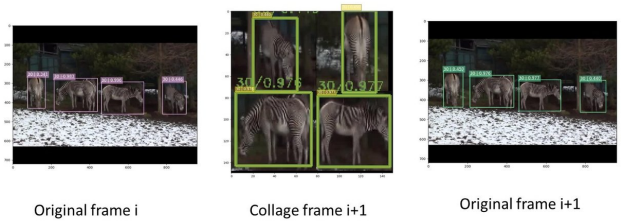
Fast object detection in videos using region-of-interest packing

19 September 2018, by Ingrid Fadelli



Sample of consecutive frames processed with the ROI packing mechanism. Credit: Athindran et al.

Researchers at the Robert Bosch Center for Data Science and Artificial Intelligence and Center for Computational Brain Research, Indian Institute of Technology Madras, and Purdue University have recently developed a new method of reducing computational requirements for object detection in videos using neural networks. Their technique, called Pack and Detect (PaD), was outlined in a paper pre-published on arXiv.

Object detection is a key aspect of many computer vision applications, such as object tracking, video summarization, and video search. While recent advances in machine learning have led to the development of increasingly accurate tools for completing this task, existing methods are still computationally very intensive. For instance, processing a video at 300 x 300 resolution using the SSD300 object detection network, with VGG16 as backbone and at 30 fps requires 1.87 trillion floating point operations per second (FLOPS).

The researchers observed that in some cases, however, most regions in a video frame are merely background, with salient objects occupying only a small fraction of the area in the frame. In addition, they found that there is a strong temporal correlation between consecutive frames. They leveraged these observations and proposed a new technique for object detection in videos that could reduce computational requirements for object detection tasks.

"We were inspired by the foveal mechanism in both biological and artificial vision systems," Athindran Ramesh Kumar, one of the researchers who carried out the study, told TechXplore. "Previous efforts pertaining to the foveal attention mechanisms in artificial vision systems focus on only one region in the image or on one object at a time. We wondered how a vision system would be if it could focus on all salient regions in the scene at once."

The object detection method devised by the researchers is hence inspired by biological vision systems. However, contrary to previous attempts, their system packs all the regions of interest together in a single frame, instead of processing them sequentially.

"The objective of our work was to speed-up object detection in videos by focusing only on the salient regions in the frame and eliminating the background clutter," Balaraman Ravindran, another researcher who carried out the study, told TechXplore. "For eliminating background clutter, we exploited the temporal correlation between adjacent frames in a video. This is a property that video compression techniques use to reduce the storage and bandwidth requirements; we use it to speed up computation."

PaD, the object detection method proposed by Ravindran and his colleagues works by processing frames at regular intervals in full size. These frames are referred to as "anchor frames." In all other frames, on the other hand, the tool identifies regions of interest based on the location in which objects were situated in the previous frame.

"These regions of interest are arranged together like in a collage, which is used as input for the

object detector," Anand Raghunathan, one of the researchers that carried out the study, told TechXplore. "The detections are then mapped back to the locations in the original image. This method is faster because the collage images are of smaller size than the full frames. We leverage the flexibility of popular object detectors such as SSD300 to process images at both full size and smaller sizes."

The researchers evaluated their method on the ImageNet VID dataset and found that it sped up times by 1.25x, with less than a 1.6 percent drop in accuracy. In addition, they observed that the time taken to process lower-sized frames was almost three times lower, with the FLOP count reduced by four times.

In addition, their study highlighted two important aspects that could inform the development of faster and less computationally intensive methods of detecting objects in videos. First, objects of interest generally only occupy a small fraction of pixels in a frame; second, there is a correlation between adjacent frames in a video.

"Our work can help make video analytics possible on resource-constrained devices at the edge of the Internet of Things by reducing computational requirements, or may improve the number of video streams that may be processed by a server in the cloud," Athindran said.

The study carried out by this team of researchers is an initial step toward the development of more effective object detection tools. They are now planning further investigations that could improve their method further.

For instance, currently, PaD selects anchor frames at regular intervals, yet the researchers could develop a mechanism that dynamically identifies these key frames. They also plan to test their technique in more resource-constrained hardware, such as smartphones, wearable devices and smart home appliances.

"We handcrafted an algorithm to infer the regions of interest and form a collage image," Ravindran said. "But a fully neural system would have neural networks that generate the collage image based on the previous frame. This is a more ambitious line of future work."

More information: Pack and Detect: Fast object detection in videos using region-of-interest packing. arXiv:1809.01701v1 [cs.CV]. arxiv.org/abs/1809.01701