USC Viterbi School of Engineering

Seminar

Ming Hsieh Department of Electrical and Computer Engineering

General Purpose and Interactive Video Analytics

Francisco Romero

PhD, Electrical Engineering

Stanford University

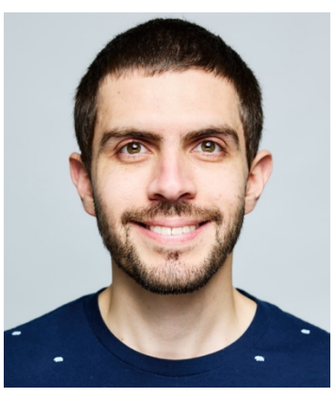
Tuesday, April 2, 2024 | 10:00am PDT | EEB 132

Zoom Meeting ID: 928 8141 1147 | Passcode: 731696

USC NetID login required

Abstract: The availability of vast video datasets and the increasing accuracy of machine learning models have made exploration of video data an exciting opportunity. Asking complex questions like “Find cases where a car takes a left turn while a pedestrian is crossing the road on a rainy night” over terabytes of videos should be possible. Recent video analytics research expects users will manually reason about their query, combine optimizations, and occasionally train models to meet their performance and accuracy goals. This is a long way from the experience users have when exploring structured data.

In this talk, I will present the design of a general purpose and interactive video analytics system. First, I will present how to automatically optimize multi-model, multi-predicate video queries with the VIVA video analytics system. VIVA allows users to express domain knowledge about model relationships. VIVA uses this knowledge to automate complex query optimization by deciding how and when it should be applied. Second, I will present how to efficiently execute video queries across heterogeneous hardware resources with INFaaS. INFaaS exposes a "model-less" interface that enables users to simply specify the performance and accuracy requirements for their applications without needing to specify a specific model-variant for each query. INFaaS efficiently navigates the large trade-off space of model-variants on behalf of users to meet application-specific objectives: (a) for each query, it selects a model, hardware architecture, and model optimizations, (b) it combines VM-level horizontal autoscaling with model-level autoscaling to reduce cost as query load varies. I will also briefly discuss how I extended INFaaS across DAGs of machine learning models with Llama: a serverless video processing framework. I will close by outlining future directions in multi-modal data analysis across heterogeneous hardware resources.



Bio: **Francisco Romero** works at the intersection of computer systems and architecture, databases, and machine learning, where his goal is to design systems that automatically make decisions on users’ behalf to optimize for their goals like cost, performance, accuracy, and resource efficiency. He recently received his PhD in Electrical Engineering at Stanford University, where his research spanned general machine learning inference, serverless computing, data systems, and datacenter scheduling. He has several publications in top-tier conferences, including a best paper at USENIX ATC 2021. His work has been deployed in production Microsoft Azure Functions and is being used for automated video analysis at a stealth company.

Host: Murali Annavaram, annavara@usc.edu