Work-in-Progress: Temporal RegionDrop – Frame Difference Sparsity for Efficient Video Inference

Youki Sada*, Seiya Shibata†, Yuki Kobayashi‡ and Takashi Takenaka§
Secure System Platform Research Laboratories
NEC Corporation
Kanagawa, Japan
Email: *youkis@nec.com, †s-shibata@nec.com, ‡y-kobayashi.hq@nec.com, §takashi.t@nec.com

***Abstract*—Convolutional Neural Networks (CNNs) require costly hardware for real-time processing due to massive multiplications. In video-based inference, many pixels remain unchanged between frames. By skipping these operations and reusing previous results, CNNs can be accelerated without accuracy loss. However, efficient implementation of sparse operations is challenging due to the need for change detection and sparse operations in each layer. To address this, we propose Temporal RegionDrop, which performs change detection once per frame and uses a common mask to reduce conversion costs between sparse and dense formats. This method is applicable to all CNNs without retraining. We evaluate our RegionDrop on AGX Orin GPU with ChangeDetectionNet, MOT15, and JHMDB datasets. RegionDrop achieves a 2.0×–6.5× speedup with minimal accuracy loss.**

***Index Terms*—neural networks, sparsity, efficient inference.**

I. Introduction

Applications using CNN (Convolutional Neural Networks) are rapidly growing in demand in surveillance, human computer interaction, etc. However, these applications require high-performance machines and costly environments to run CNNs at practical speed of tens of billions of floating-point operations (FLOPs). Research in the last few years has focused on temporal sparsity. In fixed cameras, the similarity between frames is high, and the computation of unchanged areas can be skipped by reusing the computation results of previous frames. In this paper, we propose Temporal RegionDrop to reduce the computational complexity by hardware-aware activation skipping focusing on sparsity due to frame difference.

II. Temporal RegionDrop

We propose Temporal RegionDrop that removes temporal redundancy on activation without retraining, as shown in Fig. 1. Temporal RegionDrop utilizes sparsity in all layers including convolution, pooling, activation, and upsampling. We also propose a lightweight method that minimizes the execution time for additional sparse address calculation and mask update by using a common mask for all sparse layers. Our scheme effectively accelerate CNNs for both single-camera and multi-camera case.

We provide an implementation that can be efficiently executed in an end-to-end sparse data representation at all layers from the input to the output layer. In addition, we compared with our implementation with latest cuDNN matrix multiplication algorithm.

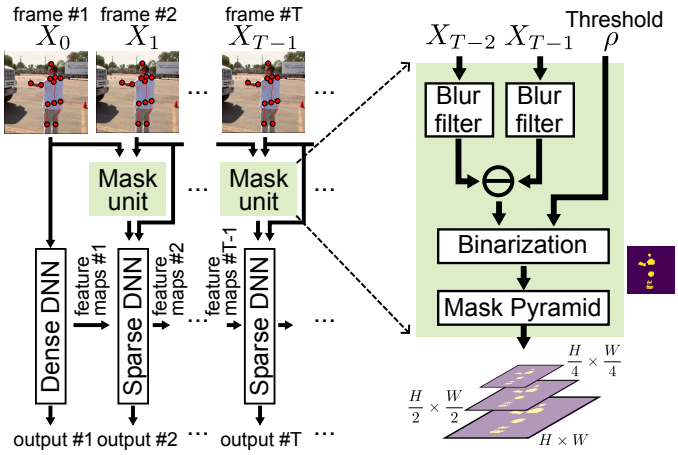

Fig. 1. Temporal RegionDrop

III. Results

Firstly, we evaluate Temporal RegionDrop for person detection [1] and YOLOv5x model. RegionDrop achieves 76.1% mAP and 419.1 FPS on AGX Orin GPU, and achieved 0.9 point of negligible accuracy degradation and a factor of 6.5 improvement in throughput compared to baseline. In object tracking [2], RegionDrop using ByteTrack achieved 4.3 times faster speed than the non-sparse one without accuracy loss. It realized real-time DNN inference with high resolution camera on edge device. Moreover, we evaluate RegionDrop on human pose estimation [3], and RegionDrop achieved 1.9 times faster speed with negligible accuracy loss. The speedup ratio was same as costly prior art [4] that requires retraining. Finally, RegionDrop realized simple and efficient frame difference computation and significant speedup on many applications not only for simple less-object scenes but also for low-resolution camera or crowded scenes.

References

[1] N. Goyette *et al.*, "Changedetection. net: A new change detection benchmark dataset," in *CVPR2012 workshops*. IEEE, 2012, pp. 1–8.
[2] L. Leal-Taixé *et al.*, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv preprint arXiv:1504.01942*, 2015.
[3] H. Jhuang *et al.*, "Towards understanding action recognition," in *ICCV*, Dec. 2013, pp. 3192–3199.
[4] A. Habibian *et al.*, "Skip-convolutions for efficient video processing," in *CVPR2021*, pp. 2695–2704.

2643-1726/24/$31.00 ©2024 IEEE
DOI 10.1109/CASES60062.2024.00012