- We show that the mono-task model trained separately for daytime and nighttime zones in an autonomous driving system achieves a higher recognition accuracy during day and night than the multi-task model trained on a full dataset.
- We propose an object detection method that switches between mono-task DNN models and a multi-task DNN model based on time information. Then, we demonstrate the effectiveness of the proposed switching model through experiments.

The contents of this study are as follows. Section II describes related work. Section III summarizes our approach, the DNN switching technique for an autonomous driving system. Section IV presents the experimental results for research questions. Section V concludes with the limitations and future research directions.

II. RELATED WORK

Driven by the remarkable advances of DNNs, several DNN models have been proposed as object detectors in autonomous driving systems. For state-of-the-art object detection models, we refer to survey [1], including discussions on baselines RetinaNet [3] and YOLOv3 [4] used in this study.

In order to improve the performance of the DNN models for an autonomous driving system, efforts have been made to make the dataset bigger and more diverse. Most of the available autonomous data so far consists of images of sunny or clear days, but models trained on clear daylight conditions cannot generalize well in bad weather. Therefore, autonomous driving datasets should cover driving images in diverse weather conditions and at different time-of-the-day.

Recently, the GAN-based style transformation [9, 10] has been actively used to augment training data of diverse weather and time. Zhang et al. [9] synthesized driving scenes with various weather conditions, and developed a metamorphic testing module for DNN-based autonomous driving systems. It transforms data into a sunny-to-snowy style, sunny-to-rainy style. Ostankovich et al. [10] propose a cycleGAN-based augmentation approach to help solve segmentation and detection problems occurring at nighttime. GAN-based domain adaption is also applied for virtual-to-real and real-to-virtual driving scenes [11, 12]. SG-GAN [11] is designed to automatically convert an unrealistic virtual-world scene to a realistic scene by employing a new soft gradient-sensitive loss and a semantic-aware discriminator. DUdrive [12] employs an unsupervised network to transform real driving images into simpler virtual images, and utilizes the virtual image to improve autonomous driving system quality.

Our work is orthogonal to both modeling and data generation efforts, discussed above, by examining and detecting whether the dataset is too diverse for training purposes, or, whether CoM holds, then aggregating mono-task models to overcome such challenges.

III. METHODOLOGY

In this work, we study problems where the subset of each mono-task model specializes, which we denote as "subdata" that can be automatically identified: in our

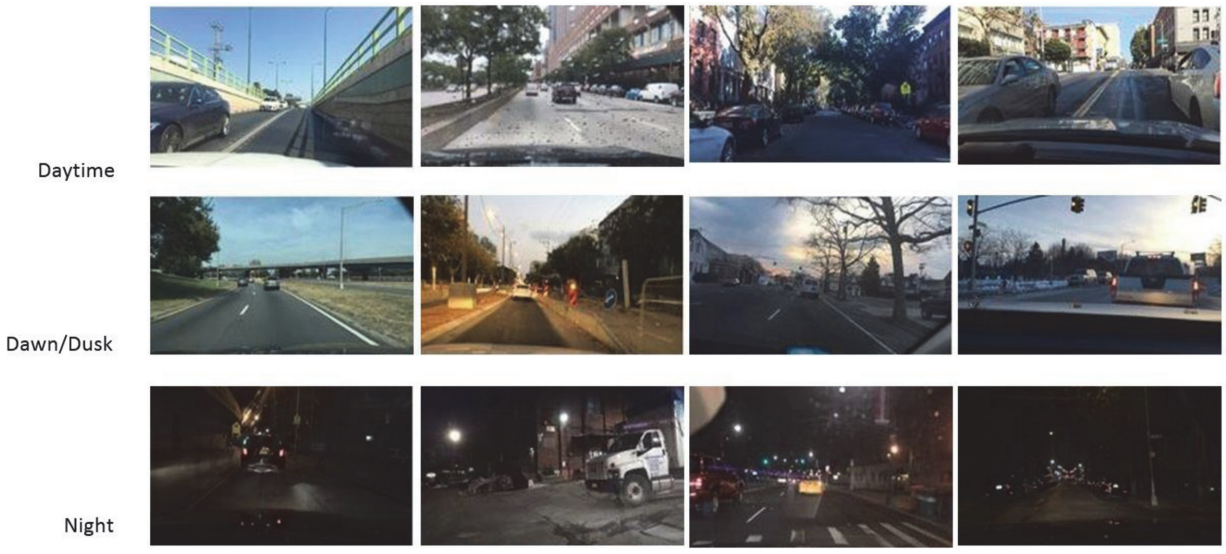

Fig. 1. Example of BDD images.

problem, it is possible to distinguish between night and day through time information and the value of the illuminance sensor.

