Accelerating Video Analytics

Joy Arulraj
Georgia Institute of Technology

MOTIVATION. The advent of inexpensive, high-quality cameras has led to a rapid increase in the volume of generated video data [19, 16]. It is now feasible to automatically analyze these video datasets at scale due to two developments over the last decade. First, researchers have designed complex, computationally-intensive deep learning (DL) models that capture the contents of a given set of video frames (*e.g.*, objects present in a particular frame [11]) [15]. Second, the computational capabilities of hardware accelerators for evaluating these DL models have increased over the last decade (*e.g.*, TPUs) [8]. We anticipate that automated analysis of videos will reduce the labor cost of analyzing video datasets in a wide range of important applications [14].

BACKGROUND. Motivated by these developments, researchers have recently proposed several novel video database management systems (VDBMSs) [2, 1, 9, 21, 4]. These systems accelerate declarative queries over videos using techniques like training a lightweight, specialized model to filter out irrelevant frames [12], or sampling a subset of important frames [10, 3]. The queries they support primarily focus on detecting objects of interest (*e.g.*, searching for frames containing atleast two cars in a surveillance video). To accelerate this query, the VDBMS may train a lightweight model to quickly filter out irrelevant frames that are unlikely to contain cars [12]. By reducing the number of invocations of the heavyweight oracle model (*i.e.*, the more accurate DL model specified by the user [12, 5]), the VDBMS speeds up the query with a tolerable drop in query accuracy.

CHALLENGES. State-of-the-art VDBMSs suffer from two limitations that constrain their utility and computational efficiency. First, these systems primarily focus on accelerating object detection queries over videos. So, they are not able to support queries associated with more complex vision tasks. For example, an important class of video analytics queries focuses on detecting and localizing *actions* – events spread across a sequence of frames (*e.g.*, "right-turn of a car") [17, 20, 6]. It is difficult to process such queries due to two reasons. First, current VDBMSs operate on individual frames (either using the lightweight filter or the heavyweight object detector). To detect an action, the VDBMS would need to identify features that span across multiple frames. Second, inference times of DL models tailored for action detection are higher than that of object detectors.

Another limitation is that it is computationally expensive for the VDBMS to train filters for each unique combination of: video content, oracle model, and predicate of interest. First, filters depend on video content (*e.g.*, day- vs night-time videos [18]). Second, the labels associated with the training frames are obtained using a specific model (*e.g.*, SSD [11]). Third, due to the limited capacity of filters, they are tailored for a specific predicate (*e.g.*, COUNT(CAR) > 2 [13]). These constraints increase the overall training cost associated with filters.

IDEAS. To tackle the first challenge, we will need to design novel algorithms for efficiently processing action queries. For instance, we could train a DL-based agent to quickly skim through video segments that are unlikely to contain the target action [7]. The agent would quickly generate proxy features of a given video segment and use them to choose the next video segment to process (*e.g.*, picking the resolution of the frames, the sampling frequency, *e.t.c.*) from a large space of possible segments. We anticipate that such task-specific optimizations will need to be developed for other vision tasks [4].

For the second challenge, it is important to develop unsupervised algorithms for sampling representative frames from a video. This will allow the VDBMS to answer ad-hoc queries using these representative frames instead of training a filter tailored for a specific predicate or oracle model. It is critical to obtain theoretical bounds on the likelihood of the representative frames satisfying the query accuracy constraint.

SUMMARY. An amalgamation of ideas in database systems, computer vision, and machine learning will help realize the vision of accelerating video analytics. We anticipate that VDBMSs will become more common in the future, and hence, the optimizations developed by the database community will become important.

REFERENCES

[1] BlazeIt. https://github.com/stanford-futuredata/blazeit.

[2] EVA. https://github.com/georgia-tech-db/Eva.

[3] J. Bang, P. Chunduri, and J. Arulraj. Eko: Adaptive sampling of compressed video data. *arXiv preprint arXiv:2104.01671*, 2021.

[4] F. Bastani, S. He, A. Balasingam, K. Gopalakrishnan, M. Alizadeh, H. Balakrishnan, M. Cafarella, T. Kraska, and S. Madden. Miris: Fast object track queries in video. In *SIGMOD*, pages 1907–1921, 2020.

[5] J. Cao, R. Hadidi, J. Arulraj, and H. Kim. Thia: Accelerating video analytics using early inference and fine-grained query planning. *arXiv preprint arXiv:2102.08481*, 2021.

[6] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, pages 1130–1139, 2018.

[7] P. Chunduri, J. Bang, Y. Lu, and J. Arulraj. Zeus: Efficiently localizing actions in videos using reinforcement learning. *arXiv preprint arXiv:2104.06142*, 2021.

[8] J. Dean, D. Patterson, and C. Young. A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro*, 38(2):21–29, 2018.

[9] B. Haynes, M. Daum, A. Mazumdar, M. Balazinska, A. Cheung, and L. Ceze. Visualworlddb: A dbms for the visual world. In *CIDR*, 2020.

[10] D. Kang, J. Emmons, F. Abuzaid, P. Bailis, and M. Zaharia. Noscope: Optimizing neural network queries over video at scale. *arXiv: Databases*, 2017.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[12] Y. Lu, A. Chowdhery, S. Kandula, and S. Chaudhuri. Accelerating machine learning inference with probabilistic predicates. In *SIGMOD*, pages 1493–1508, 2018.

[13] Y. Lu, A. Chowdhery, S. Kandula, and S. Chaudhuri. Accelerating machine learning inference with probabilistic predicates. In *SIGMOD*, pages 1493–1508, 2018.

[14] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, and A. Waldman-Brown. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[16] T. J. Sejnowski. *The deep learning revolution*. MIT press, 2018.

[17] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, pages 1049–1058, 2016.

[18] A. Suprem, J. Arulraj, C. Pu, and J. Ferreira. Odin: Automated drift detection and recovery in video analytics. *VLDB*, 13(11), 2020.

[19] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[20] H. Xia and Y. Zhan. A survey on temporal action localization. *IEEE Access*, 8:70477–70487, 2020.

[21] Y. Zhang and A. Kumar. Panorama: a data system for unbounded vocabulary querying over video. *VLDB*, 13(4):477–491, 2019.