Weight grouping method for generation management of watermarks for white-box DNN models

Ryu Furukawa, Shigeyuki Sakazawa,

***Abstract*— We propose an embedding method of multiple watermarks for derived DNN models to indicate their copyright information over several generations. Our grouping approach of DNN weight parameters used for watermarking can avoid severe interference among multiple watermarks.**

I. Introduction

Research on embedding WM (Watermark) in DNN (Deep Neural Network) models has attracted much attention, and the motivation for this is to protect the copyright of the original developer. Our research motivation is that in legitimately derived DNN models, the copyrights of both the original developer and the legitimate second-generation developer should be protected, and we want to embed watermarks for both. Therefore, it is essential to have a method that avoids serious interference between the original watermark and the second-generation watermark when developing the derived DNN model.

II. Conventional Method

WM embedding in the DNN model was first proposed by Uchida et al [1]. The method extracts the weight parameters of a particular layer in a multi-layer model (white-box setting). The principle of WM Embedding is to have a statistical bias for a particular set of weight parameters by modifying a loss function in the model learning process. The weight parameters are averaged on the output channel axis to increase resistance to node reordering attack [1].

Our previous work [2], derived from the above methods, has proposed a method for embedding multiple WMs in a model by dividing the set of weight parameters used for each generation of WM to reduce interference between WM. However, one problem with the proposed method is the small number of bits that can be embedded.

III. Proposed Method

In this paper, we consider the case of fine-tuning to develop the second/third/fourth generation models and omit the weight averaging in [1]. Then, the number of controllable weight parameters used for embedding can be increased, and thus the number of WM bits also increases. When grouping the weight parameters for each generation, we decide to group them by output channel to reduce interference between generations. The weight parameters for multiple output channels are then bundled as a group for each generation.

IV. Experimental Results

The total number of parameters in the model used in this experiment is 1,250,858, and the number of parameters in the layers used for embedding is 9216. The number of output channels is 32. The output channels are bundled 8 at a time and assigned to 4 generations. The tasks for each generation are as follows: cifar10 for the Gen1 (first generation) model, MNIST for the Gen2 model, Fashion-MNIST for the Gen3 model, and cifar10 again for the Gen4 model. After generating the Gen1 model and embedding the WM, the Gen2 model is derived from the Gen1 model by fine tuning, and the Gen3 and Gen4 models are derived sequentially in the same way. The WM detection results for each generation are shown in the following table.

TABLE I. WM Detection Results

	Gen1 WM	Gen2 WM	Gen3 WM	Gen4 WM
Gen1	ok	–	–	–
Gen2	ok	ok	–	–
Gen3	ok	ok	ok	–
Gen4	ok	ok	ok	ok

From TABLE 1, it was found that four generations of WMs can be embedded by using the weights of 8 channels for each WM embedding. In addition, when using the conventional method [2] to embed 256 bits at a time, an error of 50 bits for Gen1, 55 bits for Gen2, 61 bits for Gen3, and 68 bits for Gen4 occurred, but the new method enables embedding. The following table summarizes the accuracy.

TABLE II. Task Accuracy

	accuracy		accuracy
Gen1 w/o WM	0.7922	Gen3 w/o WM	0.9222
Gen1	0.7888	Gen3	0.9248
Gen2 w/o WM	0.9954	Gen4 w/o WM	0.7933
Gen2	0.9949	Gen4	0.7914

From TABLE 2, there is little impact on the original task in WM embedding. However, depending on the grouping, WM embedding may fail, and it is a future task to investigate the optimal grouping method.

V. Conclusion

The watermark added later does not add significant interference to the watermark of the previous generation. The proposed method successfully embeds four generation of 256-bit WMs.

References

[1] Yusuke Uchida, et al. “Embedding Watermarks into Deep Neural Networks”, Proceedings of the 2017 ACM International Conference on Multimedia Retrieval, pp.269-277, 2017.

[2] Ryu Furukawa, et al. “Generation management of white-box DNN model watermarking”, Proc.GCCE2023, pp.803-804, 2023.

* This work was supported by JSPS KAKENHI Grant Number JP 21K11896.

R. Furukawa and S. Sakazawa are with the Graduate School and Faculty of Information Science and Technology, Osaka Institute of Technology, Osaka 5730196 Japan (e-mail: shigeyuki.sakazawa@oit.ac.jp).