Seungcheol Park

Ph.D. Student at Seoul National University

ant6si@snu.ac.kr
(+82)-2-880-7263

Education

Seoul National University Mar. 2019 – *Feb. 2025
- Ph. D. student in Computer Science and Engineering
- Advisor: prof. U Kang

* Expected graduation

Seoul National University Mar. 2011 - Feb. 2019
- B.S. in Department of Naval Architecture and Ocean Engineering

(Including military service Feb. 2013 – Dec. 2014)

Research Interests

Model Compression, Pruning, Quantization, Large Language Models (LLMs)

Research Experience

Enhancing the Efficiency of Large Language Models via Extreme Compression June. 2023 – Present

• Study state-of-the-art compression algorithms for pretrained LLMs such as OPT or LLaMA v1/v2

• Our goal is to design an extreme compression algorithm by finding the best use of state-of-the-arts pruning and quantization techniques without losing unacceptable accuracy

- We are working on integrating Kprune [1], an efficient pruning algorithm, with quantization algorithms such as QLoRA (NeurIPS'23)

• Grant funded by the Youlchon Foundation

Flexible and Efficient Model Compression Method for Various Applications and Environments April. 2020 – Present

• Study compression algorithms for diverse types of models (CNN, Transformer) under diverse environments (full-data, few-shot, and zero-shot)

• For Transformer, we propose Kprune [1], an accurate retraining-free pruning algorithm

- Generate up to 58%p more accurate models after compression than existing retraining-free pruning algorithms
- Requires up to 422 times lower pruning cost than existing retraining-based pruning algorithms

• For CNN, we working on developing FCNN [link] that improves the efficiency of conventional 2D convolution operation by adopting 3D convolution operation for processing 2D images

- We perform additional convolution operations following the dimension of input channels
- The goal of FCNN is to substitute unparallelizable depthwise separable convolutions in state-of-the-art CNN architectures such as EfficientNet v2 or ConvNext v2

• Grant funded by the Korea government (MSIT)

Microsoft MIND News Recommendation Competition July 2020 – Sep. 2020

• The problem is to find the best algorithm that recommends proper news to individual users based on their click history

• **2nd prize award**

- We maximize the accuracy of our model by exploiting a self-attention mechanism to compile various information considering their importance (e.g., creating news embeddings from the

word embeddings in the title, or user embeddings from news embeddings in their history)

Pedestrian Detection Using Sensor Fusion in Indoor Environment Sep 2018. – Feb. 2019

• Study how to detect pedestrians in 3D point clouds obtained by LiDAR
• Propose Curved-Voxel Clustering (CVC), a fast and accurate clustering algorithm for 3D data points that splits 3D space with curved-voxels using spherical coordinate [2]
  - CVC shows 1.7 times faster and 30% more accurate than other clustering algorithms
• Industry-Academy Cooperation with SAMSUNG

Internship in American Bureau of Shipping @ Houston, Texas Jan. 2016 – Feb. 2016

• Research about the operations of floating liquified natural gas (FLNG)
• Sponsored by Offshore Plant Specialization University Project

Teaching Experience

Deep Learning @ SAMSUNG	Dec. 2023
Model Compression @ HYUNDAI	Nov. 2023
Deep Learning @ LG	Feb. 2023
Data Analysis @ IBK	Nov. 2022
Deep Learning @ SAMSUNG	Aug. 2022
Deep Learning @ SAMSUNG	Jul. 2021
Deep Learning @ LG	Feb. 2021

Awards and Honors

Youlchon AI Research Fellowship	Aug. 2023
2nd Prize Award in MIND News Recommendation Competition	Sep. 2020
Google Student Travel Grants Program	Nov. 2019
2nd Place in Edison Challenge	Mar. 2016

• Computational Fluid Dynamics Region
• Chairman's Award (National Research Foundation of Korea)

Publications

[1] Accurate Retraining-free Pruning for Pretrained Encoder-based Language Models, **Seungcheol Park**, Hojun Choi, U Kang, ICLR, 2024 [link]

[2] Curved-Voxel Clustering for Accurate Segmentation of 3D LIDAR Point Clouds with Real Time Performance, **Seungcheol Park**, Shuyu Wang, Hunjung Lim, and U Kang, IROS 2019 [link]

[3] Improved environmental contour methods based on an optimization of hybrid models, Junhwan Choi, Beom-Seon Jang, Jung-Ho Park, Hyeon-Jin Kim, **Seungcheol Park**, Applied Ocean research, vol.91, 2019 [link]

Talks

• Accurate Retraining-free Pruning for Pretrained Encoder-based Language Models @ The 6th Artificial Intelligence Workshop [link] Dec. 2023