Deep Learning Image Analysis System on Embedded Platform

Hyok Song, In-kyu Choi, Min-soo Ko, *Jisang Yoo
*Korean electronics Technology Institute, *Kwangwoon Univ.*
*{hsong, cig2982, kmsqwet}@keti.re.kr, *jsyoo@kw.ac.kr*

Abstract

Recent advances in the field of deep learning technologies have made it possible to develop practical video analysis systems with embedded platform that are more accurate and faster than prior embedded systems based on pattern analysis technology. In video analysis applications, object detection, face recognition, action recognition and super resolution technology is the most important functions. In this paper, we show action recognition and super resolution on embedded deep learning system. It is discovered that ResNet34 structure with 16 frames analysis is most profitable for speed and accuracy and Attention based super resolution method is enough for real-time processing. Both deep learning models optimized for speed and accuracy are operated on embedded system with Intel NPU at real-time. We introduce main technologies in chapter 1. Proposed method is shown in chapter 2 and the result is shown in chapter 3.

Keywords: Embedded system, Action recognition, Super resolution. Deep learning.

1. Introduction

Human action recognition in the real-world environment finds plenty of applications including intelligent video analysis system such as surveillance system and recent advances in the field of neural processing unit(NPU) made it possible to develop embedded video surveillance systems with video analysis function that are fast enough[1]. For applications of vision-based surveillance, it is important to early detect the occurrences of events such as human detection or abnormal actions in restricted area. Therefore, human detection and action recognition on surveillance system are the most important functions for past decades.

There are major two methods for detecting actions which are two stream framework and 3D primitive framework[2,3]. On our embedded platform, our relevant task is creating a solution that can achieve high accuracy and providing a fast inference speed both. We modify a transfer network model based on a lightweight architecture which can run in real-time on a small NPU platform. UCF101 dataset is used for learning and verifying this accuracy. The dataset consists of predicting the presence or absence of each of the 101 action classes shown in Fig 1 [4].

Fig 1. UCF101 dataset examples

Most surveillance images captured from CCTV are low resolution images or cropped images are small and coarse to identify an individual or classify objects for deep learning systems[5]. Super-resolution (SR) is a technique that can overcome this limitation to produce high resolution images of a subject. Typical Super resolution methods are using filters like interpolation filters or wavelets. Trained deep learning method was shown using upscaled bicubic interpolation[6]. We used single image super-resolution(SISR) for high frequency information lost compensation [7,8]. However, SISR also has a difficulty in the recovery of high frequency details which shows Mean squared error loss. This comes from SISR is performed in high resolution space only using deep learning instead of interpolation. The problem is solved by adapting residual building blocks[9]. Attention mechanism is first used in speech recognition and have used image generation[10]. This mechanism locates the tiny textures and solved tiny texture recovery problem.

2. Proposed method

We adopt encoder and decoder models to create a sequence to sequence system to identify actions and this use ResNet34 Transfer encoder shown in Fig 2

[11]. The encoder converts each frame of the input sequence into an embedding vector using CNN, and the decoder classifies the action label of the given input clips by collecting temporal information between frames using the multi-head attention module. The test dataset UCF 101 shows that the location of objects which make actions is usually on the center of the images. We focus on the location of the main object for action recognition models shown in Fig. 3 [12,13]. Action recognition based on attention analysis can infer the actions potentially in videos by focusing only on the relevant places in each frame.

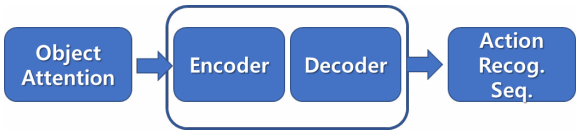

Fig 2. Seq2Seq structure

Fig 3. Object attention

SISR model has inpainting problem and also our model has the difficulty in the recovery of tiny textures. This can be solved using high frequency information or edge data extracted from the input image. Our base model also adapted attention based method with SISR model. We extracted grayscale edge information from the low resolution input image and adapt to attention mask from attention producing network additionally. The mask and edge information work as a feature enhancing high frequency. It also gives enhancement edge areas and tiny texture areas and removes smoothness.

