Human Action Classification through Image Vectorization and Self-supervised Learning

Seo-El Lee
Department of Public Safety Bigdata
Kyonggi University, Korea

Do-Eun Choe
Department of Civil Engineering
New Mexico State University
NM, USA

Kyungyong Chung
Division of AI Computer Science and Engineering
Kyonggi University, Korea

ABSTRACT

An accurate and reliable deep-learning model requires a large amount of labeled data, while, in practice, there are limitations in collecting labeled data of our particular interest. The proposed method first detects objects within images through the You Only Look Once method and removes backgrounds based on the bounding boxes targeting human objects. As a next step, image vectorization is performed using a CNN encoder pre-trained through the Kinetics dataset. It classifies human behaviors within images through self-supervised learning to resolve the labeled data insufficiency problem and enhances object detection.

1 Introduction

You Only Look Once (YOLO) is one of the most popular deep learning technologies used for object detection/recognization. While this supervised learning trained by labeled data can be used to create a model, it requires a large amount of labeled data. In the case where the amount or quality of labeled data is insufficient, it becomes difficult to accurately learn a model. In addition, the traditional method may result in a model that only detects the features of our interest from the data similar to the learned data, instead of learning the generalized features applicable to unseen data. To resolve this problem, in this research, a novel architecture is proposed with steps of the followings: (1) detect objects using YOLO; (2) improve the accuracy by applying a pretext task; & (3) perform auto-labeling through the clustering technology. The proposed architecture enables us to utilize a vast amount of non-labeled data to enhance object detection and recognition performance. This overcomes the limitations of YOLO which requires a large number of labeled data. The self-supervised classification can be performed through the clustering technique.

2 Human Action Classification through Image Vectorization and Self-supervised Learning

YOLO is a deep-learning model for object detection that can be used for image background removal. It uses CNN to detect and classify objects within images and generates bounding boxes around the detected objects. In this study, a YOLO model is used to remove all regions other than the persons detected in a video. The Kinetics dataset is a large-scale video dataset used to pre-train the CNN encoder for image vectorization. The Kinetics dataset pre-trains the CNN encoder, and this enables the encoder to learn to recognize and extract meaningful features from images such as human shapes, movements, and interactions between objects. Then, the involved information is compressed into a high-dimensional vector and is expressed as an image in compressed form. Then, as such vectors can be used for tasks such as clustering and classification, it is possible to more accurately express images compared to existing methods such as pixel-based expression. This process enables the algorithm to capture meaningful information from images and improves the performance and accuracy of image-based tasks. The obtained vector is grouped with similar vectors and is entered into clustering algorithms such as K-Means that form clusters. As each cluster represents mutually different motion classes, it can automatically classify image motions. The self-supervised classification through image vectorization provides a method that utilizes a pre-trained CNN encoder and a clustering algorithm to automatically classify image motions without having to label human objects. This model significantly reduces the time and costs required for label designation and improves the expandability of the action classification task. Fig.1 is the result of t-SNE vectorization.

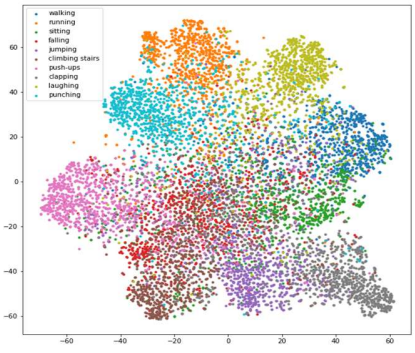

Figure 1: The result of t-SNE vectorization

3 Conclusion

In this study, a novel architecture is proposed to overcome the limitations of YOLO-based object detection technology, by combining object vector extraction process, pretext task, and clustering technology. It is expected to overcome the labeled data insufficiency problem, and, thereby, utilize a large amount of non-labeled data to improve object detection and recognition.

This work was supported by the GRRC program of Gyeonggi province. [GRRC KGU 2020-B03, Industry Statistics and Data Mining Research]