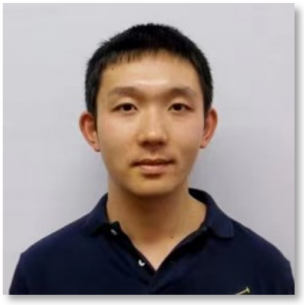
You're Invited to ECE's guest speaker Series

Tuesday, June 8, 2021 | 11:00 AM | Held via Zoom

# Efficient AI Seminar

## Transformer Efficiency: From Model Compression to Training Acceleration



**Dr. Yu Cheng**
Principal Researcher
Microsoft Research

**Abstract** - In recent years, transformer has become ubiquitous in the deep learning field. However, as new generation transformers grow more and more into behemoth size, it becomes increasingly challenging to deploy them in resource-deprived environments. In this talk, I will review our recent work towards transformer efficiency. Specifically, several approaches to compress transformer models, such as pruning, knowledge distillation will be presented. In addition, I will introduce the method to accelerate the training of transformers. Most of these approaches have been deployed in Microsoft products, in both CV and NLP fields.

**Bio:** Yu Cheng is a Principal Researcher at Microsoft Research. Before that, he was a Research Staff Member at IBM Research/MIT-IBM Watson AI Lab. He got a Ph.D. degree from Northwestern University in 2015 and a Bachelor degree from Tsinghua University in 2010. His research focus covers deep learning in general, with specific interests in model compression/ efficiency, deep generative model, and adversarial learning. Currently, he focuses on productionizing these techniques to solve challenging industry problems in computer vision, natural language processing and multimodal learning.