Fig. 1 shows examples of driving images at different times such as daytime, dawn/dusk and night, provided by the BDD100k dataset [8]. Day and night driving images can be easily distinguished. However, in the case of dawn/dusk, the distinction is a little vague even with the human eye; sometimes, they may be perceived as daytime or night images. Thus, we do not consider such ambiguous images in generating mono-task models.

Let D be the full dataset on which the object detection model will be trained. Then, we divide the set D into $D_{daytime}$, D_{night}, and $D_{etc.}$, where each represents daytime data, nighttime data, and the data that is vague to distinguish whether it is daytime or nighttime. Then, these three subsets are all disjoint.

$$D = \{x | x \in D_{daytime} \text{ or } D_{night} \text{ or } D_{etc.}\} \quad (1)$$

$$\emptyset = D_{daytime} \cap D_{night} \cap D_{etc.} \quad (2)$$

Fig. 2 illustrates the overview of our approach. Excluding the vague data, we create two mono-task models, $M_{daytime}$ and M_{night}, independently trained model instances for each subdata, $D_{daytime}$ and D_{night}. Note that these subspecialized mono-task models, $M_{daytime}$ and M_{night}, share the same model architecture as the original multi-task model, M_{all}: they are just trained using different subsets of the original training data. Intuitively, we can consider the mono-task model to perform particularly better in smaller subtasks, rather than performing well in general. Therefore, if it is determined that it is definitely day or nighttime, it is advantageous for accuracy to make inferences by switching

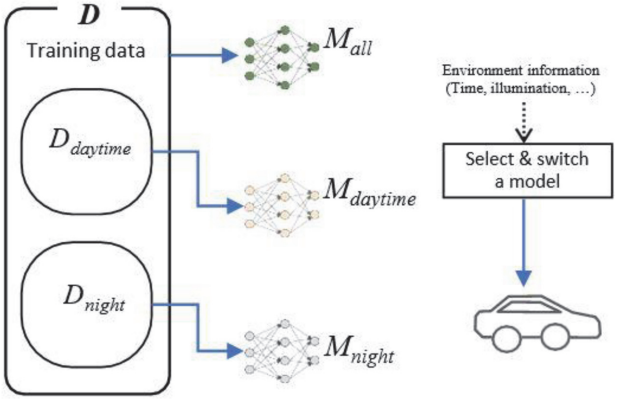

Fig. 2. Overview of our approach.

to a specialized mono-task model at day or nighttime rather than using a multi-task model trained using the entire data.

Note that our approach still requires the traditional multi-task model, M_{all}. We keep a multi-task model, for ambiguous boundary cases, such as dawn/dusk images in Fig. 2, where M_{all} is better. Based on this concept, the DNN model switching is repeated in the order of M_{all}, $M_{daytime}$, M_{all}, and M_{night} per day.

Not surprisingly, this type of switching incurs additional memory overhead for maintaining multiple models. However, in a safety-critical system such as autonomous driving, accuracy is typically considered more critical over such an overhead.

IV. EMPIRICAL STUDY

We follow the empirical setup [7] to compare multi-

Table 2. Number of images in the BDD dataset

	Training data	Validation data
Daytime	36,800	5,258
Nighttime	28,028	3,929
Dawn/Dusk	5,033	1,788
Unknown	139	35
Total	70,000	10,000

task and mono-task models (Section IV-A), for validating CoM (Section IV-B) and overcoming such a curse (Section IV-C).

A. Setup

For dataset, we use BDD [8], which consists of 70,000 training data and 10,000 validation data. We chose the BDD dataset because it provides a large number of nighttime driving images, compared to other datasets where daytime driving images are dominant.

Table 2 shows the statistics of this dataset, where we consider two major categories, Daytime and Nighttime, and treat the rest as "etc".

For object detection models, we use YOLOv3 (https://github.com/qqwweee/keras-yolo3) and RetinaNet (https://github.com/fizyr/keras-retinanet), both on keras, to share common components. Training DNN models for autonomous systems is usually a very expensive task. For this reason, we have only covered those two models, of the mainstream object detection models, in this experiment. We are cautiously anticipating that the results would be similar if other models were used.

