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

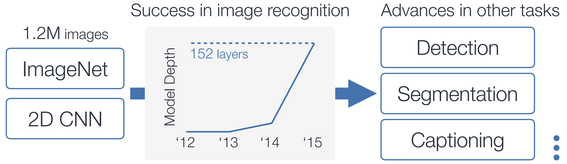
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
AIST

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

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

INTRODUCTION

Using very deep **2D CNNs** trained on **ImageNet** generates outstanding progress.



Can very deep **3D CNNs** trained on **Kinetics** retrace the successful history?

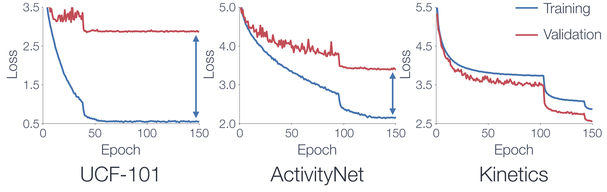
⟶ **Kinetics** should be as large-scale as ImageNet.

We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets.

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=16$_{\text{frames}}$×112$_{\text{pixels}}$×112$_{\text{pixels}}$, optimization=SGD, data augmentation= (multi-scale spatial crop from 4 corners and 1 center, random temporal crop)

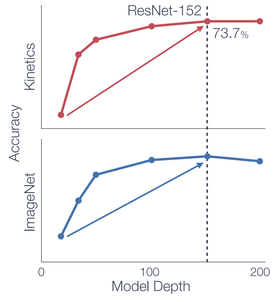
RESULT 1 | Analyses on each dataset



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

Method	Top-1	Top-5	Average
ResNeXt-101	-	-	74.5
ResNeXt-101 (64f)	-	-	**78.4**
CNN+LSTM	57.0	79.0	68.0
Two-stream CNN	61.0	81.3	71.2
C3D w/ BN	56.1	79.5	67.8
RGB-I3D	68.4	88.0	78.2
Two-stream I3D	71.6	90.0	**80.8**

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

ResNeXt-101 (64f) outperformed RGB-I3D even though the input size is still four times smaller than that of I3D.

RESULT 4 | Analyses of fine-tuning

Method	Dim	UCF-101	HMDB-51
ResNeXt-101	3D	90.7	63.8
ResNeXt-101 (64f)	3D	**94.5**	**70.2**
C3D	3D	82.3	-
Two-stream I3D	3D	**98.0**	**80.7**
Two-stream CNN	2D	88.0	59.4
TDD	2D	90.3	63.2
ST Multiplier Net	2D	94.2	68.9
TSN	2D	94.2	69.4

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

GITHUB

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