Rank Selection of CP-decomposed Convolutional Layers with Variational Bayesian Matrix Factorization

Marcella Astrid[1], Seung-Ik Lee[1,2], Beom-Su Seo[2]

[1]*University of Science and Technology, Daejeon, South Korea*

[2]*Electronics and Telecommunications Research Institute, Daejeon, South Korea*

marcella.astrid@ust.ac.kr, the_silee@etri.re.kr, bsseo@etri.re.kr

***Abstract*— Convolutional Neural Networks (CNNs) is one of successful method in many areas such as image classification tasks. However, the amount of memory and computational cost needed for CNNs inference obstructs them to run efficiently in mobile devices because of memory and computational ability limitation. One of the method to compress CNNs is compressing the layers iteratively, i.e. by layer-by-layer compression and fine-tuning, with CP-decomposition in convolutional layers. To compress with CP-decomposition, rank selection is important. In the previous approach rank selection that is based on sensitivity of each layer, the average rank of the network was still arbitrarily selected. Additionally, the rank of all layers were decided before whole process of iterative compression, while the rank of a layer can be changed after fine-tuning. Therefore, this paper proposes selecting rank of each layer using Variational Bayesian Matrix Factorization (VBMF) which is more systematic than arbitrary approach. Furthermore, to consider the change of each layer's rank after fine-tuning of previous iteration, the method is applied just before compressing the target layer, i.e. after fine-tuning of the previous iteration. The results show better accuracy while also having more compression rate in AlexNet's convolutional layers compression.**

***Keyword*— Convolutional Neural Networks, Compression, CP-decomposition, Rank Selection, Variational Bayesian Matrix Factorization**

Marcella Astrid received the BS in computer engineering from Multimedia Nusantara University, Tangerang, Indonesia, in 2015, and the MEng in computer software from University of Science and Technology (UST), Daejeon, South Korea, in 2017, and in the same university, is currently working toward PhD degree in computer science. Her recent interests include deep learning and computer vision.

Seung-Ik Lee received his MS and PhD in computer science from Yonsei University, Seoul, South Korea, in 1997 and 2001, respectively. He is currently working for ETRI (Electronics and Telecommunications Research Institute), South Korea. His research interests include machine learning, deep learning, and reinforcement learning..

Beom-Su Seo received his MS in computer science and statistics from University of Seoul, South Korea, in 1998. He is currently working for ETRI (Electronics and Telecommunications Research Institute), South Korea as a project leader on the environment perception for unmanned vehicles and the safety verification and validation for service robots.