Letter from the Rising Star Award Winner

I am honored to receive the 2022 IEEE TCDE Early Career Award "for contributions to the design of query processing engines for non-volatile memory and video database systems." I am thankful to my nominator, letter writers, and the award committee members, as well as my students, colleagues, collaborators, and sponsors who have helped shape our research agenda. I want to use this opportunity to present an overview of our ongoing work on video database systems and highlight the importance for our community to rethink video database systems.

There is a lot of excitement in industry and academia around building Data Lakes for Big Data. It is important to note that video data will be a significant source of Big Data. For example, it would take a lifetime to watch all the YouTube videos uploaded in a single hour. To extract insights hidden in these videos, it is essential to automatically analyse them at scale. There are several real-world applications of video analytics ranging from climate prediction to urban planning.

Video Database Systems 1.0 Database researchers have long recognized the potential of organizing and querying video databases. QBIC from IBM Research and Chabot from Berkeley were pioneering projects in the video database systems space in the 1990s. These systems were designed around the observation that traditional database systems are tailored for alphanumeric, structured data. To address the limitations of traditional database systems, these systems sought to transform each frame into a numeric feature representation, and then retrieve frames similar to the queried frame based on the distance between the feature vectors. While this technique could return images with similar color or texture, it could often return semantically irrelevant results (e.g., searching for a bluish-white image of a beach might result in a bluish-white image of a living room). This is because the computer vision techniques in the early 2000s could not robustly extract richer semantics from the image – like the fact that the image is that of a beach and not just a generic bluish-white image.

What is New Now? This is no longer a problem, thanks to advances in computer vision over the last decade with the availability of large labeled datasets like ImageNet that are used to train complex neural networks like ResNet. This has led to a resurgence of interest in video database systems 2.0 at several institutions like Georgia Tech, Google, Microsoft, MIT, Stanford, and Washington. Unlike video database systems 1.0, these systems leverage deep learning models to extract richer semantics from each frame, ranging from the locations of different objects inside an image to detecting emotions.

Challenges in Video Database Systems 2.0 However, there are several challenges that these systems must tackle. These include (1) usability, (2) computational cost, (3) accuracy guarantees, and (4) type of queries. First, it is still challenging for a domain expert to set up a query pipeline for video analytics as that requires low-level imperative programming across multiple libraries and frameworks (e.g., OpenCV, PyTorch, Pandas, etc.). So, to find red-colored SUVs in a collection of images, the user needs to write around a hundred lines of Python code requiring expertise in computer vision and systems. Second, deep learning models are expensive to run. Naïvely running a model on every video frame is cost-prohibitive at scale. For example, the estimated cost for processing one month's video from a camera using Google's Cloud Vision API amounts to $225K and 350$ K for the image classification and the object localization tasks, respectively. The cost increases with the complexity of the vision task.

Third, unlike traditional database systems, accuracy is neither guaranteed nor expected in video database systems. This is because the model may return wrong inference results due to incorrect training data. Errors accumulate in complex queries that rely on multiple deep learning models. Furthermore, users themselves may require different accuracy targets. For example, an officer doing a license plate search requires higher accuracy than an urban planner analyzing traffic flow patterns. Fourth, users are interested in issuing semantically richer queries. Most work in our community has focused on only detecting "objects" of interest. However, users might be interested in querying for "actions" of interest. It is not sufficient to look at each frame independently to answer such action queries. For example, the model requires context across a sequence of frames to distinguish between a left and a right turn of a car.

EVA Database System I will next discuss how we tackle some of these challenges in the EVA database system

we are developing at Georgia Tech. EVA is an end-to-end database system tailored for videos [https://georgia-tech-db.github.io/eva/index.html]. We are redesigning all the layers of the system, starting from the storage manager up to the query optimizer. EVA supports a declarative SQL-like language with video-specific functionality to address the usability challenge. This allows domain experts to easily issue video analytics queries without requiring low-level imperative programming. Furthermore, they can reuse the query language across different applications. The user only needs to write a short SQL query to find red-colored SUVs in a collection of images. To lower the computational cost, EVA optimizes declarative queries using a Cascades-style optimizer. The optimizer is tailored for user-defined functions that wrap around deep learning models. It automatically materializes and reuses the results of expensive user-defined functions (i.e., models). EVA uses symbolic analysis of predicates to identify opportunities for reusing results. Its optimizer uses a reuse-aware cost function to guide decisions like predicate reordering and model selection.

Third, EVA leverages the flexibility in accuracy targets to reduce query processing time. In particular, it selects the appropriate "physical" deep learning model for a given "logical" vision task (e.g., object detection) to meet the accuracy requirement (e.g., YOLOv4 vs. Faster-RCNN). It processes different chunks of a given video using different models in the ensemble to lower query processing time while meeting the accuracy requirement. Lastly, EVA supports complex action-based queries. Action localization task requires a more expensive neural network than object detection. So, the query optimizer trains a reinforcement learning agent to adaptively pick video chunks that are then fed to the action classifier. This is an exciting combination of Data Management for ML and ML for Data Management. The agent learns to tune three knobs: sampling rate, chunk length, and resolution, for adaptively constructing the next video chunk.

In conclusion, I hope this letter got you interested in exploring the space of video database systems. Our work, recognized by this award, is but one part of this tidal wave. It will take our entire community to unlock the full potential of video database systems for practitioners. We can make a significant difference in several open problems – data cleaning, multi-modal query processing, fairness, holistic query optimization, multi-tier storage management, provenance, etc. I hope we will rise to the occasion and help develop the next generation of video database systems.

Joy Arulraj
Georgia Tech, USA