УДК 811.111:005.5

Folynskov D., Folynskova E., Khomenko S.

# Improving the Efficiency of the Video Analytics Algorithm

Belarusian National Technical University
Minsk, Belarus

One of the biggest barriers to video analytics is having sufficient computational resources. An alternative to increasing hardware performance for video analytics is improving algorithm efficiency.

Efficiency allows you to run more accurate and powerful analytics with less hardware and fewer constraints. Often, video surveillance devices have constraints on size, power consumption, hardware, etc. that demand either greater efficiency or suffering worse performance.

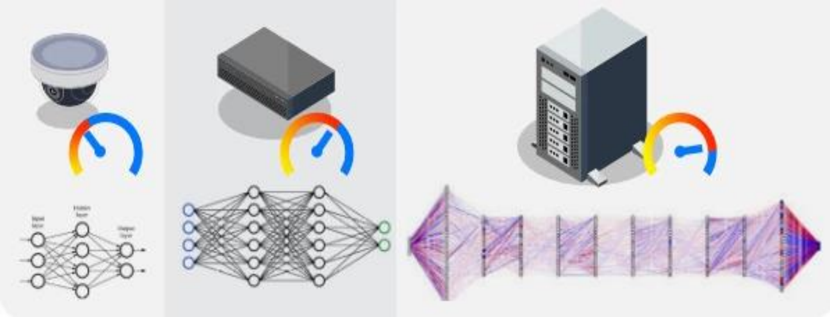


Fig. 1 – Efficient Algorithms Decrease Processing Power

The difference in processing load, for a given accuracy, can be a 10X or 100X. If a developer does not have efficient algorithms, this typically results in less sophisticated algorithms and degraded accuracy. This image in Figure 1 shows how efficient algorithms affect processing power.

Video surveillance is often done in real-time of surveillance means processing needs to happen quickly. With the average frame rate for surveillance recording at ~15fps, even processing analytics at a fraction of that (e.g., 5fps) would require a max of 200 milliseconds processing. But if processed on too few frames, running subjects may be missed. As such, an algorithm's accuracy decreases if it is missing objects (false negatives) because it is processing frames too slowly. Video analytic algorithm efficiency is commonly defined in frames per second (FPS) processed.

While strong performing deep learning, algorithms are more accurate than heuristics and machine learning methods, they generally have a higher computational cost processing the same number of frames per second.

Most video surveillance deep learning algorithms use convolutional neural networks (CNNs) because they are accurate at detecting or classifying features/objects in images.

As such, the goal of most algorithm developers is to continually increase CNN efficiency while maintaining a high level of accuracy. This is commonly achieved by:

- Using specific categories of CNNs
- Training CNNs to detect a limited number of object types Decreasing CNN size.

Network pruning and quantization are 2 methods AI analytics developers use to improve efficiency. Pruning eliminates nodes or edges from the neural network, which makes the analytic run faster. Ideally, the removed edges and nodes are redundant and do not improve accuracy, but if important nodes or edges are removed, accuracy will drop. Graphically, the network pruning process is shown in Figure 2.

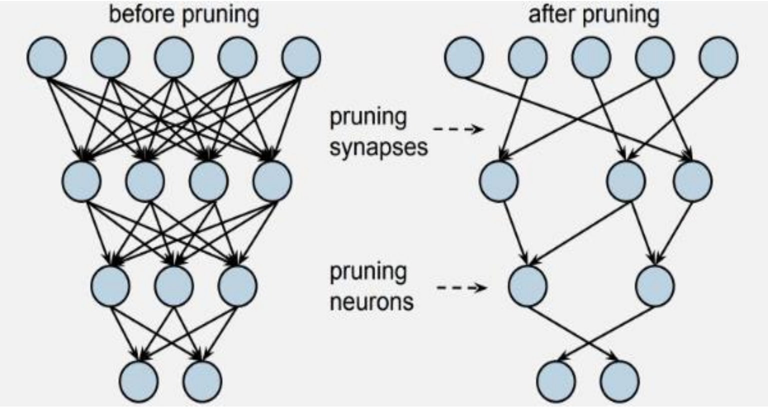
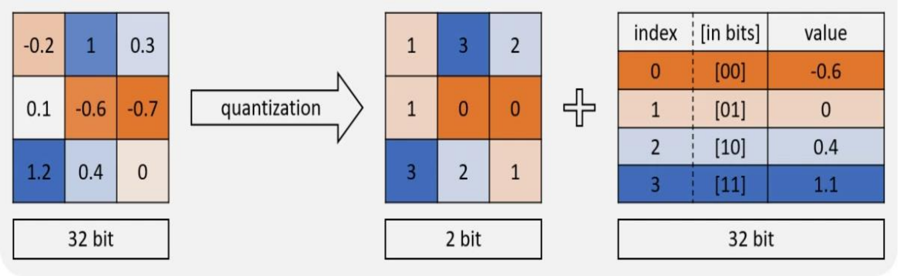


Fig. 2 – Network Pruning

Quantization compresses the neural network by round large numbers with decimals to smaller integer representations. The rounding is designed to keep most of the information of the original number while reducing the size, keeping accuracy high. This simplifies the calculations that are required, decreasing the memory required and improving speed. Graphically, the quantization process is shown in Figure 3.



| | | |
|---|---|---|
| -0.2 | 1 | 0.3 |
| 0.1 | -0.6 | -0.7 |
| 1.2 | 0.4 | 0 |

32 bit

| | | |
|---|---|---|
| 1 | 3 | 2 |
| 1 | 0 | 0 |
| 3 | 2 | 1 |

2 bit

| index | [in bits] | value |
|---|---|---|
| 0 | [00] | -0.6 |
| 1 | [01] | 0 |
| 2 | [10] | 0.4 |
| 3 | [11] | 1.1 |

32 bit

Fig. 3 – Quantization

However, pruning and quantization can both decrease accuracy and require additional training/development resources.There is an art to getting the best results, and trial-and-error is typically used during AI development to avoid over-pruning or over-quantization, which results in low processing and memory requirements, but compromises accuracy.

References:

1. Kuntai Du, Ahsan Pervaiz, Xin Yuan, Aakanksha Chowdhery, Qizheng Zhang, Henry Hoffmann, and Junchen Jiang. Server-Driven Video Streaming for Deep Learning Inference. – Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM). – 2020. – p. 557–570.
2. Axis. AI in video analytics. – 2021. – p. 2–16,
3. Video Analytics Algorithms / Efficiency [Electronic resource]. – Mode of access: https://ipvm.com/reports/analytics-eff. – Date of access: 01.03.2021.