Deep Learning Model Protection using Negative Correlation-based Watermarking with Best Embedding Regions

Sayoko Kakikura*, Hyunho Kang**, Keiichi Iwamura*

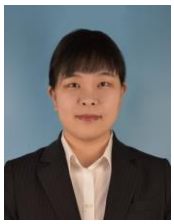
**Department of Electrical Engineering, Tokyo University of Science, Japan*

***Department of Electronic Engineering, National Institute of Technology, Tokyo College, Japan*

kakikura_sayoko@sec.ee.kagu.tus.ac.jp, kang@tokyo-ct.ac.jp, iwamura@ee.kagu.tus.ac.jp

***Abstract*—Deep learning has been used in several fields, such as image classification and data analysis. Training a high-performance model is expensive; thus, its property value is high. Watermarking is a representative technology that provides intellectual property protection for models. In this study, we proposed white-box watermarking using a modified Barni's method (our previous study) for image watermarking. Our method is applicable to pre-trained models because the watermark is embedded in the parameters of the network without training. The proposed method embeds multiple watermarking into neural networks using different keys. We evaluated the method using ResNet-50 trained on CIFAR-10 datasets and confirmed that our watermarking method has high fidelity and robustness against model compression and retraining. The experimental results reveal that our proposed approach can embed up to 10 watermarks with less than 0.1% loss of accuracy. They also indicate the method can completely detect watermarks even after 90% of the parameters are pruned and then transfer learned with CIFAR-100.**

***Keyword*— Copyright protection, deep learning model protection, digital watermarking**



Sayoko Kakikura was born in Japan 1998, received her B.S. degree in field of electrical engineering from Tokyo University of Science, Japan, in 2021. She is currently pursuing the M.S. degree with Tokyo University of Science, Japan. Her main research interests include watermarking and deep learning.

Hyunho Kang was born in Korea, 1973, received his Ph.D. from the University of Electro-Communications, Tokyo, in 2008. He is currently an Associate Professor in the Department of Electronic Engineering at National Institute of Technology, Tokyo College, Japan. His main interests are machine learning, deep learning, information security applications, multimedia security (steganography, digital watermarking), biometric security and physical unclonable functions.

Keiichi Iwamura was born in Japan, 1958, received B.S. and M.S. degrees in Information Engineering from Kyushu University in 1980 and 1982, respectively, and the Ph.D. degree from The University of Tokyo. He is currently a Professor with Tokyo University of Science. His research interests include coding theory, information security, and digital watermarking.