Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh

**INTRODUCTION**

Using very deep 2D CNNs trained on ImageNet generates outstanding progress. However, can very deep 3D CNNs trained on Kinetics retrace the successful history? Kinetics should be as large-scale as ImageNet. We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets.

We believe that 3D CNNs trained on Kinetics have the potential to contribute to significant progress in computer vision for videos.

**EXPERIMENTAL CONFIGURATION**

- **Network architectures**: ResNet-18, 34, 50, 101, 152, 200, 200 (pre-act), Wide ResNet-50, ResNeXt-101, DenseNet-121, 201
- **Datasets**: UCF-101, HMDB-51, ActivityNet, Kinetics
- **Implementation**: input size=16frames×112pixels×112pixels, optimization=SGD, data augmentation=(multi-scale spatial crop from 4 corners and 1 center, random temporal crop)

**RESULT 1 | Analyses on each dataset**

Training ResNet-18 on Kinetics did not result in overfitting. It is possible for Kinetics to train deep 3D CNNs.

**RESULT 2 | Analyses of deeper networks**

Changing the model depth from 18 to 200. Improvements of accuracy continued until reaching the depth of 152. Kinetics has sufficient data for training of deep 3D CNNs, and enables training of up to ResNet-152, similar to ImageNet.

**RESULT 3 | Comparisons with SOTA on Kinetics**

Simple 3D architectures pretrained on Kinetics outperform complex 2D architectures.

**RESULT 4 | Analyses of fine-tuning**

ResNeXt-101 achieved the highest accuracy in the models examined in this study.

**GITHUB**

- Training and testing 3D CNNs
- Classifying videos and extracting features of them using pretrained models

Codes and pretrained models are available. 3D-ResNets-PyTorch | GitHub: https://bit.ly/2JeBgCN