The training was conducted on GeForce RTX 2080 GPU and intel i7-9700 CPU hardware specifications. As a learning parameter, the batch size was set at 32 and a maximum of 100 epochs was performed. For accuracy metrics, we follow the convention of object detection, to adopt mean average precision (mAP), comparing the ground-truth bounding box to the detected box and returns a score over 0 to 1.

B. RQ1: Does CoM hold?

For each DNN, we generated three trained models, one multi-task model, Mall, and two mono-task models, M_{day} and M_{night}. The multi-task model, M_{all}, is trained using all of the 70,000 training images in Table 2. The mono-task models, M_{day} and M_{night}, are trained only on daytime and nighttime subsets in Table 2. For Mall training, it took about a week in our experimental environment, and M_{day} and M_{night} took less than that.

The comparison of object recognition accuracy was performed using the validation data in Table 2. Table 3 shows the mAP scores of mono-task models (M_{day} and M_{night}) compared with M_{all}. For the 5,258 daytime validation data, M_{day}showed the highest mAP value, and for the 3,928 nighttime validation data, M_{night} showed the highest performance.

Table 3. mAP of mono-task models, compared with M_{all} (unit: %)

DNN model	Validation data		
	Daytime	Nighttime	All
YOLOv3			
M_{day}	**24.60**	18.11	22.85
M_{night}	17.41	**22.93**	19.15
M_{all}	22.88	22.49	**22.87**
RetinaNet			
M_{day}	**29.33**	19.18	26.62
M_{night}	21.63	**24.30**	22.32
M_{all}	27.56	23.44	**26.81**

These results confirm the CoM—that is, M_{all} is outperformed by M_{day} and M_{night} in daytime and nighttime instances, respectively. Not surprisingly, M_{day} and M_{night} cannot generalize to untrained scenario of nighttime and daytime instances, respectively.

The last "all" column of Table 3 shows the mAP values using a total of 10,000 validation data including all images such as daytime, nighttime, dawn/dusk, and unknown. For all data, M_{all} shows the highest mAP value. However, it only means that the M_{all} model works fine on average. Experimental results suggest using M_{day} and M_{night} for day and night, respectively, and the M_{all} in other situations. These findings were consistent for both YOLOv3 and RetinaNet models.

C. RQ2: Can we overcome CoM?

Our goal is to outperform M_{all}, by switching between mono-task models without CoM. From now on, we call the model set by our switching approach as M_{switch}.

Table 4 compares the mAP values of the existing method, M_{all}, and our proposed method, M_{switch}. The upper "All data" row represents the mAP values using the 10,000 BDD validation data. It shows our proposed model M_{switch} achieves mAP scores of 24.15 and 27.74 for YOLOv3 and RetinaNet, to outperform state-of-the-arts with 22.87 and 26.81.

We also categorize the results by object types. The BDD data consists of a total of 10 object categories such as bike and bus. For all object types, M_{switch} showed higher object detection performance than Mall in most cases. In Table 4, the number in parentheses on the right of the object type name means the number of ground-truth objects of the corresponding type. Among the 10 object types, only four types (car, person, traffic light, and traffic sign) contain more than 10,000 objects. In the

Table 4. mAP of our approach, M_{switch}, compared with M_{all} (unit: %)

	YOLOv3		RetinaNet	
	M_{all}	M_{switch}	M_{all}	M_{switch}
All data	22.87	**24.15**	26.81	**27.74**
bike (1,007)	**17.50**	16.45	20.71	**21.78**
bus (1,597)	31.07	**32.68**	17.16	**18.10**
car (102,540)	47.09	**47.63**	64.03	**64.36**
motor (452)	10.37	**13.94**	8.04	**9.32**
person (13,265)	24.49	**26.59**	43.40	**43.62**
rider (649)	14.48	**16.99**	20.58	**24.40**
traffic light (26,891)	21.65	**22.93**	22.12	**22.63**
traffic sign (34,915)	30.57	**31.05**	35.47	**36.24**
train (15)	0.00	0.00	0.00	0.00
truck (4,247)	31.50	**33.27**	36.60	**37.00**

case of "train" object type, the training and validation data was so small that no objects were detected with both M_{all} and M_{switch}. Although Yolov3 showed a different result for the "bike" type, it is considered that the number of objects to be detected is only about 1,000, which is difficult to generalize. For object types that contain a sufficient number of objects, more than 10,000, M_{switch} undoubtedly outperformed M_{all} in mAP values.