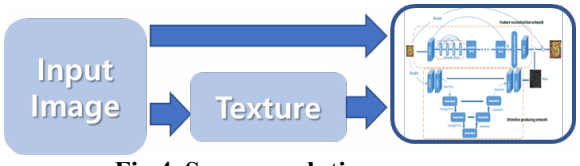

Fig 4. Super resolution process

OpenVINO is an open-source toolkit for optimizing and deploying deep learning models[14]. It optimizes deep learning models for vision, audio and so on from Pytorch and TensorFlow. Our action recognition model and super resolution model is converted to OpenVINO intermediate representation (IR) for model optimization including quantization, pruning, etc and data format managing. The process of OpenVINO is shown in Fig 5.

Fig 5. OpenVINO IR formatting

3. Experimental results

We evaluate the inference time which is one of the important issues for embedded platforms. Our embedded system uses Intel NPU, we use OpenVINO toolkit for optimizing the model and Pytorch framework.

Table 1: Inference time of AR

ResNet-34-VTN-RGB	56 FPS
Stacked RGB+RGBDiff ResNet-34 VTN	51 FPS
ResNet-50-VTN-RGB	49 FPS
Ours	80 FPS

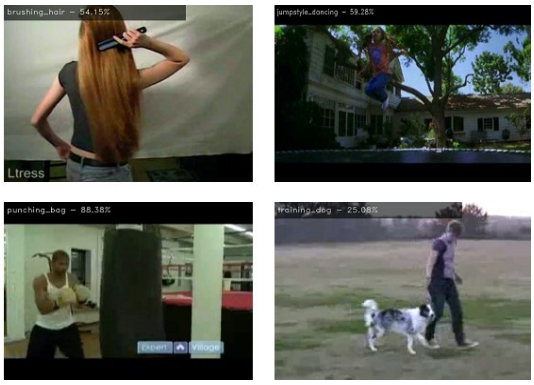

Fig 6. Action recognition results

Table 1 shows the inference time that employ the proposed and similar approaches and fig 6 shows the results. Our method shows 80.98 FPS which is faster than real-time speed promising for edge computing.

We also evaluate the inference time of super resolution on our embedded platform using OpenVINO toolkit for optimizing the model and Pytorch framework. Fig 7 shows super resolution result inferred on the embedded platform. The inference time is 54 msec.

Fig 7. Super resolution results

Table 2: Inference time of SR

Ours	54 msec

4. Future works

We evaluate the inference time which is one of the important issues for Intel NPU platforms. Real-time processing implementation on the platform is satisfied. Our next goal is re-identification and de-identification implementation on the platform for stand alone embedded platform.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-00804, Media production technology using learning based directing methods)

References

[1] Amrutha, C. V., C. Jyotsna, and J. Amudha. "Deep learning approach for suspicious activity detection from surveillance video." 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2020.

[2] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pages 568–576, 2014.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015.

[4] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402 , 2012.

[5] Lin, Frank, et al. "Investigation into optical flow super-resolution for surveillance applications." WDIC 2005: APRS Workshop on Digital Image Computing: Workshop Proceedings. University of QLD, 2005.

[6] C. Dong, C. C. Loy, K. He et al., “Learning a deep convolutional network for image super-resolution,” in ECCV, 2014, pp. 184–199.

[7] Pesavento, Marco, Marco Volino, and Adrian Hilton. "Attention-based multi-reference learning for image super-resolution." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[8] Liu, Yuan, et al. "An attention-based approach for single image super resolution." 2018 24Th international conference on pattern recognition (ICPR). IEEE, 2018.

[9] K. He, X. Zhang, S. Ren et al., “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[10] K. Gregor, I. Danihelka, A. Graves et al., “Draw: a recurrent neural network for image generation,” Computer Science, pp. 1462–1471, 2015.

[11] https://github.com/openvinotoolkit

[12] Sharma, Shikhar, Ryan Kiros, and Ruslan Salakhutdinov. "Action recognition using visual attention." arXiv preprint arXiv:1511.04119 (2015).

[13] Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov. Exploiting image-trained CNN architectures for unconstrained video classification. CoRR, abs/1503.04144, 2015.

[14] OpenVINO Toolkit. https://software.intel.com/en-us/openvino-toolkit.