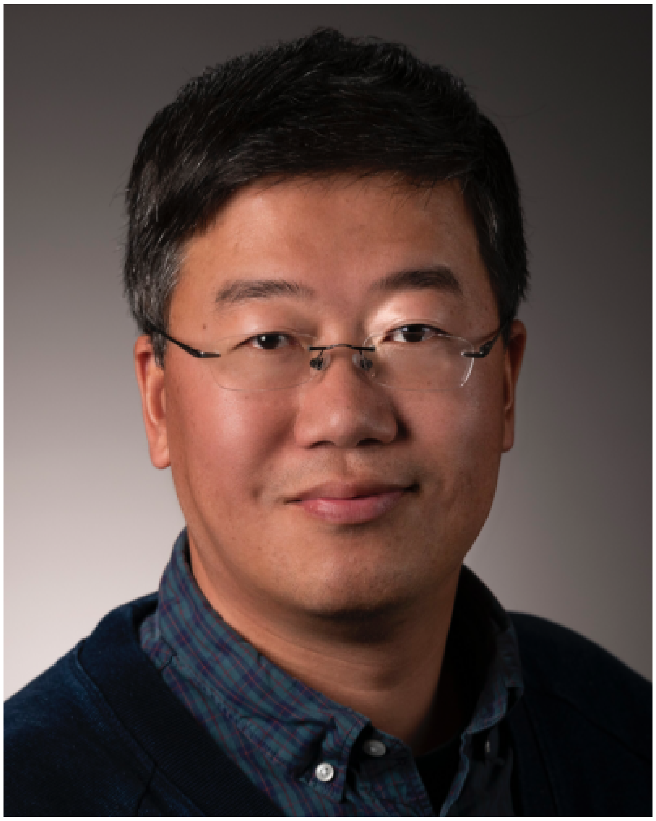
CSE@UCR *colloquium*

Friday, January 9th | 11:00 – 11:50 AM | SSC 329

Adaptive Inference in Large Language Models

Transformer-based large language models (LLMs) have achieved remarkable success, yet many challenges remain. In this talk, I will address a fundamental question: Do all tokens require the same amount of computation within a Transformer? I will share insights into this question and introduce our dynamic layer-skipping and attention-skipping algorithm for adaptive inference in pre-trained LLMs, where different tokens are generated using varying numbers of Transformer layers and attention heads. Our findings show that many layers can be automatically skipped without degrading output quality. These skipped layers reveal a substantial amount of underutilized compute within Transformers, which can be further exploited to enable the generation of multiple tokens using only a subset of layers. We refer to this inference paradigm as Direct Multi-Token Decoding (DMTD). Unlike speculative decoding, our method introduces no additional parameters, no auxiliary routines, and requires no post-generation verification. Despite being trained on a limited dataset, it has demonstrated promising results on a fine-tuned Qwen3-4B model, achieving up to a 2x speedup with only minor performance degradation. Scaling analysis suggests further gains with larger training datasets. At the end of the talk, I will also briefly introduce our other work including agent declarative language and AI4science.



Xifeng Yan

University of California, Santa Barbara

Xifeng Yan is a professor at the University of California, Santa Barbara, where he holds the Venkatesh Narayanamurti Chair in Computer Science. He received his Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign in 2006 and was a research staff member at the IBM T. J. Watson Research Center from 2006 to 2008. His current research focuses on exploring foundation models in artificial intelligence, leveraging these models for knowledge discovery, and developing cross-disciplinary applications. His work has been widely cited and recognized with numerous honors. His team developed the first Transformer-based time series forecasting model, initiating a new research direction in the field.