Problem statement

The problem statement of the project is clear to me, but some terminology is not so obvious when you see this for the first time. Also, interesting how significant this problem is in real life. For example, what is the inference time of current models? Is it significant and unacceptable for real life?

Methods

The part related to the methods of change point detection and anomaly detection shows the models used to solve this problem, but I think it is a good idea to show where this model could be used.

Related works

I do not understand the part about video and image compression. In my opinion, such a technique could reduce the size of the data. However, for applying deep neural networks, we need to decompress the image into the raw format, and this operation potentially could increase the inference time.

To obtain low dimensionality image representation, we also need to apply a convolution neural network and apply the KL-CPD method. How could this technique improve the efficiency of solving this problem?

Pruning methods are based on the assumption that some part of the weight equals zero. The next step is to make matrix multiplication so that we do not use these zeroes in computations. It is probably reasonable to add spare matrix storage and multiplications methods references.

Ideas

Probably one of the most straightforward ideas is to drop some frames. It looks like 30 frames per second probably are too much for solving the problem.

There are several methods to improve the efficiency of neural networks by reducing the floating-point accuracy – mixed precision or quantization. Also, hardware frameworks to reduce computation complexity like TensorRT and e.t.c.