Collusion Resistant Watermarking for Deep Learning Models Protection

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 many 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 providing 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. Additionally, the proposed method embeds multiple watermarking into neural networks using different keys. We evaluated the method using two different networks: 5-layer convolutional neural networks trained on MNIST and ResNet-50 trained on CIFAR-10 datasets. The experimental results show that our proposed approach can embed 10 watermarks with less than 0.1% loss of accuracy, and it detects them completely even after 90% of the parameters are pruned.**

***Keyword*— Copyright protection, Multi-layer neural network, Watermarking, White box watermarking**

Sayoko Kakikura received her Bachelor of Engineering (B.E), in field of electrical engineering, from Tokyo University of Science, Japan, in 2021. She is currently pursuing the M.E degree with Tokyo University of Science, Japan.
Her main research interests include watermarking and deep learning.

Hyunho Kang is currently an Associate Professor in the Department of Electronic Engineering at National Institute of Technology, Tokyo College, Japan; he has held this position since April 2017. He received his Ph.D. from the University of Electro-Communications, Tokyo, in 2008. From 2008 to August 2010, he was a Researcher/Assistant Professor at Chuo University, Tokyo, where he was part of a team that developed Biometric Security technologies. From September 2010 to March 2013, he was an AIST Postdoctoral Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan, where his research work focused mainly on the evaluation of physical unclonable functions. From April 2013 to March 2017, he was an Assistant Professor in the Department of Electrical Engineering at Tokyo University of Science, Japan.
His main interests are machine learning, deep learning, information security applications, multimedia security (steganography, digital watermarking), biometric security and physical unclonable functions.
Dr. Kang is a senior member of Institute of Electronics, Information and Communication Engineers (IEICE) and a member of Information Processing Society of Japan (IPSJ).

Keiichi Iwamura received B.S. and M.S. degrees in Information Engineering from Kyushu University in 1980 and 1982, respectively. During 1982–2006, he was with Canon Inc. He received a Ph.D. from Tokyo University. He is now a Professor at the Tokyo University of Science.
His subjects are coding theory, information security, and digital watermarking.
Dr. Iwamura is a fellow of IEICE and a fellow of IPSJ.