Fig. 3 shows qualitative comparisons of M_{switch} and M_{all}, from object-detection results, one taken in late afternoon image, and another at night. For example, for a nighttime image, M_{switch} picks M_{night}, which can identify cars with light reflections in the dark, while Mall fails to do so. Similarly, for daytime, M_{switch}, selecting M_{day}, captures two people, while M_{all} fails to capture the person (on the right) due to darker clothing. This confirms the findings in Table 3, where specialized models for day/time is more effective, such that M_{switch}, by switching between the two, can outperform M_{all} in Table 4.

Fig. 4 shows precision-recall curves of object types—car and traffic signs. We chose these two object types because they contain the largest number of objects in the BDD dataset. Specifically, Fig. 4(a)–4(b) reports the curves for car and traffic signs at nighttime, respectively, and Fig. 4(c)–4(d) for daytime. For both types, we can see clear trends that specialized models, represented in green line trained with night data, perform the best for nighttime instances (Fig. 4(a) and 4(b)), while the orange line performs best for daytime instances (Fig. 4(c) and 4(d)), and outperforms multi-task models. Meanwhile, daytime images have a negative training impact (or, CoM) for nighttime instances, and vice versa, performs the worst. In contrast, M_{switch} in Fig. 4(e)-4(f) can avoid such negative effects

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we validated CoM in automonomous driving, and showed the effectiveness of switching mono-task models.

To further improve effectiveness, we suggest the following research questions as future directions. First, to improve mono-task models, we may distill knowledge between such models, as found effective in multi-

Fig. 3. Qualitative Results for M_{switch} and M_{all}.

