論 文 の 内 容 の 要 旨

論文題目　Video Representation Learning for Action Recognition and Retrieval
（行動認識・検索のための映像表現学習）

氏　　名　陶 砺

The development of digital devices makes it much more convenient for ordinary people to create, edit, and share videos. However, many steps have to be done manually and are time-consuming to ensure high quality when processing videos because video understanding is necessary for many tasks. Many methods have been proposed to extract good videos representations and applied to video understanding tasks, such as action recognition, video retrieval, etc.

Video representation learning is the most fundamental task in video understanding. Good video representations contain sufficient information and can help with a lot of video-related downstream tasks. To obtain good representations, many recent works have required additional calculation on hand-crafted motion features, even though the computation of convolutional neural networks is already very high. And large annotated videos are necessary for specific tasks. In this thesis, we have tackled two video representation learning paradigms (i.e., supervised learning and self-supervised learning) and proposed solutions to obtain good video representations without increasing the complexity of models.

First, we address the task of supervised action recognition. We propose a new data modality with 3D convolutional neural networks, which requires stacked frames (i.e., video clips) as input data. We confirm that by simply replacing traditional RGB video clips with stacked frame differences, the network can extract better temporal information. Greater generalization ability can also be ensured when applying this kind of video representation to other video-related tasks.

Second, we propose a novel learning framework in video self-supervised learning, which can help learn good video representations without any annotations. Intra-negative samples are generated to benefit contrastive learning. We show that by introducing negative samples by breaking the temporal relations while maintaining the spatial similarities, the network can focus more on the temporal clues, resulted in better performance when applied to the downstream video understanding tasks.

Third, we try to bridge the gap between contrastive learning and pretext tasks in video self-supervised learning. We demonstrate that a simple combination of contrastive learning and pretext tasks with proper training strategies can contribute to better video representations than that on their own. We validate the generality of this combination, explore the potential mechanism, and try to reach as closer to the performance limits of traditional video self-supervised learning methods, which are much better than corresponding baselines as reported in the original papers.