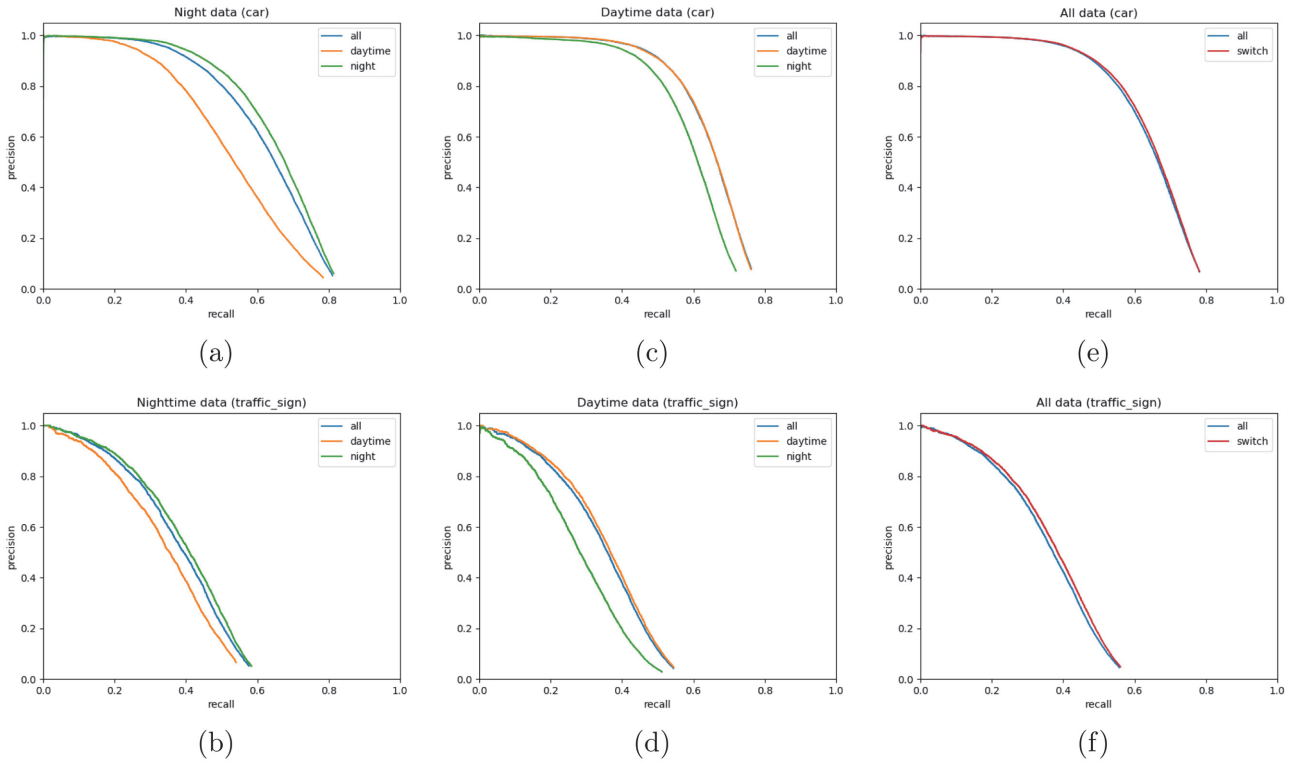

Fig. 4. Precision-recall graph. (a) (b) (c) (d) (e) (f).

linguality [13]. Second, deciding which mono-task model to switch to is trivial in BDD100k dataset, analogous to the setting of PackNet [14] in continual learning, where input-task is labelled. In other datasets, such information can be missing in testing data, for which we can add a classifier to inform on the model to use, as done by StackNet [15] in continual learning. Lastly, we did not study whether dividing into two tasks, daytime and nighttime, is optimal. Deciding how many models to divide, considering also a possibility of transfer learning between models, is a promising future direction.

ACKNOWLEDGMENTS

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00769, Neuromorphic Computing Software Platform for Artificial Intelligence Systems).

REFERENCE

1. L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837-128868, 2019.
2. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91-99, 2015.
3. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 779-788.
4. T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2999-3007.
5. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Computer Vision – ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 21-37.
6. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, et al., "Unsupervised cross-lingual representation learning at scale," 2019 [Online]. Available: https://arxiv.org/abs/1911.02116.
7. P. Rust, J. Pfeiffer, I. Vulic, S. Ruder, and I. Gurevych, "How good is your tokenizer? on the monolingual performance of multilingual language models," 2020 [Online]. Available: https://arxiv.org/abs/2012.15613.
8. F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, et al., "BDD100K: a diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 2633-2642.
9. M. Zhang, Y. Zhang, L. Zhang, C. Liu, and S. Khurshid,

"DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems," in *Proceedings of 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Montpellier, France, 2018, pp. 132-142.
10. V. Ostankovich, R. Yagfarov, M. Rassabin, and S. Gafurov, "Application of cycleGAN-based augmentation for autonomous driving at night," in *Proceedings of 2020 International Conference Nonlinearity, Information and Robotics (NIR)*, Innopolis, Russia, 2020, pp. 1-5.
11. P. Li, X. Liang, D. Jia, and E. P. Xing, "Semantic-aware Grad-GAN for virtual-to-real urban scene adaption," 2018 [Online]. Available: https://arxiv.org/abs/1801.01726.
12. L. Yang, X. Liang, T. Wang, and E. Xing, "Real-to-virtual domain unification for end-to-end autonomous driving," in *Computer Vision - ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 553-570.
13. J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 4825-4829.
14. A. Mallya and S. Lazebnik, "PackNet: adding multiple tasks to a single network by iterative pruning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 7765-7773.
15. J. Kim, J. Kim, and N. Kwak, "StackNet: stacking feature maps for continual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, 2020, pp. 975-982.

Yu-Seung Ma

Yu-Seung Ma received the B.S., M.S., and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1998, 2000, and 2005, respectively. In February 2005, she joined in the Artificial Intelligence Research Laboratory of the Electronics and Telecommunications Research Institute (ETRI), Korea, where she is currently a principal researcher. Her research interests include program testing, software engineering, and machine learning.

Seung-won Hwang

Seung-won Hwang received the Ph.D. degree from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2005. She is currently a professor of computer science and engineering with Seoul National University, Seoul, South Korea. Prior to this, she was a Faculty with POSTECH, Pohang, South Korea and Yonsei University, Seoul, South Korea. Her research interests include language understanding from web-scale data, text, and knowledge.

Hojae Han

Hojae Han is a student pursuing a Ph.D. in Computer Science and Engineering from the Seoul National University, Republic of Korea. He received his B.S. and M.S. degrees in Computer Science from the Yonsei University, Republic of Korea in 2018 and 2021, respectively. His research interests are Natural Language Processing, Code Generation, Training Data Optimization, and Information Retrieval.